The Handbook of Phonetic Sciences

Second Edition

Edited by

William J. Hardcastle, John Laver and Fiona E. Gibbon

The Handbook of Phonetic Sciences

Blackwell Handbooks in Linguistics

This outstanding multi-volume series covers all the major subdisciplines within linguistics today and, when complete, will offer a comprehensive survey of linguistics as a whole.

Already published:

The Handbook of Child Language
Edited by Paul Fletcher and Brian MacWhinney

The Handbook of Phonological Theory
Edited by John A. Goldsmith

The Handbook of Contemporary Semantic Theory
Edited by Shalom Lappin

The Handbook of Sociolinguistics
Edited by Florian Coulmas

The Handbook of Phonetic Sciences, Second Edition
Edited by William J. Hardcastle, John Laver, and Fiona E. Gibbon

The Handbook of Morphology
Edited by Andrew Spencer and Arnold Zwicky

The Handbook of Japanese Linguistics
Edited by Natsuko Tsujimura

The Handbook of Linguistics
Edited by Mark Aronoff and Janie Rees-Miller

The Handbook of Contemporary Syntactic Theory
Edited by Mark Baltin and Chris Collins

The Handbook of Discourse Analysis
Edited by Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton

The Handbook of Language Variation and Change
Edited by J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes

The Handbook of Historical Linguistics
Edited by Brian D. Joseph and Richard D. Janda

The Handbook of Language and Gender
Edited by Janet Holmes and Miriam Meyerhoff

The Handbook of Second Language Acquisition
Edited by Catherine J. Doughty and Michael H. Long

The Handbook of Bilingualism
Edited by Tej K. Bhatia and William C. Ritchie

The Handbook of Pragmatics
Edited by Laurence R. Horn and Gregory Ward

The Handbook of Applied Linguistics
Edited by Alan Davies and Catherine Elder

The Handbook of Speech Perception
Edited by David B. Pisoni and Robert E. Remez

The Blackwell Companion to Syntax, Volumes I–V
Edited by Martin Everaert and Henk van Riemsdijk

The Handbook of the History of English
Edited by Ans van Kemenade and Bettelou Los

The Handbook of English Linguistics
Edited by Bas Aarts and April McMahon

The Handbook of World Englishes
Edited by Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson

The Handbook of Educational Linguistics
Edited by Bernard Spolsky and Francis M. Hult

The Handbook of Clinical Linguistics
Edited by Martin J. Ball, Michael R. Perkins, Nicole Müller, and Sara Howard

The Handbook of Pidgin and Creole Studies
Edited by Silvia Kouwenberg and John Victor Singler

The Handbook of Language Teaching
Edited by Michael H. Long and Catherine J. Doughty

The Handbook of Phonetic Sciences

Second Edition

Edited by

William J. Hardcastle,
John Laver, and
Fiona E. Gibbon

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This second edition first published 2010

Edition history: Blackwell Publishers Ltd (1e, 1997)

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical, and Medical business to form Wiley-Blackwell.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ,
United Kingdom

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data
The handbook of phonetic sciences/edited by William J. Hardcastle, John Laver,
Fiona E. Gibbon. – 2nd ed.
p. cm. – (Blackwell handbooks in linguistics)
Includes bibliographical references and index.
ISBN 978-1-4051-4590-9 (hardcover : alk. paper) 1. Phonetics–Handbooks,
manuals, etc. I. Hardcastle, William J., 1943– II. Laver, John. III. Gibbon, Fiona E.
P221.H28 2009
414′.8–dc22
2009033872

A catalogue record for this book is available from the British Library.

Set in 10/12pt Palatino by Graphicraft Limited, Hong Kong
Printed in Singapore

1 2010

For Peter Ladefoged and Gunnar Fant, who led the field

Contents

Contributors

Hermann Ackermann
University of Tübingen

Janet Mackenzie Beck
Queen Margaret University, Edinburgh

Mary E. Beckman
Ohio State University

Rolf Carlson
KTH Royal Institute of Technology, Stockholm

Anne Cutler
Max Planck Institute for Psycholinguistics, Nijmegen
MARCS Auditory Laboratories, University of Western Sydney

Barbara L. Davis
University of Texas

Daniel P. W. Ellis
Columbia University

John H. Esling
University of Victoria

Edda Farnetani
Centro di Studio per le Richerche di Fonetica del CNR, Padova

Janet Fletcher
University of Melbourne

Paul Foulkes
University of York

Christer Gobl
Trinity College Dublin

Björn Granström
KTH Royal Institute of Technology, Stockholm

Helen M. Hanson
Union College, New York

Jonathan Harrington
University of Munich

Hajime Hirose
Kitasato University

Simon King
University of Edinburgh

Anders Löfqvist
Haskins Laboratories, New Haven

James M. McQueen
Max Planck Institute for Psycholinguistics, Nijmegen
Radboud University Nijmegen

Brian C. J. Moore
University of Cambridge

Ailbhe Ní Chasaide
Trinity College Dublin

John J. Ohala
University of California at Berkeley

Daniel Recasens
Universitat Autònoma de Barcelona

Steve Renals
University of Edinburgh

James M. Scobbie
Queen Margaret University, Edinburgh

Christine H. Shadle
Haskins Laboratories, New Haven

Anne Smith
Purdue University

Kenneth N. Stevens
Massachusetts Institute of Technology

Maureen Stone
University of Maryland

Jennifer J. Venditti
San Jose State University

Dominic Watt
University of York

Wolfram Ziegler
City Hospital, Bogenhausen, Munich

Preface to the Second Edition

It is now over 10 years since the publication of the first edition of *The Handbook of Phonetic Sciences*. Since then the phonetic sciences have developed substantially and there are now many more disciplines taking a professional interest in speech-related areas. This multidisciplinary orientation continues to be reflected in the second edition.

In this second edition, 32 leading researchers have contributed 22 chapters in 5 major sectors of the contemporary subject. As with the first edition, an elementary knowledge of the field is assumed and each chapter presents an overview of a key area of the expertise which makes up the wide range of the phonetic sciences today.

There are a number of chapters retained from the first edition which have been substantially updated by the authors. These include the chapters by Stone, Shadle, Hirose, Mackenzie Beck, Farnetani and Recasens, Löfqvist, Gobl and Ní Chasaide, Stevens and Hanson, Moore, McQueen and Cutler, Ohala, Carlson and Granström. Other topic areas from the first edition have been given completely new treatment by newly commissioned authors (chapters by Harrington, Ackermann and Ziegler, Smith, Davis, Ellis, Renals and King). There are also two new chapters covering sociophonetics (Scobbie, Foulkes, and Watt) and phonetic notation (Esling). To reflect the increasing significance of the area of prosody in the phonetic sciences we have also included two commissioned chapters covering the areas of timing and rhythm (Fletcher), and tone and intonation (Beckman and Venditti).

For readers with complementary interests in phonology and clinical phonetics and linguistics the companion volumes to this handbook, *The Handbook of Phonological Theory* (Goldsmith, 2010, 2nd edn.) and *The Handbook of Clinical Linguistics* (Ball, Perkins, Müller, & Howard, 2008) are recommended.

We would like to thank a number of colleagues for their assistance with editorial work, including Annabel Allen, Pauline Campbell, Erica Clements, Sue Peppe, and Sonja Schaeffler. Special thanks are also due to Anna Oxbury for her meticulous and thoughtful copy-editing.

The editors

Introduction

WILLIAM J. HARCASTLE, JOHN LAVER, AND FIONA E. GIBBON

As with the first edition, the book is divided into five major sections. The first part begins with an account of the main measurement techniques, methodologies, and instruments found in experimental phonetic laboratories. The next part explores aspects of the anatomical and physiological framework for normal and disordered speech production. The third and largest part of the book focuses on the acquisition of speech and theories and models of speech production and perception. The fourth part deals with the linguistic motivation of much research in the phonetic sciences in covering a number of key areas of linguistic phonetics. The final part returns to experimental approaches to the phonetic sciences but this time focusing on speech signal processing and engineering in an overview of the main developments in speech technology. There are extensive pointers to further reading in each chapter.

Part I has four chapters on the topic of Experimental Phonetics. The section begins with a critical evaluation by Maureen Stone on current laboratory techniques that measure the oral vocal tract during speech. The focus is on instruments that measure the articulators directly and indirectly. Indirect measurements come from instruments that are remote from the structures of interest such as imaging techniques (e.g., X-ray, MRI, and ultrasound). Direct measurements come from instruments that contact the structures of interest, such as, point-tracking devices and electropalatography. References are made to current research using each instrument in order to indicate its applications and strengths.

Experimental approaches to speech production are explored further by Christine Shadle in the next chapter on the aerodynamics of speech. This chapter begins by defining aerodynamics and reviews the basic concepts of fluid statics and dynamics (including turbulence), and aerodynamically distinct vocal tract behaviors are discussed. This is followed by a section covering measurement methods, divided into basic methods such as pressure and flow velocity measurement, and speech-adapted methods such as the Rothenberg mask and methods for measuring or estimating lung volume and subglottal pressure, and the use of hot-wires to measure flow velocities in the vocal tract. A final section describes models of speech production that incorporate aerodynamics.

Acoustic phonetics is the subject of the third chapter by Jonathan Harrington. This new chapter provides an overview of the acoustic characteristics of consonants and vowels from the perspective of a broad range of research questions in experimental phonetics and laboratory phonology. Various procedures for the phonetic classification of the acoustic speech signal are reviewed including the identification of vowel height and backness from various transformed acoustic spaces, the derivation of place of articulation in oral stops from burst and locus cues, and techniques for distinguishing between fricatives based on parameterizing spectral shape. These techniques are informed by a knowledge of speech production and are related to speech perception, and they also establish links to pattern classification in signal processing.

Investigating the physiology of laryngeal structures is the subject of the final chapter in this section. In this chapter, Hajime Hirose describes specialized, newly developed techniques for observing laryngeal behavior during speech production, including flexible fiberscopy, high-speed digital imaging, laryngeal electromyography, photoglottography, electroglottography, and magnetic resonance imaging. Basic behaviors of the laryngeal structures are described with reference to the results of observation obtained by the above techniques and the nature of laryngeal adjustments that take place under different phonetic conditions.

Part II contains three chapters on biological perspectives and opens with an exploration by Janet Mackenzie Beck on organic variation and the ways it affects the vocal apparatus. She points to two main sources of variation in speech performance: phonetic variation resulting from differences in the way individuals use their vocal apparatus, and organic variation depending on individual differences in inherent characteristics of the vocal organs. The chapter focuses on organic variation bringing together information from a variety of sources, anatomical, physiological, anthropological. Three main types of differences in the structure of the vocal apparatus are discussed: the life-cycle changes within an individual; genetic or environmental factors which differentiate between individuals; and differences which result from trauma or disease.

Hermann Ackermann and Wolfram Ziegler in their chapter on brain mechanisms underlying speech motor control begin with an overview of the topic. Their discussions draw upon data derived from three approaches, namely, electrical surface stimulation of the cortex, lesion studies in patients with neurogenic communication disorders, and functional imaging techniques. These discussions are preceded by a review of experimental studies in subhuman primates addressing the corticobulbar representation of orofacial muscles as well as the cerebral correlates of vocal behavior.

The final chapter in Part II is by Anne Smith and concerns the development of neural control for speech. She gives an integrative overview of studies of the development of the neuromotor processes involved in controlling articulatory movements for speech. The area of speech motor development has not been critically reviewed recently and this chapter provides a detailed summary of major advances in understanding the time course of maturation of speech motor control processes, which, contrary to earlier claims, are not adult-like until late adolescence. Discussions of theoretical issues in speech motor development, such as the units involved in the language–motor interface and the issues of neural plasticity and sensitive periods in speech motor development, portray important, ongoing debates in this area.

Part III contains seven chapters on the topic of modeling speech production and perception. The first is a chapter on speech acquisition by Barbara Davis. She addresses the question of how young children integrate biology and cognition to achieve the necessary capacities for the phonological component of linguistic communication. The chapter outlines how contemporary theoretical perspectives and research paradigms consider the nature of speech acquisition. These include formalist phonological perspectives representing a consistent strand of proposals on acquisition of sound patterns in languages. She contrasts this approach with functionalist phonetic science perspectives that have focused on biological characteristics of the developing child and the ways in which these capacities contribute to emergence of complex speech output patterns.

The chapter by Edda Farnetani and Daniel Recasens presents an overview of the current knowledge concerning coarticulation and connected speech processes. The authors address the nature of coarticulatory and assimilatory processes in connected speech, and explore the foundations and predictions of the most relevant theoretical models of labial, velar, and lingual coarticulation (feature spreading, time-locked, locus equation, adaptive variability, window model, and coarticulatory resistance). They describe the significant theoretical and experimental progress in understanding contextual variability, which is reflected in continuously evolving and improving models, and in increasingly rigorous and sophisticated research methodologies.

Theories and models of speech production are developed further by Anders Löfqvist, particularly from the point of view of spatial and temporal control of speech movements. In his chapter, theoretical and empirical approaches to speech production converge in their focus on understanding how the different parts of the vocal tract are flexibly marshaled and coordinated to produce the acoustic signal that the speaker uses to convey a message. He outlines a variety of experimental paradigms and how these are applied to the problem of coordination and control in motor systems with excess degrees of freedom.

An area of key theoretical and technical importance is the nature of the voice source and how it varies in speech. The chapter by Christer Gobl and Ailbhe Ní Chasaide is concerned with acoustic aspects of phonation and its exploitation in speech communication. The early sections focus on the source signal itself, on analysis techniques, and provide acoustic descriptions of different voice qualities. The later sections describe how variations in the voice source are associated with segmental or suprasegmental aspects of the linguistic code, and discuss the role of voice quality in the paralinguistic signaling of emotion, mood, and attitude. The sociolinguistic function in differentiating among linguistic, regional, and social groups is briefly outlined, as well as its important role in speaker identification.

The next chapter by Kenneth Stevens and Helen Hanson focuses on articulatory–acoustic relations as the basis of distinctive contrasts. The chapter provides a physical basis for the inventory of binary distinctive features or phonological contrasts that are observed in language. The chapter is a major update on the quantal nature of speech, and the authors show how aerodynamic and acoustic properties of speech production lead to quantal relations between the articulatory parameters and the acoustic consequences of these variations. The chapter also proposes how listeners might extract additional enhancing cues as well as cues relating to the defining quantally-based properties of the acoustic signal in running speech. Other approaches that have been proposed to account for variability in speech are also described.

The final two chapters in Part III deal with aspects of auditory processing and speech perception. The first chapter by Brian Moore reviews selected aspects of auditory processing, chosen because they play a role in the perception of speech. The review is concerned with basic processes, many of which are strongly influenced by the operation of the peripheral auditory system and which can be characterized using simple stimuli such as pure tones and bands of noise. He discusses the resolution of the auditory system in frequency and time, as revealed by psychoacoustic experiments. A consistent finding is that the resolution of the auditory system usually markedly exceeds the resolution necessary for the identification or discrimination of speech sounds. This partly accounts for the fact that speech perception is robust, and resistant to distortion of the speech and to background noise.

James McQueen and Anne Cutler in their chapter focus on the cognitive processes involved in speech perception. They describe how recognition of spoken language involves the extraction of acoustic-phonetic information from the speech signal, and the mapping of this information onto cognitive representations. They focus on our ability to understand speech from talkers we have never heard before, and to perceive the same phoneme despite acoustically different realizations (e.g., by a child's voice versus an adult male's). They show how processing of segmental, lexical and suprasegmental information in word recognition contributes significantly to listeners' processing decisions.

The five chapters in Part IV cover different aspects of linguistic phonetics, and begins with two new chapters on speech prosody. Janet Fletcher explores rhythm and timing in speech with a particular focus on how durational patterns of segments and syllables contribute to the signaling of stress and/or accent and prosodic phrasing in different languages. The chapter summarizes the contribution of durational patterns of segments, morae, and syllables to the rhythm and tempo of spoken language, and evaluates the different kinds of metrics that are often used in experimental investigations. What emerges is a complex picture of how speech unfolds in time, and crucially how the temporal signatures of prosody in a language are often accompanied by additional qualitative acoustic and articulatory modifications, rather than just adjustment of measurable duration alone.

In the second chapter on speech prosody, Mary Beckman and Jennifer Venditti examine tone and intonation. The authors begin by reviewing the ways in which pitch patterns are represented in work on tone and intonation. A key point in this review is that symbolic representations are phonetically meaningful only if they are tags for parameter settings in an analysis-by-synthesis model of f_0 contours. The most salient functions of lexical contrast, prosodic grouping, and prominence marking are described in a way that makes clear that many aspects of the pitch pattern can simultaneously serve one, two, or all three of these functions. The authors conclude by suggesting that broad-scale typologies that differentiate only between two or three language "types" (e.g., "tone languages") are overly simplistic.

The next chapter by John Ohala explores the relation between phonetics and phonology. In tracing the history of this relationship from the early part of the last century, he shows it has been affected by theoretical frameworks such as structuralist phonology, in which more attention was given to relations between sounds at the expense of substance of sounds. It is proposed that in order to explain sound patterns in language, phonology needs to re-integrate scientific phonetics (as well as psychology and sociolinguistics). The author provides examples where principles of aerodynamics and acoustics are used to explain certain common sound patterns.

John Esling's chapter on phonetic notation reviews the theoretical constructs of how speech sounds are transcribed using phonetic notation. He presents the International Phonetic Alphabet (IPA) as a common core of standard usage that transcribers of language can universally refer to and understand. Orthographic, iconic, and alphabetic notation are differentiated, and the phonetic relationships between sets of symbols are addressed. A revised version of the IPA consonant chart is developed, as well as a novel way of looking at the IPA vowel chart. Place of articulation, manner of articulation, vowel classification, and secondary articulation are discussed where they present challenges to notational conventions. He also discusses notation for stress and juncture, strength of articulation, voice quality, and clinical usage for transcribing disordered speech.

The last chapter in Part IV is on sociophonetics. In this chapter, Paul Foulkes, James Scobbie, and Dominic Watt provide an overview of sociophonetics as an area of the phonetic sciences which takes into account the systematic subtle differences in phonetic systems which attach to social groups. This structured variation informs theoretical debate in fields such as sociolinguistics, phonetics, phonology, psycholinguistics, typology, and diachronic linguistics. In their chapter, Foulkes, Scobbie, and Watt survey work which touches on all these areas, although sociolinguistics features most strongly. The chapter addresses both production and perception studies, before moving on to consider contemporary methodological issues and the general theoretical implications that arise from the literature.

Part V contains three chapters that are concerned with issues relating to speech technology. Most speech technology applications rely on digital signal processing and Daniel Ellis presents an introduction to the topic of signal processing for speech. His chapter emphasizes an intuitive understanding of signal processing in place of a formal mathematical presentation. He begins with familiar daily experiences of resonance and oscillation, for instance as seen in a pendulum, and builds up to the ideas of decomposing signals into sinusoids (Fourier analysis), filtering, and the familiar speech-related tools of the spectrogram and cepstral coefficients. All of this is done without a single equation, but in a way that may help cement insights even for readers already familiar with more technical presentations.

The next chapter, by Rolf Carlson and Björn Granström, is a survey of speech synthesis systems. They review some of the more popular approaches to speech synthesis and show how it is no longer simply a research tool but has many everyday applications. They describe current trends in speech synthesis research and point to some present and future applications of text-to-speech technology.

Part V concludes with a chapter on automatic speech recognition by Steve Renals and Simon King. They define automatic speech recognition as the task of transforming an acoustic speech signal to the corresponding sequence of words. Their chapter provides on overview of the statistical, data-driven approaches which now comprise the state-of-the-art. The chapter outlines the decomposition of the problem into acoustic modeling and language modeling and provides a flavor of some of the technical details that underpin this research field, as well as outlining some of the major open challenges.

We would like to conclude by offering our warmest thanks to all the contributors. We believe that the 22 chapters in the second edition of this handbook give an exciting as well as a representative flavor of the productive multidisciplinary research that typifies the phonetic sciences today.

Part I Experimental Phonetics

1 Laboratory Techniques for Investigating Speech Articulation

MAUREEN STONE

This chapter discusses current laboratory techniques that measure the oral vocal tract during speech. The focus is on instruments that measure the articulators directly and indirectly. Indirect measurements come from instruments that are remote from the structures of interest such as imaging techniques. Direct measurements come from instruments that contact the structures of interest, such as, point-tracking devices and electropalatography. Although some references are made to current research using each instrument, to indicate its applications and strengths, the list of studies is not comprehensive as the goal is to explain the instrument.

Measuring the vocal tract is a challenging task because the articulators differ widely in location, shape, structural composition, and speed and complexity of movement. First, there are large differences in tissue consistency between soft tissue structures (tongue, lips, velum) and hard tissue structures (jaw, palate), which result in substantially different movement complexity. In other words, the fluid deformation of the soft structures and the rigid movements of the bones need different measurement strategies. Second, measurement strategies must differ between structures visible to superficial inspection, such as the lips, and structures deep within the oral cavity, such as the velum. Third, articulator rates of motion vary, so that an instrument with a frequency response appropriate for the slow-moving jaw will be too slow for the fast-moving tongue tip. The final and perhaps most important measurement complication is the interaction among articulators. Some articulatory behaviors are highly correlated, and distinguishing the contributions of each player can be quite difficult. The most dramatic example of this is the tongue–jaw system. It is clear that jaw height is a major factor in tongue tip height. However, the coupling of these two structures becomes progressively weaker as one moves posteriorly, until in the pharynx, tongue movement is only minimally coupled to jaw movement if at all. Thus, trying to measure the contribution of the jaw to tongue movement becomes a difficult task.

It is difficult to devise a transducer that can be inserted into the mouth, which will not in some way distort the speech event. Thus, the types of instruments used in the vocal tract need to be unobtrusive, such as by resting passively against a surface (e.g., electropalatography), by being small and positioned on noncontact surfaces (e.g., pellet tracking systems), or by not entering the vocal tract at all (e.g., imaging techniques).

Instruments that enter the oral cavity must meet certain criteria. They need to be unaffected by temperature change, moisture, or air pressure. Affixatives must be unaffected by moisture, nontoxic, able to stick to expandable, moist surfaces, and must be removable without tearing the surface tissue. Devising instruments that are noninvasive, unobtrusive, meet the above criteria, and still measure one or more components of the speech event is so difficult that most researchers prefer to study the speech wave and infer physiological events from it. However, since those inferences are based on, and refined by, physiological data, it is critical to add new physiological databases, lest models of the vocal tract and our understanding of speech production stagnate.

In recent times, physiological measurements have improved at an extraordinary pace. Imaging techniques are revolutionizing the way we view the vocal tract by providing recognizable images of structures deep within the pharynx. They also provide information on local tissue movement and control strategies. Point-tracking systems and palatographic measurements have transformed our ideas about coarticulation by revealing inter-articulator relationships that could only in the past be addressed theoretically. Applications to linguistics and rehabilitation are now ongoing. This chapter considers indirect measurements, that is, imaging techniques, and direct measurements such as point-tracking techniques, and tongue–palate measurement devices

1 Imaging Techniques

The internal structures of the vocal tract are difficult to measure without impinging upon normal movement patterns. Imaging techniques overcome that difficulty because they register internal movement without directly contacting the structures. Four well-known imaging techniques have been applied to speech research: X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound. Imaging systems provide recordings of the entire structure, rather than single points on the structure.

1.1 X-ray

X-ray is the most well known of the imaging systems. It is important because it was the first widely used imaging system and most of our historical knowledge about the pharyngeal portion of the vocal tract came from X-ray data. To make a lateral X-ray image, an X-ray beam is projected from one side of the head through all the tissue, and recorded onto a plate on the other side. The resulting image shows the head from front to back and provides a lengthwise view of the tongue.

A frontal or anterior–posterior (AP) X-ray is made by projecting the X-ray beam from the front of the head through to the back of the head and recording the image on a plate behind the head. The resulting images provide a cross-sectional view of the oral cavity. Prior to the advent of MRI considerable research was done using X-ray imaging. More recent X-ray studies are based on archival databases.

X-ray data have contributed to many aspects of speech production research. Many vocal tract models are based on X-rays (cf. Fant, 1965; Mermelstein, 1973; Harshman et al., 1977; Wood, 1979; Hashimoto & Sasaki, 1982; Maeda, 1990). X-rays have also been used to study normal speech production (Kent & Netsell, 1971; Kent, 1972; Kent & Moll, 1972), nonspeech motions (Kokawa et al., 2006), motor control strategies (Lindblom et al., 2002; Iskarous, 2005), language differences (cf. Gick, 2002b; Gick et al., 2004), and speech disorders (Subtelny et al., 1989; Tye-Murray, 1991).

Usually soft tissue structures such as the tongue are difficult to measure with X-rays, because the beam records everything in its path including teeth, jaw, and vertebrae. These strongly imaged bony structures obscure the fainter soft tissue. Another limitation of X-ray is that unless a contrast medium is used to mark the midline of the tongue, it is difficult to tell if the visible edge is the midline surface of the tongue or a lateral edge. This is particularly problematic during speech, because the tongue is often grooved or arched. Finally, the potential hazards of overexposure have reduced the collection of large quantities of X-ray data. There is, however, public availability of archival X-ray databases for research use. One such database (Munhall et al., 1994a, 1994b) was compiled by Advanced Technologies Research Laboratories, Kyoto, and is available from http://psyc.queensu.ca/~munhallk/05_database.htm.

1.2 Tomography

Tomography is a fundamentally different imaging method from projection X-ray in that it records slices of tissue. Three tomographic techniques used in speech research are Computed Tomography, Magnetic Resonance Imaging, and Ultrasound Imaging. These slices are made by projecting a thin, flat beam through the tissue in one of four planes: sagittal, coronal, oblique, and transverse (see Figure 1.1). The mid-sagittal plane is a longitudinal slice, from top to bottom, down the median plane, or midline, of the body (dashed line – upper right). The para-sagittal plane is parallel to the midline of the body and off-center (not shown). The coronal plane is a longitudinal slice perpendicular to the median plane of the body. The oblique plane is inclined between the horizontal and vertical planes. Finally, the transverse plane lies perpendicular to the long axis of the body, and is often called the transaxial, or in MRI, the axial plane.

1.2.1 Computed Tomography (CT) Computed Tomography uses X-rays to image slices (sections) of the body as thin as 0.5 mm or less. Tomographic images

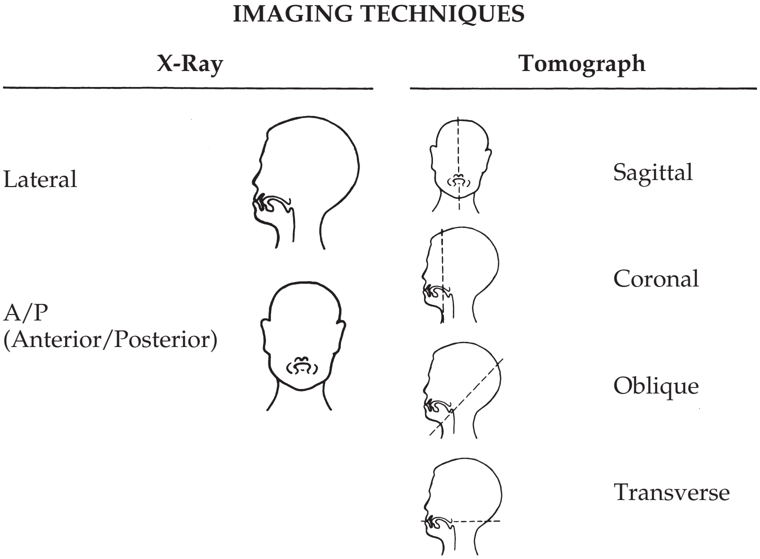


Figure 1.1 Scan types used in through-transmission and tomographic imaging. There are two X-ray angles contrasted with four tomographic scanning planes.

made in coronal planes are made by projecting very thin X-ray beams through a slice of tissue from multiple origins. The scanner rotates around the body taking these images and a computer creates a composite, including structures that are visible in some scans but obscured in others. Using this technique, tissue slices can be collected rapidly, 15 Hz or faster, and multiple slices can be collected simultaneously. CT images soft tissue more clearly than X-rays because it produces a composite X-ray. By digitally summing a series of scans, the composite section has sharper edges and more distinct tissue definition. From the multislice datasets, planar sections can be reconstructed in any direction. CT images can produce excellent resolution of soft and hard tissue structures. Figure 1.2, for example, is a reconstructed image of the midsagittal plane of the vocal tract. Bone appears bright white in the image, soft tissue structures are gray. In this figure, the junction of the velum and hard palate can be seen to be quite complex. The soft tissue below the hard palate widens before the velum emerges as a freestanding object. It is clear from this image that the shape of the palatine bone is not well reflected in the soft tissue. Measures of the palate bone made from an MRI or ultrasound image will differ from measurements made directly in the mouth or from dental impressions. Without this image, those differences would be hard to interpret.

Another method of CT data collection is Spiral CT. Spiral CT collects multiple slices at the same time by collecting a single spiral-shaped slice instead of multiple flat planar slices. In the mid 1980s, the cable and drum mechanism for

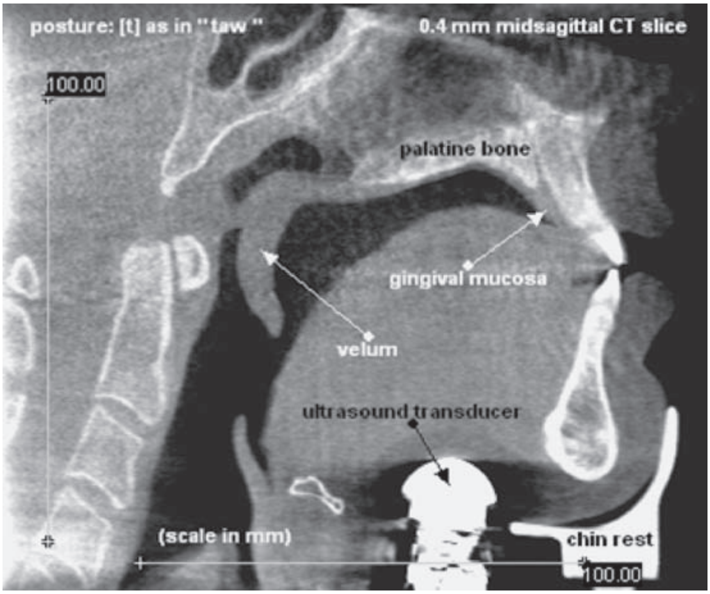


Figure 1.2 Midsagittal CT of vocal tract reconstructed from axial images. Bone is white; soft tissue is gray. (Reproduced courtesy of Ian Wilson)

powering the rotation of the CT machine was replaced with a slip ring. The slip ring allows the CT scanner to rotate continuously, creating a spiral image. Spiral CT scans have very high resolution, but currently take 20–30 seconds to create, and hence are too slow for imaging continuous speech, though excellent for static images (Lell et al., 2004).

Electron Beam CT was developed to measure calcium deposits around coronary arteries. Its principles are similar to CT, but it uses an electron "gun" instead of regular X-ray. EBCT collects a set of parallel images that are reconstructed as a 3D volume. EBCT is a fast acquisition technique and therefore has been used to collect vocal tract images for datasets requiring short acquisition times. For example, Tom et al. (2001) scanned the entire vocal tract in under 90 seconds, to compare vocal tract shapes during falsetto and chest registers.

Although CT has been used to image the vocal tract, it is not the instrument of choice for speech research because of radiation exposure and because MRI provides much the same information, albeit at a lower spatial and temporal resolution. In fact, the major limitation of CT is that it has more radiation exposure than traditional X-ray, because it images thinner slices, and each slice is scanned several times to collect multiple images. Another limitation is that the subject is supine or prone, so gravitational effects on the subject differ from upright. On the positive side, 3D reconstructions can be made and sliced in any plane, and images are clear and easy to measure.

1.2.2 Magnetic Resonance Imaging (MRI) Another tomographic technique is Magnetic Resonance Imaging, which uses a magnetic field and radio waves rather than X-rays to image a section of tissue. There are a number of MRI procedures that yield a variety of information: high-resolution MRI, cine MRI, tagged-snapshot MRI, tagged-cine MRI, diffusion tensor MRI, and functional MRI. All of these use identical hardware: typically 1.5 or 3 Tesla machines. The differences lie in the software algorithms, which are designed to exploit different features of the relationship between the hydrogen proton, magnetic fields, and radio waves.

An MRI scanner consists of electromagnets that surround the body and create a magnetic field. MRI scanning detects the presence of hydrogen atoms, which occur in abundance in water and, therefore, in human soft tissue. Figure 1.3 depicts the MRI process. Picture (a) represents hydrogen protons spinning about

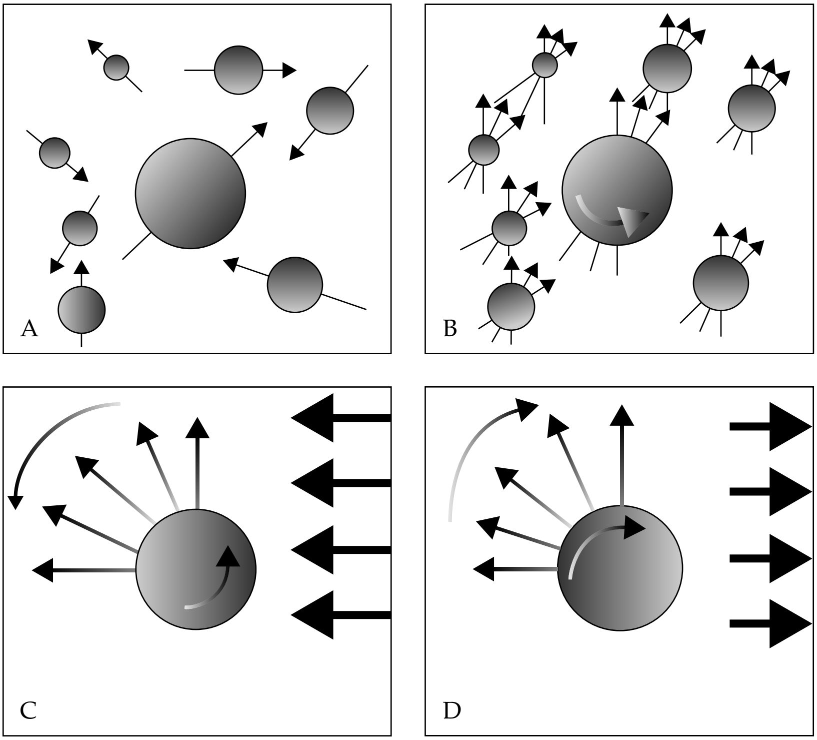


Figure 1.3 MRI recording of the amount of hydrogen in tissue. Hydrogen protons spin about axes that are oriented randomly (A). The MRI magnet causes them to align to the long axis of the body, but with a small precession (wobble) (B). A radio-frequency pulse knocks them out of alignment (C). As the protons realign to the magnet (D) they emit a radio pulse that is read by the scanner.

their axes, which are oriented randomly. (b) shows what happens when a magnetic field is introduced. The protons' axes align along the direction of the field's poles. Even when aligned, however, the protons wobble, or precess. In (c) a short-lived radio pulse, vibrating at the same frequency as the precession, is introduced. This momentarily knocks the proton out of alignment. (d) shows the proton realigning, within milliseconds, to the magnetic field. As the proton realigns, it emits a weak radio signal of its own. Period (d) is when the MR image is "read." The radio signals are summed until the protons return to position (b). The resulting data are constructed into an image that reflects the hydrogen content (i.e., the amount of water or fat) of the different tissues. Because the proton emissions are weak, the process is repeated many times and the data are summed into a single image. If the process is repeated for several minutes, while the subject holds still, high-resolution images result.

MRI measurement of oral structures has replaced X-ray for many research applications. MR images have been used to detail developmental vocal tract anatomy and function (Xue & Hao, 2003; Vorperian et al., 2005). MRI also has provided quite accurate extraction of vocal tract surfaces (Story et al., 1996). These surfaces have been used to calculate 3D vocal tract volumes for modeling geometry to acoustic relationships (Tameem & Mehta, 2004; Story, 2005). Extracted edges have also been used to model 3D structures within the vocal tract. Serrurier and Badin (2005) modeled velar position for French vowels from MRI and CT images. Engwall (2003) modeled tongue position for Swedish vowels from MRI, Electromagnetic Articulography (EMA), and Electropalatography (EPG). Story et al. (1996) modeled vocal tract airway shapes for 18 English phonemes from MRI. MRI is very good at characterizing different types of soft tissue and therefore is quite successful in identifying tumors and soft tissue pathology. For example, Lenz et al. (2000) used MRI and CT together to stage oral tumors and Lam et al. (2004) had good success using MRI T1 and T2 weighted images to determine tumor thickness.

Two types of MRI are used particularly to characterize tissue: high-resolution MRI (hMRI) and diffusion tensor MRI (DTI). Figure 1.4 shows a high-resolution sagittal MRI image of the vocal tract at rest. The vocal tract appears black, as do the teeth, since neither contains water. Water and fat, both of which are high in hydrogen, are found in marrow, seen in the palate and mandible. Muscles are visible in the tongue, velum, and lips. The other method of characterizing soft tissue is diffusion tensor MRI (DTI), which measures 3D fiber direction, typically in *ex-vivo* structures. DTI, developed in the early 1990s, visualizes fiber direction by measuring random thermal displacement of water molecules in the tissue. The direction of greatest molecular diffusion parallels the local fiber direction. DTI has virtually microscopic spatial resolution and distinguishes tissue fibers with their orientations for any muscles. A fiber map can be drawn and superimposed on an MRI structural image. The fiber map is 3D and can differentiate among nerve fiber pathways and detail anatomical structures based on their fiber architecture. There are limitations of this technique that impede the measurement of oropharyngeal structures. First, when fiber directions cross within a

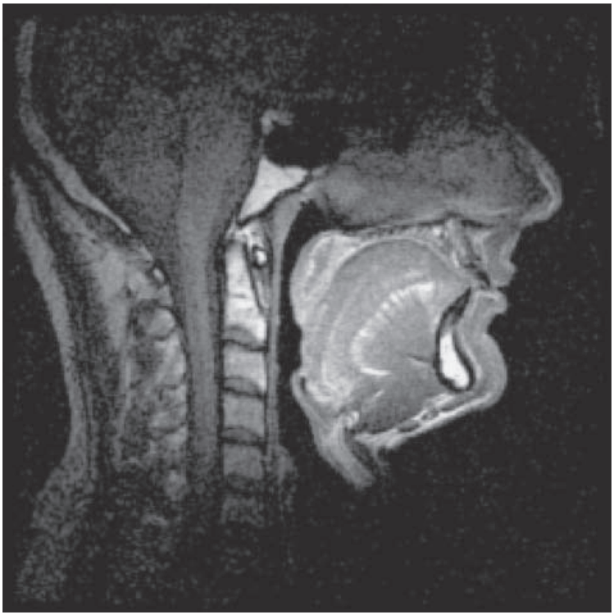


Figure 1.4 High-resolution MRI (hMRI) of the midsagittal vocal tract at rest.

single voxel (3D pixel), visualization of the underlying fiber structure is reduced. Fiber interdigitation is typical in oral musculature, especially the tongue, lips, and velum. Second, DTI is sensitive to motion and the structure must remain immobile for several minutes to record a volumetric scan. Using long collection times, DTI has been used to study the excised tongues of animals (Wedeen et al., 2001) and humans (Gilbert & Napadow, 2005). In addition, DTI can be used *in vivo* with cooperative subjects to collect data in as little as 3–5 minutes. Figure 1.5 shows a fiber map indicating the fan-like fibers of the genioglossus muscle, which run from superior–inferior to anterior–posterior in direction. This image was taken from an *in vivo* human tongue at rest (Shinagawa et al., 2008, 2009).

When measuring vocal tract motion, Cine-MRI is of particular interest. Cine-MRI is similar to other cine techniques, such as videofluoroscopy or movies, in that it divides a moving event into a number of still frames. Because MRI sums proton emissions over time, it typically takes a long time to reconstruct a single image, and collecting data during speech motion is challenging. Cine-MRI is often done by having the subject repeat a task multiple times and summing data from each frame across repetitions, similar to ensemble averaging. This technique has been used to compare vocal tract behaviors during speech production (Magen et al., 2003), especially vowel production (Hasegawa-Johnson et al., 2003; Story, 2005; McGowan, 2006). However, the subject must produce the repetitions very

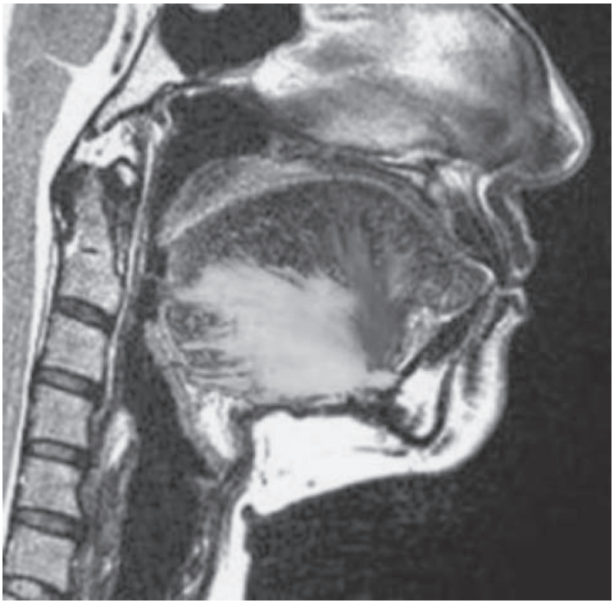


Figure 1.5 Diffusion Tensor Image shows fiber directions of the genioglossus muscle (light gray). Fibers are overlaid on a high resolution image of the head.

precisely to prevent image blurring. It is also possible to compare MRI vocal tract data to speech acoustics, either by collecting the speech wave independently from the (quite noisy) MRI data collection, or by using subtraction microphones within the MRI machine itself (cf., NessAiver et al., 2006). Some Cine-MRI data have been collected with single repetitions of a task (Mády et al., 2002; Narayanan et al., 2004). These images usually have a reduced frame rate or a reduced image quality, but typically have sufficient spatial resolution to extract surfaces of vocal tract structures. Faster frame rates and good image quality require several repetitions per slice (Stone et al., 2001a). Rapid collection of MRI frames using a single repetition has been reported using spiral MRI (Narayanan et al., 2004). This method acquires images in interleaved spirals rather than planes. Although Cine-MRI does not produce the quality of anatomy seen in high resolution (hMRI) images, the speed of pellet-tracking systems, or the level of muscle detail seen in DTI, it is able to answer many speech questions that were previously unstudied. Therefore, Cine-MRI is very popular in the study of speech.

Concurrent with the development of Cine-MRI, was the development of Tagged Snapshot MRI. Tags are created by applying a spatial gradient to the tissue, which demodulates the spinning protons in alternating planes. In the demodulated planes the protons spin out of phase with the rest of the tissue. The demodulated protons are invisible to the machine when the image is read, thus the invisible proton planes appear as black stripes on the image (see Figure 1.6). After tags are

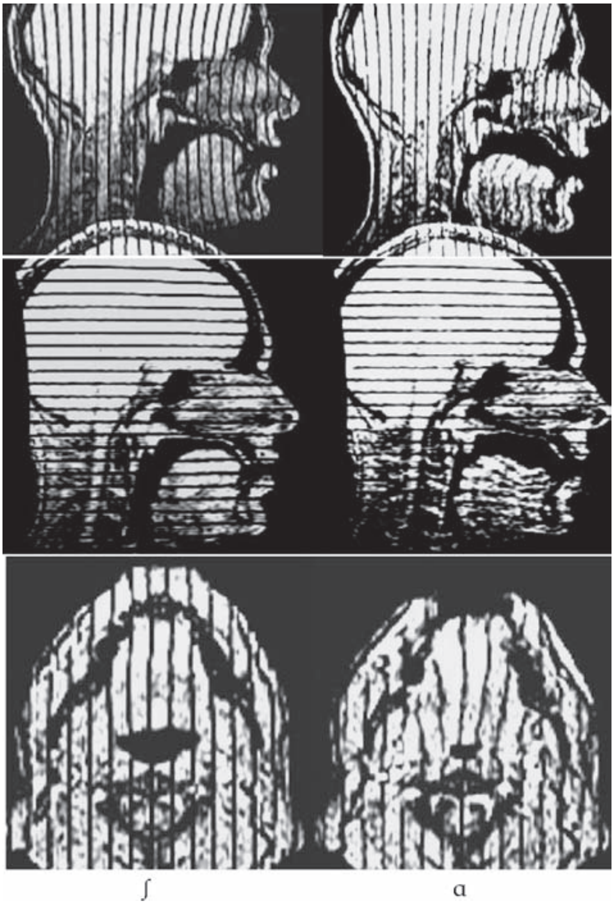


Figure 1.6 Tagged MR images in three planes. The left column shows the reference state of the tongue, just after the tags were applied during the sound /ʃ/ The right column shows the tongue in the deformed state, after the tongue has moved into /ɑ/ position.

applied to the tissue, the object of interest, such as the tongue or lips, is moved. Because magnetization stays with the tissue, the tags deform to exactly the same extent as the tissue. Figure 1.6 shows reference frames in three planes taken during /ʃ/ and deformed frames taken during the /ɑ/ in "sha." In Tagged Snapshot MRI only two images are read. One is before the motion (reference) and the other is during or after the motion (deformed). A grid of tags is created by applying tags in horizontal and vertical planes in immediate succession prior to reading the image, or by combining two datasets of orthogonal tags collected separately. Niitsu et al. (1994) examined tag positions from rest (before) and a vowel position (after) to derive the direction of the movement. Napadow et al. (1999a, 1999b)

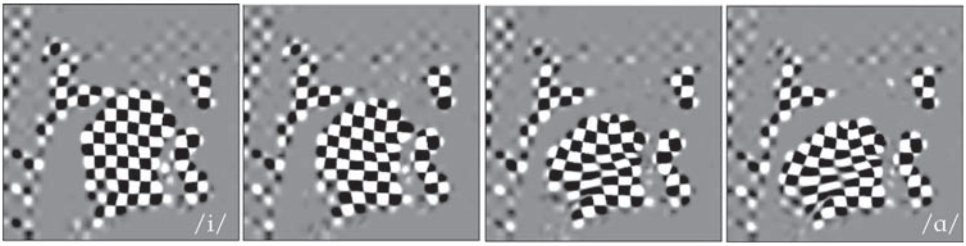


Figure 1.7 Tagged Cine-MRI sequence shows motion from /i/ to /ɑ/. Checkerboard deformation shows local expansion, compression, and shear.

studied swallowing and nonspeech tongue motions by having subjects repeat the task multiple times and each time collecting a deformed image at a later time. These images were then put into a pseudo-motion sequence reflecting the movement.

Cine-MRI and Tagged Snapshot MRI can be combined to form Tagged Cine-MRI (tMRI). tMRI captures internal tissue motion over time during the performance of a task. The first applications of tMRI were studies of the heart's motion and internal tissue characteristics (Zerhouni et al., 1988; Axel and Dougherty, 1989). Continuous tongue deformation during speech has also been measured using tMRI. Figure 1.7 shows deformation of the midsagittal tongue between the two vowels /i/ and /ɑ/. The black and white squares are a visualization device to better depict the deformations. The change in their shapes over time demonstrates features of local tissue deformation. From these images tags can be tracked to directly measure positions and motion of all tissue points in the tongue (Parthasarathy et al., 2007). From the tissue point motions one can calculate displacement, velocity, and local strain (compressions and expansions).

Functional MRI (fMRI) is a method of MRI scanning that measures changes in blood flow in the brain. Increased blood flow characterizes increased uptake of blood in the active region of the brain. Because blood is high in hydrogen, the local increase in activity can be measured by MRI. The premise is that spatially distinct, distributed areas of the brain are connected into functional networks organized to produce specific tasks. If these networks can be imaged, they can be mapped geographically to detail brain function during various behaviors. Increased neural activity in a region causes increased demand for oxygen, and that oxygen is brought to the region by blood. Replacement of deoxygenated blood with oxygenated blood produces a more uniform magnetic environment, which increases the longevity of the MRI signal. fMRI signals are snapshots that are collected at 10–20 Hz. Multislice recording of the brain and multiple repetitions can be recorded. Some limitations exist in the use of fMRI for speech research. First, the visible effects of the increase in oxygen occur some time after the event itself (0.5–8.0 seconds), which results in poor temporal resolution. Second, signals can vary even with no change in brain state. To overcome the latter problem,

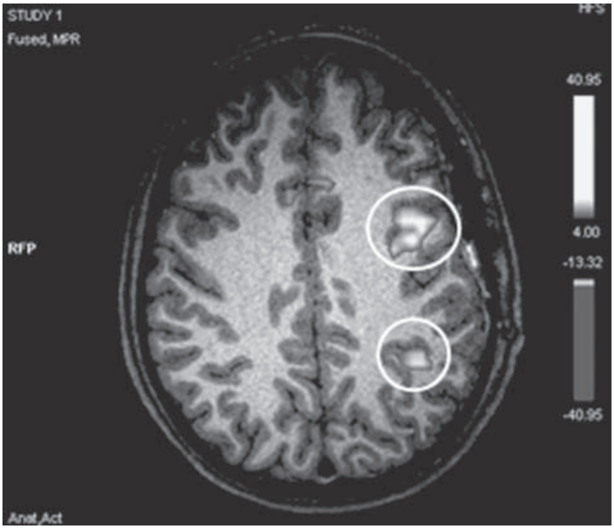


Figure 1.8 Axial fMRI scan showing regions of the brain that are active while subject thinks of words. The white regions (circled) indicate activity in the left hemisphere. (Photo courtesy of Rao Gullapalli, University of Maryland Medical School)

fMRI usually compares sets of images before and after the task. Despite these difficulties, research on speech and language is being done including studies of cortical aspects of speech production (Gracco et al., 2005), speech perception (Specht et al., 2005; Pulvermuller et al., 2006; Uppenkamp et al., 2006), and speech disorders (Ackermann & Riecker, 2004; Bonilha et al., 2006). Figure 1.8 shows an axial fMRI scan of the brain during a speech task. The task is to think of as many words as possible that use a specific letter, in a 24-second period. The task is repeated several times and modeled to bring out the contrast between the on and off states. Active regions are circled.

As with all instruments, MRI has several drawbacks. The first is the slow capture rate that results from summing the weak radio signals emitted by each proton. Thus, high-resolution images and DTI require long periods of immobility for a good image, and fMRI has a slow response time. Cine- and tMRI require summation of multiple, very precise repetitions for optimal images. A second drawback is the width of the section. Whereas CT sections are less than 1 mm wide, and ultrasound sections are less than 2 mm, MRI sections are usually 5–10 mm wide. A tomographic scan compresses a three-dimensional volume into two dimensions, which is like flattening a cylinder into a circle. For example, in a slice that is 5 mm wide, items that are actually 5 mm apart will appear to be in the same plane. Thus, in the transverse plane, the hyoid bone and epiglottis might appear in the same slice even though one is several millimeters below the other. Narrower widths in MRI sections are possible, but require longer exposure time. A third drawback is that many subjects, as many as 30 percent, experience claustrophobia and cannot tolerate the procedure. Fourth, metal clamps, tooth crowns, and steel implants quench the signal creating a diffuse dark spot surrounding the metal. Final drawbacks for MRI are that the subject must be lying supine or prone, which changes the location of gravity with respect to the oral structures and normal agonist–antagonist muscle relationships. Despite these nontrivial drawbacks, MRI's strengths make it an important instrument in speech production research.

1.2.3 Ultrasound Ultrasound produces an image by using the reflective properties of sound waves. A piezoelectric crystal stimulated by an electric current emits an ultra high-frequency sound wave. The crystal both emits a sound wave and receives the reflected echo. The sound wave travels through the soft tissue and reflects back when it reaches an interface with tissue of a different density, like bone, or when it reaches air. The best reflections are perpendicular to the beam (see Figure 1.9). In order to see a section of tissue rather than a single point, one needs an array transducer. In an array transducer, up to 128 crystals fire sequentially, imaging a section of tissue that is rectangular or wedge-shaped. The image size is proportional to the size of the transducer and frequency of the crystals, the wedge angle may be up to 140 degrees. The returning echoes are processed by an internal computer and displayed as a video image.

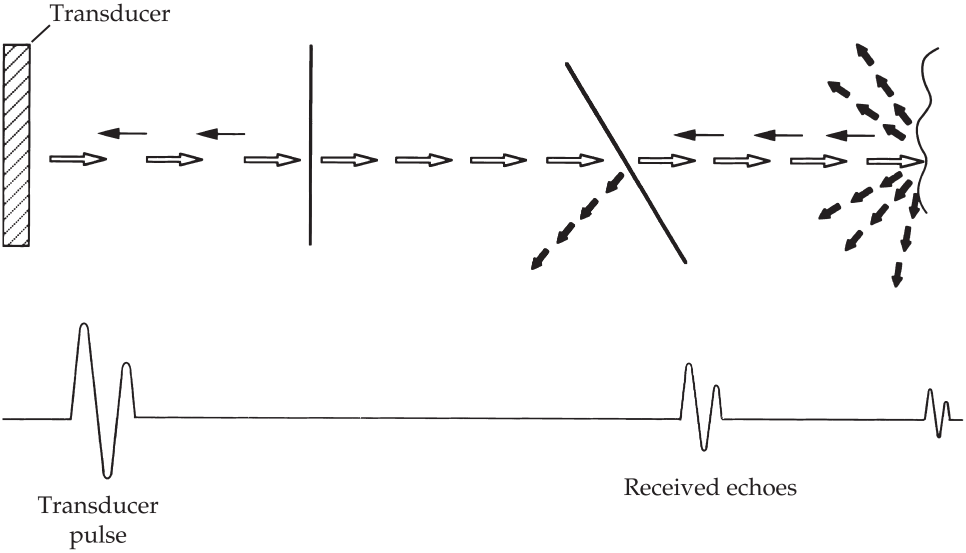


Figure 1.9 Schematic of ultrasound beam emitted from a transducer and reflected by surfaces of different angles. The best reflections are perpendicular to the beam. Soft tissue typically reflects and refracts sound in multiple directions.

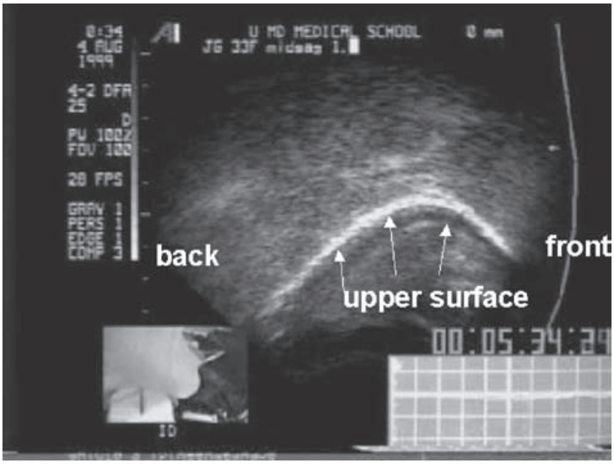


Figure 1.10 An ultrasound image of the sagittal (lengthwise) tongue. The white line is the upper surface of the tongue.

Figure 1.10 shows a sagittal image of the tongue in a 90-degree wedge-shaped scan. To create such an image, the transducer is placed below the chin and the beam passes upward through a 1.9 mm thick section of the tongue. When the sound reaches the air at the surface of the tongue, it reflects back creating a bright white line. The black area immediately below is the tongue body. The tongue surface is the lower edge of the white line. Interfaces within the tongue are also visible. Figure 1.11 depicts the tongue in coronal section. The tongue surface is thinner in cross-section and herein contains a small midsagittal depression. Measurement error on such images is at most 1 pixel (Stone et al., 1988). Although ultrasound is typically used to study tongue motion, it has been used on occasion to study other structures, such as the lateral pharyngeal wall (Parush & Ostry, 1993; Miller & Watkin, 1997), or vocal folds (Munhall & Ostry, 1985; Ueda et al., 1993).

The vocal folds may be imaged by placing the ultrasound transducer at the front of the neck, at the thyroid notch (Adam's apple), and pointing it directly back in the transverse plane. Glottal stops, tumors, and slow-moving behaviors can be seen this way. However, the vocal folds have a very fast vibration rate, at least 80 Hz, and usually much more. The sampling rate of the fastest ultrasound machines is about 90 Hz. Therefore, vibration may be seen during phonation, but it is undersampled, and in individual frames cannot be measured. Other instruments, such as Electroglottography, are more accurate methods for measuring vocal fold vibration.

Ultrasound presents specific challenges when measuring the tongue. The first challenge is that up to 1 cm of the tongue tip may not be captured in the image, because the ultrasound beam is reflected at the floor of the mouth and the sound wave doesn't enter the tongue tip. The tip may be imaged, however, if there is sufficient saliva in the mouth, if the tongue is resting against the floor, or if the

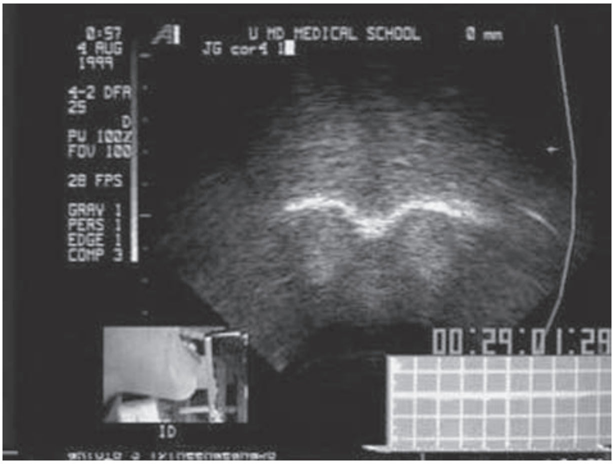


Figure 1.11 An ultrasound image of the coronal (crosswise) tongue. The upper surface has a midline groove.

transducer is posterior and angled forward (Stone, 2005). The second limitation is the inability to see beyond a tissue–air or tissue–bone interface. Since the tissue–air interface at the tongue's surface reflects the sound wave, the structures on the far side of the vocal tract, such as the palate and pharyngeal wall, cannot be imaged. Similarly, when ultrasound reaches a bone, the curved shape refracts the sound wave creating an acoustic shadow or dark area. Thus, the jaw and hyoid bones appear as shadows and their exact position cannot be reliably measured.

Despite these limitations, a large number of studies successfully use real-time ultrasound to study tongue movements. Normal speech production and exploration of tongue surface features are the most common applications, (cf. Davidson, 2006; Slud et al., 2002; Chiang et al., 2003). Ultrasound has also been the basis of tongue surface models in the sagittal plane (Green & Wang, 2003), coronal plane (Slud et al., 2002), and in 3D (Watkin & Rubin, 1989; Yang Stone & Lundberg, 1996; & Stone, 2002; Bressmann et al., 2005b). Because ultrasound is well tolerated by subjects it is well suited to the study of disorders (Bressmann et al., 2005a; Schmelzeisen et al., 1996) and to studies of children (Ueda et al., 1993). Ultrasound is also an excellent tool to study swallowing because it is noninvasive and does not affect the swallow (Miller & Watkin, 1997; Chi-Fishman et al., 1998; Watkin, 1999; Peng et al., 2000; Soder & Miller, 2002). Moreover, it is now possible to align tongue position with the hard palate, if the transducer and head are held still (Epstein & Stone, 2005). This alignment allows better interpretation of the ultrasound data. Finally, applications of ultrasound to linguistics are increasing because portable machines allow ultrasound to be used in fieldwork (cf. Gick, 2002a). Ultrasound provides large quantities of time–motion data with a single repetition, thin slices, and a noninvasive method. Many machines are now digital and can collect 90 or more scans per second, which is fast enough to measure most tongue motions, though not fast enough to measure vocal fold vibration. A second advantage is that ultrasound involves no known biological hazards, since the transduction process involves only sound waves.

Although most ultrasound machines scan a single tissue slice, some transducers scan multiple slices and reconstruct 3D volumes. Visualizing 3D tongue movements facilitates understanding of tongue surface motion. This is important because the tongue is anisotropic from front to back and medial to lateral. Therefore no single slice fully represents its motion. Some commercial ultrasound machines have 4D transducers that create 3D ultrasound volume sequences. These transducers contain a standard 2D array transducer, which sequentially scans a series of tissue slices by mechanically stepping through a series of angles, scanning a slice at each step. If the stepping is slow, the resulting scans and steps are combined into a static 3D volume. A 3D movie is created when the scanning, stepping, and combining are done very quickly, at multiple times per second. This technology is very promising even though at present the scan rates are fairly slow. To capture the entire tongue, a large volume must be imaged, and the scan rate is usually limited to about 6 Hz for a volume large enough to capture the entire tongue. Alternatively, multiple 2D ultrasound scans can be collected with a faster frame rate, and combined into a 3D tongue surface reconstruction as seen in Figure 1.12 (Stone & Lundberg, 1996; Bressmann et al., 2005b).

Ultrasound, like the other imaging techniques, captures the inaccessible parts of the tract (e.g., the pharynx), and measures planes rather than tissue points to provide comprehensive and detailed information about vocal tract structures.

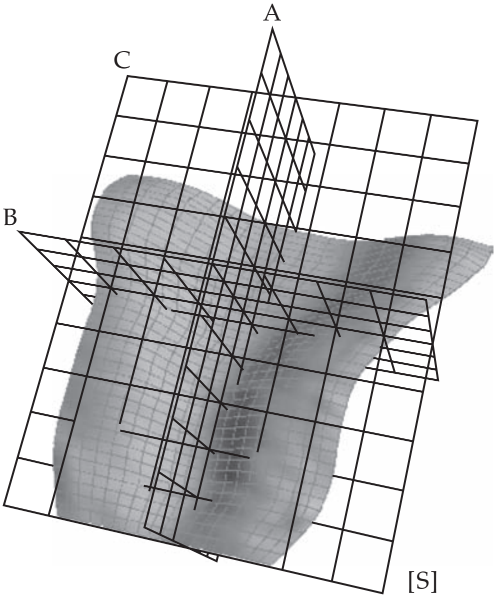


Figure 1.12 A 3D reconstruction of the tongue surface from multiple ultrasound slices.

Although the sampling rates of imaging techniques tend to be slower than point-tracking systems, imaging systems are now widely available in hospitals and laboratories. Moreover, portable machines, used in fieldwork, provide expanded and varied linguistic datasets.

2 Point-Tracking Measurements of the Vocal Tract

Point-tracking systems have different strengths from imaging techniques; they measure individual fleshpoints by affixing pellets to the articulators and tracking their movement over time. Typically, multiple articulators are measured at the same time, and tracking speed is fast, so that inter-articulator timing measures are quite good. Well-known tracking systems can be external to the oral cavity, such as Optotrak or Vicon, or both external and internal, such as the articulometer and the X-ray microbeam. External tracking systems can directly measure markers on the face and lips using video or light emitting diode (LED) tracking. Tracking systems operating within the vocal tract use pellets or receivers tracked by magnetic fields or X-rays.

2.1 Electromagnetic Articulometer (EMA)

Several names refer to similar point-tracking devices; these are Electromagnetic Midsagittal Articulometer (EMMA), Electromagnetic Articulometer (EMA), and Articulograph. Several such instruments, using roughly similar principles are currently in use (Perkell et al., 1992; Van Lieshout et al., 1994; Zierdt et al., 1999). The instruments track tissue-point movement in and around the oral cavity by measuring the movement of small receiver coils through alternating magnetic fields.

In the 2D EMA systems, three transmitter coils form an equilateral triangle in the midsagittal plane. They are suspended around the subject's head using a clear plastic assembly. Each transmitter coil is driven at a different sinusoidal frequency to generate alternating magnetic fields of different frequencies. Small insulated receiver coils are attached with adhesive to oral and facial structures of interest at midline. The alternating magnetic fields induce an alternating signal in the receiver coils. The voltage of this signal is inversely related to the distance between the transmitter and the receiver coil. A computer algorithm uses these distances to calculate the location of the receiver coil as it moves in x–y space over time. The best resolution in the field space is found in the center, i.e., in the oropharyngeal region, where measurement resolution is calculated at less than 1 mm.

EMA's strength is its ability to measure multiple articulators simultaneously, at a rapid sampling rate, making it popular in the fields of linguistics, speech motor control, swallowing, and speech pathology. In linguistics, studies have focused on articulation and inter-articulator programming, among other topics. In English, Byrd et al. (2005) examined effects of sentence position on consonant articulation. In German, Fuchs et al. (2006) studied differences in the control of

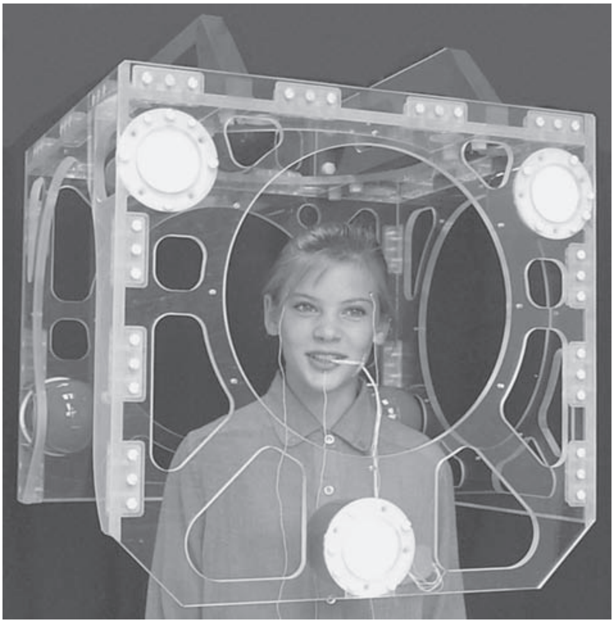


Figure 1.13 3D Electromagnetic articulograph tracks markers in the mouth using a magnetic field. (Photo courtesy of Carstens Medizinelektronik, Inc.)

fricatives and stops, and Kuhnert and Hoole (2004) found similar patterns of reduction in English and German. EMA studies have also effectively studied motor control (cf. Kaburagi & Honda, 1996; Van Lieshout & Moussa, 2000; Perkell & Zandipour, 2002; McClean & Tasko, 2003), swallowing (Nicosia et al., 2000; Steele & Van Lieshout, 2004), and speech disorders (Katz & Verma, 1994; Goozée, et al., 2000; Bose et al., 2003).

A 3D EMA instrument (the Articulograph AG500) is commercially available from Carstens, Inc., Munich, Germany (Zierdt et al., 2000; Hoole et al., 2003). In this 3D system, a clear acrylic cube surrounds the speaker's head (see Figure 1.13). The cube contains six transmitters, in a spherical configuration, each of which produces a magnetic field at a different frequency. Within the oral cavity sensors are affixed to the articulators of interest. Each sensor produces an alternating current that varies with its relative distance from the six transmitters. Sensor location is represented by five parameters: three positions (x, y, z), and two angles (azimuth and elevation). The angular parameters mean that tilt of the sensor no longer introduces artifact; instead tilt is incorporated into the two measurement angles. The position of each sensor is subtracted from a sensor affixed to the head. The 3D machine has a spatial resolution of 1 mm and an angular accuracy of one degree.

Two limitations of point-tracking systems are that sensors need to be affixed to the structures, potentially interfering with speech or swallowing. In addition, only points are measured, so the behavior of the entire articulator is largely inferred. This is most problematic for the soft tissue structures like the lips, tongue, and velum, whose movements are more fluid than rigid.

The biggest advantages of point-tracking systems are the rapid tracking rate, and the ability to track multiple articulators simultaneously. Because of these two features, interaction among the articulators can be measured, and questions about inter-articulator timing and programming can be answered. In addition, the 3D articulograph is one of the few instruments that does not impede head motion, allowing natural movement. Although there has been some concern over possible health consequences from exposure to magnetic fields, the articulometer poses minimal biological hazards as it uses short exposure times and low field strengths (cf. Hasegawa-Johnson, 1998). Nonetheless, its use is not recommended with pregnant women.

2.2 *X-ray Microbeam*

The X-ray Microbeam tracks tissue-points on the surface of the articulators, resulting in data similar to 2D EMA. The method is quite different, however, as the X-ray Microbeam uses a very thin X-ray beam to track the motion of small gold pellets. The pellets are affixed to one or more articulators using dental adhesive. The beam is 0.4 mm thick and the pellets are 2–3 mm in diameter. Gold pellets are used because gold is an inert metal and because, as the X-ray dosage is very small, only a very dense metal can be detected. The system was designed to reduce radiation dosage to well below that of a dental X-ray, to avoid radiosensitive areas, such as the eyes, and to reduce secondary photon scatter. The X-ray beam focuses primarily on the pellets so that the surrounding tissue receives only minimal radiation.

The direction of the beam is computer controlled. The X-ray beam originates at one side of the subject, passes through the subject's head, and is detected by a scintillation counter on the far side. Up to 1,000 pellet positions are sampled per second. Thus, if 10 pellets are used, they can each be sampled 100 times per second. Differential sampling rates are also possible. The pellets are sampled initially in rest position to determine baseline x–y displacement. A computer algorithm causes the beam to scan the area in which the pellet is predicted to move. The prediction is based on the pellet's previous displacement, velocity, and acceleration. Each pellet is scanned in order, and position accuracy of the beam is 62 microns.

Many studies have collected X-ray Microbeam data, or used data available in an archival database from the University of Wisconsin, Madison (Westbury, 1994). X-ray Microbeam data have been used to study swallowing (cf. Tasko et al., 2002), articulator interaction (Westbury et al., 2002), motor control (Tasko & Westbury, 2002, 2004), and speech disorders (Weismer et al., 2003).

The strengths of the X-ray Microbeam are rapid sampling rate and accuracy of tracking, which make this an excellent system for examining timing related coarticulatory effects, kinematic parameters such as velocity and acceleration, and the intercoordination of the articulators. In addition, the technique is unobtrusive and the low radiation dosage allows for reasonably large data sets to be collected on each subject.

In addition to the limitations found in EMA, the X-ray Microbeam is further limited by being unique, and therefore difficult to access for most investigators and their subjects. An additional disadvantage is that although the radiation dosage is low, an X-ray based system does contains some biological hazard. In addition, only two-dimensional data are collected, and movements off-plane are lost or induce error. Finally, a problem common to all pellet-tracking devices is the near impossibility of affixing pellets to the pharynx, velum, and pharyngeal tongue due to the gag reflex and poor accessibility, thus limiting its use with many subjects. The X-ray Microbeam archival database is available at www.medsch.wisc.edu/~milenkvc/pdf/ubdbman.pdf.

2.3 Optotrak

It is easier to track points on the face than those within the oral cavity, because the markers maintain visual contact with the sensors at all times. Point-tracking systems, such as Optotrak (Northern Digital, Waterloo, ON, Canada) and Vicon (Vicon Motion Systems Inc., Lake Forest, CA, USA) use optical measurements of LED markers in three-dimensional (*x*–*y*–*z*) space. Thus, they measure left-to-right movement as well as anterior-to-posterior and superior-to-inferior. The system consists of (1) markers placed on surface structures, (2) sensors that track their position, and (3) a unit that controls the timing of marker emissions and sensor processing. The markers are tracked at high sampling frequencies; for example, Optotrack samples at 100 kHz divided by the number of markers. Embedded in the center of each Optotrak marker is a small semiconductor chip that emits an infrared signal. A strip of three camera-like sensors tracks the movement of the infrared emissions of the markers. Spatial resolution is also excellent, positional accuracy is 0.1 mm on the *x*- and *y*-axes and 1.5 mm on the *z*-axis at 2.25 meters. The sensors are pre-aligned on a single unit so that they measure the position of each marker in 3D space.

These external tracking instruments track lip, jaw, and face motion in 3D and are well-suited to examining the complex relationships between them, such as the cues provided by facial motions in hyperarticulated versus normal speech (Maeda & Toda, 2003). The instruments are noninvasive and are convenient to use. Their disadvantage is that they are limited to external use, unlike the X-ray Microbeam and the articulometer, because the sensors must maintain "visual" contact with the markers. Therefore, Optotrak and Vicon cannot be used inside the mouth and so do not reveal structures within the vocal tract. Two more limitations are common to all point-tracking systems. First, only tissue-points are tracked, not the entire structure. For rigid structures such as the jaw, this is not a problem and the entire structure can be reconstructed. However, the flexible deformation of soft tissue structures, such as the tongue, lips, and velum, is incompletely measured and represented. It is also possible that the markers or their attached wires may interfere, at least minimally, with truly natural speech.

3 Measurement of Tongue–Palate Interaction

3.1 *Tongue–palate contact*

Electropalatography (EPG) directly measures tongue–palate contact in real time during the motions of speech and swallowing. EPG data are quite different from both point-tracking and imaged data. Small metal electrodes (about 0.5–1.5 mm diameter) are embedded into an acrylic pseudopalate and are activated when contacted by the tongue; this contact completes an electric circuit in the body. The electrodes measure on/off contact, and though the activation threshold can be changed, subtle changes in pressure are not recorded. Thin wires (approx. 42 gauge) attached to each electrode, exit the palate by winding behind the back molars bilaterally, and running forward along the outer surface of the teeth and out through the corners of the lips. A ground electrode completes the circuit. The pseudopalate is electrically isolated from the computer that drives it, and the wires are driven by an AC current of less than five microamps. When the tongue contacts an electrode, the circuit is completed and the contact registered. EPG is high-dimensional as it dynamically records multiple contact points between two structures (tongue and palate) from the entire palate surface. A recent system, WinEPG (Articulate Instruments, Edinburgh), uses 62 electrodes embedded into a very thin (0.5 mm) acrylic pseudo-palate, which is molded to fit the palate of the speaker (Figure 1.14). The electrodes are sampled at 100−200 Hz (Wrench, 2007).

Data from EPG provide a unique perspective on the interaction of the tongue and palate, and their relationship to vocal tract shape in the palatal area (Scobbie et al., 2004). EPG was used initially, over three decades ago, to study linguistics (Hardcastle, 1972) and speech disorders (Fletcher et al., 1975). Since then it has been used as a clinical tool for on-site and remote speech therapy (Gibbon et al., 1998), to study cleft palate (Gibbon et al., 2001), cerebral palsy (Gibbon & Wood, 2003), Parkinson's Disease (McAuliffe et al., 2006), open bite (Suzuki et al., 1981), Traumatic Brain Injury (Hartelius et al., 2005), glossectomy (Suzuki, 1989), and apraxic speech (Hardcastle & Edwards, 1992).

EPG has also been used to answer linguistic questions, often with EMA, as the two methods provide complementary information about tongue motion and tongue−palate contact. One application that is well established across languages is the documentation of tongue−palate contact patterns for alveolar sounds. These have been studied extensively in such languages as Japanese (Miyawaki et al., 1974), Norwegian (Moen et al., 2004), Italian (Farnetani & Recasens, 1993), Catalan (Recasens & Pallarès, 2001), German (Kohler, 1976), and English (Ellis & Hardcastle, 2002). Velar sounds are less studied since the traditional pseudo-palate usually ends before the soft palate (Hardcastle, 1985). However, some velar patterns have been studied (Suzuki & Michi, 1986) as has the behavior of the tongue during labials (Gibbon et al., 2007). Coarticulation is observed well in EPG because of its high spatial resolution and fast sampling rates. Studies have examined locus equations (Tabain, 1998), assimilation in consonant sequences (Ellis & Hardcastle, 2002), and the interference of prosody in vowel-to-vowel coarticulation (Fletcher, 2004).

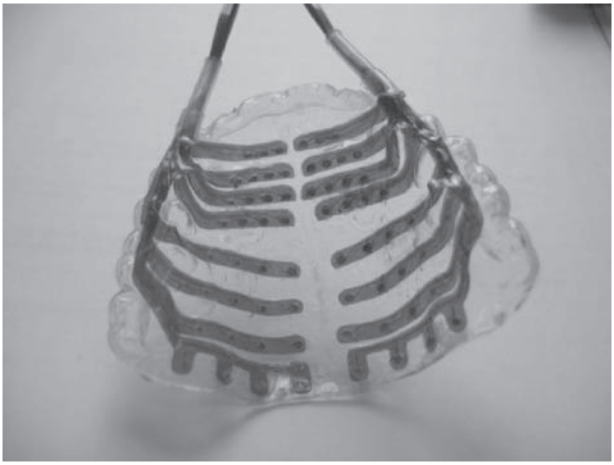


Figure 1.14 A pseudopalate containing 64 electrodes that record tongue–palate contact. (Photo courtesy of Alan Wrench)

The drawback of EPG is that data are collected only when the tongue touches the palate. Thus information is lost when the jaw lowers the tongue away from the domain of the palate, as occurs during mid and low vowels. However, during continuous speech the tongue is in fairly continuous contact with the palate at one location or another. EPG thus contributes greatly to the study of speech, especially the study of lingual consonants, constriction shapes and sizes, increasing its value in studies of language and disorders.

3.2 *Tongue–palate pressure*

Pressure palatography (PPG) measures moment-to-moment changes in tongue–palate pressure. The transducers embedded in the acrylic palate measure pressure instead of contact. Several methods have been used to transduce this pressure. The earliest and most commonly used transducers were strain gauges. A strain gauge is a transducer that is mounted on an object, such as a beam or diaphragm, that is capable of deformation. As the object deforms the gauge also deforms. As the gauge deforms, or strains, the strain produces a change in resistance, which is converted into a change in voltage. The change in voltage is an analog waveform that is amplified by a wheatstone bridge and recorded in analog or digital form.

Miniature strain gauges were first used to measure tongue–palate pressure over 30 years ago. McGlone and Proffit (1972) inserted two strain gauges into an artificial palate. Changes in pressure due to tongue–palate contact were measured. Despite their success in delineating pressure differences in /t/ versus /d/, tongue–palate pressure studies were not published again for 20 years. In recent years, strain gauge pressure sensors have been successfully used for the assessment of the tongue in oral functions, such as deglutition (Chiba et al., 2003; Ono et al., 2004) and articulation (Tsuga et al., 2003). PPG allows the measurement of tongue force changes over time, including time between the onset of linguopalatal contact to the time of maximum pressure, the maximum pressure and the position–force ratio. Pressure transducers have been used to determine hemispheric dominance (Shinagawa et al., 2003), and to record tongue behavior in patients who have had glossectomy (Yoshioka et al., 2004) and rapid palatal expansion (Kucukkeles & Ceylanoglu, 2003).

In addition to strain gauges other methods of measuring pressure changes have been used experimentally. Wakumoto and colleagues (1998) devised a pressure recording "sheet" composed of two polyester layers separated by regions of air and pressure-sensing cells. A pressure-sensing cell has a layer of "pressure-sensitive ink" embedded between two electrodes that measure its resistance to pressure. A cell is 3 mm diameter with a sensitivity of 173–2734 Pa (Wakumoto et al., 1998). The sheets of electrodes are 0.1 mm thick and are glued to an acrylic pseudopalate of 0.5 mm. This work has used up to 16 sensors per palate and has sensitivity enough to distinguish the lingual pressures of /t/ and /d/. Murdoch et al. (2004) developed pressure sensors that contain a magnet and a Hall effect transducer (HET). HETs produce a voltage when in the presence of a magnetic field that is proportional to the magnetic field. When the tongue touches one of these sensors a cantilever beam is deflected, which changes the distance between the magnet and the HET. This varying distance results in proportionally varied voltage output. One HET sensor is 4.4 mm x 6.2 mm, and five were embedded in the prototype pseudopalate. A single subject trial was inconclusive because, as with all the PPG transducers, the limited number of sensors reduces sensitivity to subtle tongue-pressure changes. Continued work is expected to improve these techniques.

Instrumental studies of physiology are challenging and, no single instrument provides total vocal tract information. However, the importance of these instruments lies in the critical role they play in cutting-edge studies of speech motor control, speech disorders, phonetics, phonology, and even speech processing. The data acquired by these instruments is in great demand for two reasons. First, the data keep increasing our knowledge of speech physiology, speech disorders, and coarticulation strategies. They reveal how articulation and rhythm are organized and controlled, what aspects of speech are common among languages, and the nature of differences between speakers, languages, and disorders. Second, the data are critical in testing current theories and models. The facts and models of an era are supported or disproved by the data from new instrumentation. Forty years ago, the sound spectrograph destroyed the notion that speech sounds were independent, concatenated segments. Today, measures of speech physiology are challenging our ideas of how vocal tract constrictions are achieved and what features of them are acoustically salient. Similarly, we wish to find out what components of the articulatory gesture are carried in the speech wave and decoded by the listener. A better understanding of how the vocal tract produces speech can also improve synthetic speech and provide strategies for machine recognition of speech. These and other issues, which are in the forefront of speech research, can be addressed and perhaps resolved using the instruments described in this chapter. The exciting leaps in knowledge provided by physiological data far outweigh the difficulties associated with the use of these instruments.

REFERENCES

Ackermann, H. & Riecker, A. (2004) The contribution of the insula to motor aspects of speech production: A review and a hypothesis. *Brain and Language*, 89, 320–8.

Axel, L. & Dougherty, L. (1989) Heart wall motion: Improved method of spatial modulation of magnetization for MR imaging. *Radiology*, 172, 349–50.

Bonilha, L., Moser, D., Rorden, C., Baylis, G. C., & Fridriksson, J. (2006) Speech apraxia without oral apraxia: Can normal brain function explain the physiopathology? *Neuroreport*, 17, 1027–31.

Bose, A., Lieshout, P. H. H. M. van, & Square, P. A. (2003) Speech coordination in individuals with aphasia and normal speakers. *Brain and Language*, 87, 158–9.

Bressmann, T., Heng, C.-L., & Irish, J. C. (2005a) Applications of 2D and 3D ultrasound imaging in speech-language pathology. *Journal of Speech-Language Pathology and Audiology*, 29, 158–68.

Bressmann, T., Uyn, C., & Irish, J. C. (2005b) Analyzing normal and partial glossectomee tongues using ultrasound. *Clinical Linguistics and Phonetics*, 19, 35–52.

Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005) Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America*, 118, 3860–73.

Chiang, Y. C., Lee, F. P., Peng, C. L., & Lin, C. T. (2003) Measurement of tongue movement during vowels production with computer-assisted B-mode and M-mode ultrasonography. *Otolaryngology-Head and Neck Surgery*, 128, 805–14.

Chiba, Y., Motoyoshi, M., & Namura, S. (2003) Tongue pressure on loop of transpalatal arch during deglutition. *American Journal of Orthodontics and Dentofacial Orthopedics*, 123, 29–34.

Chi-Fishman, G., Stone, M., & McCall, Gerald N. (1998) Lingual action in normal sequential swallowing. *Journal of Speech Language and Hearing Research*, 41, 771–85.

Davidson, L. (2006) Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Journal of the Acoustical Society of America*, 120, 407–15.

Djamshidpey, H., Waneland, I., Krull, D., & Lindblom, B. (1998) X-ray analyses of speech: Methodological aspects. *Proceedings of Fonetik '98, 11th Swedish Phonetics Conference*, Stockholm, pp. 168–71.

Ellis, L. & Hardcastle, W. J. (2002) Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data. *Journal of Phonetics*, 30, 373–96.

Engwall, O. (2003) Combining MRI, EMA & EPG measurements in a three-dimensional tongue model. *Speech Communication*, 41, 303–29.

Epstein, M. A. & Stone, M. (2005) The tongue stops here: Ultrasound imaging of the palate. *Journal of the Acoustical Society of America*, 118, 2128–31.

Fant, G. (1965) Formants and cavities. *Proceedings of the 5th International Congress of Phonetic Sciences* (pp. 120–41). Munster, 1964. Basel/New York: S. Karger.

Farnetani, E. & Recasens, D. (1993) Anticipatory consonant-to-vowel coarticulation in the production of VCV sequences in Italian. *Language and Speech*, 36, 279–302.

Fletcher, J. (2004) An EMA/EPG study of vowel-to-vowel articulation across velars in Southern British English. *Clinical Linguistics and Phonetics*, 18, 577–92.

Fletcher, S. G., McCutcheon, M. J., & Wolf, M. B. (1975) Dynamic palatometry. *Journal of Speech and Hearing Research*, 18, 812–19.

Fuchs, S., Perrier, P., Geng, C., & Mooshammer, C. (2006) What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents. In J. Harrington & M. Tabain (eds.), *Speech production: Models, Phonetic Processes and Techniques*. New York: Psychology Press, pp. 149–64.

Gibbon, F., Crampin, L., Hardcastle, W. J. et al. (1998) CleftNet Scotland: A network for the treatment of cleft palate speech using EPG. *International Journal of Language and Communication Disorders*, 33, supplement, 44–9.

Gibbon, F., Hardcastle, W. J., Crampin, L., Reynolds, B., Razzell, R., & Wilson, J. (2001) Visual feedback therapy using electropalatography (EPG) for articulation disorders associated with cleft palate. *Asia Pacific Journal of Speech, Language and Hearing*, 6, 53–8.

Gibbon, F., Lee, A., & Yuen, I. (2007) Tongue palate contact during bilabials in normal speech. *Cleft Palate-Craniofacial Journal*, 44, 87–91.

Gibbon, F. & Wood, S. (2003) Using electropalatography (EPG) to diagnose and treat articulation disorders associated with mild cerebral palsy: A case study. *Journal of Clinical Linguistics and Phonetics*, 17, 365–74.

Gick, B. (2002a) The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, 32, 113–22.

Gick, B. (2002b) An X-ray investigation of pharyngeal constriction in American English schwa. *Phonetica*, 59, 38–48.

Gick, B., Wilson, I., Koch, K., & Cook, C. (2004) Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, 61, 220–33.

Gilbert, R. J. & Napadow, V. J. (2005) Three-dimensional muscular architecture of the human tongue determined in vivo with diffusion tensor magnetic resonance imaging. *Dysphagia*, 20, 1–7.

Goozée, J., Murdoch, B. E., Theodoros, D. G., & Stokes, P. D. (2000) Kinematic analysis of tongue movements in dysarthria following traumatic brain injury using electromagnetic articulography. *Brain Injury*, 14, 153–74.

Gracco, V. L., Tremblay, P., & Pike, B. (2005) Imaging speech production using fMRI. *Neuroimage*, 26, 294–301.

Green, J. R. & Wang, Y.-T. (2003) Tongue surface movement patterns during speech and swallowing. *Journal of the Acoustical Society of America*, 113, 2820–33.

Hardcastle, W. J. (1972) The use of electropalatography in phonetic research. *Phonetica*, 25, 197–215.

Hardcastle, W. J. (1985) Some phonetic and syntactic constraints on lingual co-articulation during /kl/ sequences. *Speech Communication*, 4, 247–63.

Hardcastle, W. J. & Edwards, S. (1992) EPG-based descriptions of apraxic speech errors. In R. Kent (ed.), *Intelligibility in Speech Disorders: Theory, Measurement and Management*. Philadelphia: John Benjamins, 287–328.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977) Factor analysis of tongue shapes. *Journal of the Acoustical Society of America*, 62, 693–707.

Hartelius, L., Theodoros, D., & Murdoch, B. (2005) Use of electropalatography in the treatment of disordered articulation following traumatic brain injury: A case study. *Journal of Medical Speech–Language Pathology*, 13, 189–204.

Hasegawa-Johnson, M. (1998) Electromagnetic exposure safety of the Carstens Articulograph AG100. *Journal of the Acoustical Society of America*, 104, 2529–32.

Hasegawa-Johnson, M., Pizza, S., Alwan, A., Cha, J. S., & Haker, K. (2003) Vowel category dependence of the relationship between palate height, tongue height, and oral area. *Journal of Speech, Language, and Hearing Research*, 46, 738–53.

Hashimoto, K. & Sasaki, K. (1982) On the relationship between the shape and position of the tongue for vowels. *Journal of Phonetics*, 10, 291–9.

Hoole, P., Zierdt, A., & Geng, C. (2003) Beyond 2D in articulatory data acquisition and analysis. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, Barcelona, 265–8.

Iskarous, K. (2005) Patterns of tongue movement. *Phonetica*, 33, 363–82.

Kaburagi, T. & Honda, M. (1996) A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes. *Journal of the Acoustical Society of America*, 99, 3154–70.

Katz, W. F. & Verma, S. (1994) Kinematic evidence for compensatory articulation by normal and nonfluent aphasic speakers. *Brain and Language*, 47, 357–60.

Kent, R. D. (1972) Some considerations in the cinefluorographic analysis of tongue movements during speech. *Phonetica*, 26, 16–32.

Kent, R. D. & Moll, K. L. (1972) Cinefluorographic analyses of selected lingual consonants. *Journal of Speech and Hearing Research*, 15, 453–73.

Kent, R. D. & Netsell, R. (1971) Effects of stress contrasts on certain articulatory parameters. *Phonetica*, 24, 23–44.

Kohler, K. (1976) The instability of word-final alveolar plosives in German: An Electropalatographic investigation. *Phonetica*, 33, 1–30.

Kokawa, T., Saigusa, H., Aino, I., et al. (2006) Physiological studies of retrusive movements of the human tongue. *Journal of Voice*, 20, 414–22. Epub November 21, 2005.

Kucukkeles, N. & Ceylanoglu, C. (2003) Changes in lip, cheek, and tongue pressures after rapid maxillary expansion using a diaphragm pressure transducer. *Angle Orthodontist*, 73, 662–8.

Kühnert, B. & Hoole, P. (2004) Speaker-specific kinematic properties of alveolar reductions in English and German. *Clinical Linguistics and Phonetics*, 18, 559–75.

Lam, P., Au-Yeung, K. M., Cheng, P. W., et al. (2004) Correlating MRI and histologic tumor thickness in the assessment of oral tongue cancer. *American Journal of Roentgenology*, 182, 803–8.

Lell, M. M., Greess, H., Hothorn, T., Janka, R., Bautz, W. A., & Baum, U. (2004) Mutiplanar functional imaging of the larynx and hypopharynx with multislice spiral CT. *European Journal of Radiology*, 14, 2198–205.

Lenz, M., Greess, H., Baum, U., Dobritz, M., & Kersting-Sommerhoff, B. (2000) Oropharynx, oral cavity, floor of the mouth: CT and MRI. *European Journal of Radiology*, 33, 203–15.

Lieshout, P. H. H. M. van, Alfonso, P. J., Hulstijn, W., & Peters, H. F. M. (1994) Electromagnetic midsagittal articulography (EMMA). In F. J. Maarse, A. E. Akkerman, A. N. Brand, L. J. M. Mulder, & M. J. Van der Stelt (eds.), *Applications, Methods and Instrumentation* (pp. 62–76). Lisse, The Netherlands: Swets & Zeitlinger.

Lieshout, P. H. H. M. van & Moussa, W. (2000). The assessment of speech motor behaviors using electromagnetic articulography. *The Phonetician*, 81, 9–22.

Lindblom, B., Sussman, H., Modarresi, G., & Burlingame, E. (2002) The trough effect: Implications for Speech Motor Programming. *Phonetica*, 59, 245–62. *Using data collected in Stockholm by* Branderud, P, Lundburg, H, Lander, J.

Mády, K., Sader, R., Zimmermann, A., et al. (2002) Assessment of consonant articulation in glossectomee speech by dynamic MRI. *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP '2002*, Denver, USA. 961–4.

Maeda, S. (1990) Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. L. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 131–49). Dordrecht: Kluwer.

Maeda, S. & Toda, M. (2003) Mechanical properties of lip movements: How to characterize different speaking styles? *The 15th International Congress of Phonetic Sciences (ICPhS 2003)*, Barcelona, 189–92.

Magen, H. S., Kang, A. M., Tiede, M. K., & Whalen, D. H. (2003) Posterior pharyngeal wall position in the production of speech. *Journal of Speech, Language, and Hearing Research*, 46, 241–51.

McAuliffe, M. J., Ward, E. C., & Murdoch, B. E. (2006) Speech production in Parkinson's disease, I: An electropalatographic investigation of tongue–palate contact patterns. *Clinical Linguistics and Phonetics*, 20, 1–18.

McClean, M. D. & Tasko, S. M. (2003) Association of orofacial muscle activity and movement during changes in speech rate and intensity. *Journal of Speech, Language, and Hearing Research*, 46, 1387–400.

McGlone, R. & Proffit, W. (1972) Correlation between functional lingual pressure and oral cavity size. *Cleft Palate Journal*, 9, 229–35.

McGowan, R. S. (2006) Perception of synthetic vowel exemplars of 4-year-old children and estimation of their corresponding vocal tract shapes. *Journal of the Acoustical Society of America*, 120, 2850–8.

Mermelstein, P. (1973) Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53, 1070–82.

Miller, J. L. & Watkin, K. L. (1997) Lateral pharyngeal wall motion during swallowing using real time ultrasound. *Dysphagia*, 12, 125–32.

Miyawaki, K., Kiritani, S., Tatsumi, I. F., & Fujimura, O. (1974) Palatographic observation of VCV articulations in Japanese. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 8, 51–7.

Moen, I., Simonsen, H. G., & Lindstad, A. M. (2004) An electronic database of Norwegian speech sounds: Clinical aspects. *Journal of Multilingual Communication Disorders*, 2, 43–9.

Munhall, K. G. & Ostry, D. J. (1985) Ultrasonic measurement of laryngeal kinematics. In I. R. Titze & R. C. Scherer (eds.), *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control* (pp. 145–62), Denver: Denver Center for the Performing Arts.

Munhall, K., Vatikiotis-Bateson, E., & Tohkura, Y. (1994a) X-ray film database for speech research. *Journal of the Acoustical Society of America*, 98, 1222–4.

Munhall, K., Vatikiotis-Bateson, E., & Tohkura, Y. (1994b). X-ray film database for speech research. Technical report TR-H-116, ATR Human Information Processing Laboratories, Kyoto.

Murdoch, B. E., Goozée, J. V., Veidt, M., Scott, D. H., & Meyers, I. A. (2004) Introducing the pressure-sensing palatograp: The next frontier in electropalatography. *Clinical Linguistics and Phonetics*, 18, 433–45.

Napadow, V. J., Chen, Q., Wedeen, V. J., & Gilbert, R. J. (1999a) Biomechanical basis for lingual muscular deformation during swallowing. *American Journal of Physiology*, 277, G695–G701.

Napadow, V. J., Chen, Q., Wedeen, V. J., & Gilbert, R. J. (1999b) Intramural mechanics of the human tongue in association with physiological deformations. *Journal of Biomechanics*, 322, 1–12.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., & Byrd, D. (2004) An approach to real-time magnetic resonance imaging for speech prodcution. *Journal of the Acoustical Society of America*, 115, 1771–6.

NessAiver, M., Stone, M., Parthasarathy, V., Kahana, Y., Kots, A., & Paritsky, A. (2006) Recording high quality speech during tagged Cine MRI studies using a fiber optic microphone. *Journal of Magnetic Resonance Imaging*, 23, 92–7.

Nicosia, M. A., Hind, J. A., Roecker, E. B., et al. (2000) Age effects on the temporal evolution of isometric and swallowing pressure. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 55, M634–M640.

Niitsu, M., Kumada, M., Campeau, N. G., Niimi, S., Riederer, S. J., & Itai, Y. (1994) Tongue displacement: Visualization with rapid tagged magnetization-prepared MR imaging, *Radiology*, 191, 578–80.

Ono, T., Hori, K., & Nokubi, T. (2004) Pattern of tongue pressure on hard palate during swallowing. *Dysphagia*, 19, 259–64.

Parthasarathy, V., Prince, J. L., Stone, M., Murano, E., & NessAiver, M. (2007) Measuring tongue motion from tagged Cine-MRI using harmonic phase (HARP) processing. *Journal of the Acoustical Society of America*, 121, 1, 491–504.

Parush, A. & Ostry, D. J. (1993) Lower pharyngeal wall coarticulation in VCV syllables. *Journal of the Acoustical Society of America*, 94, 715–22.

Peng, C. L., Jost-Brinkmann, P. G., Miethke, R. R., & Lin, C. T. (2000) Ultrasonographic measurement of tongue movement during swallowing. *Journal of Ultrasound in Medicine*, 19, 15–20.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992) Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements, *Journal of the Acoustical Society of America*, 92, 3078–96.

Perkell, J. & Zandipour, M. (2002) Economy of effort in different speaking conditions, II: Kinematic performance spaces for cyclical and speech movements. *Journal of the Acoustical Society of America*, 112, 1642–51.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006) Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 7865–70.

Recasens, D. & Pallarès, M. D. (2001) Coarticulation, blending and assimilation in Catalan consonant clusters. *Journal of Phonetics*, 29, 273–301.

Schmelzeisen, R., Ptok, M., Schonweiler, R., Hacki, T., & Neukam, F. W. (1996) Re-establishment of speech and swallowing function following extensive tumour resections in the head and neck. *Laryngo-Rhino-Otologie*, 75, 231–8.

Scobbie, J. M., Wood, S. E., & Wrench, A. A. (2004) Advances in EPG for treatment and research: An illustrative case study. *Clinical Linguistics and Phonetics*, 18, 373–89.

Serrurier, A. & Badin, P. (2005) Towards a 3D articulatory model of velum based on MRI and CT images. *ZAS Papers in Linguistics*, 40, 195–211 (Zentrum für Allgemeine Sprachwissenschaft, Sprachwissenschaft, Typologie und Universalienforschung).

Shinagawa, H., Murano, E. Z., Zhuo, J., et al. (2008) Human tongue muscle fiber tracking during rest and tongue protrusion with oral appliance: A preliminary study with diffusion tensor imaging. *Acoustic Science and Technology*, 29, 291–4.

Shinagawa, H., Murano, E. Z., Zhuo, J., et al. (2009) Effect of oral appliances on genioglossus muscle tonicity seen with diffusion tensor imaging: A pilot study. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontology*, 107, 57–63.

Shinagawa, H., Ono, T., Ishiwata, Y., et al. (2003) Hemispheric dominance of tongue control depends on the chewing-side preference. *Journal of Dental Research*, 82, 278–83.

Slud, E., Stone, M., Smith, P. J., & Goldstein, M. (2002) Principal components representation of the two-dimensional coronal tongue surface. *Phonetica*, 59, 10.

Soder, N. & Miller, N. (2002) Using ultrasound to investigate intrapersonal variability in durational aspects of tongue movement during swallowing. *Dysphagia*, 17, 288–97.

Specht, K., Rimol, L. M., Reul, J., & Hugdahl, K. (2005) “Soundmorphing”: A new approach to studying speech perception in humans. *Neuroscience Letters*, 384, 60–5.

Steele, C. & Lieshout, P. H. H. M. van (2004) Use of electromagnetic midsagittal articulography in the study of swallowing. *Journal of Speech, Language, and Hearing Research*, 47, 342–52.

Stone, M. (2005) A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 19, 455–502.

Stone, M., Davis, E., Douglas, A., et al. (2001a) Modeling the motion of the internal tongue from tagged cine-MRI images. *Journal of the Acoustical Society of America*, 109, 2974–82.

Stone, M., Davis, E., Douglas, A., et al. (2001b) Modeling tongue surface contours from cine-MRI images. *Journal of Speech, Language, and Hearing Research*, 44, 1026–40.

Stone, M. & Lundberg, A. (1996) Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, 99, 3728–37.

Stone, M., Shawker, T., Talbot, T., & Rich, A. (1988) Cross-sectional tongue shape during the production of vowels. *Journal of the Acoustical Society of America*, 83, 1586–96.

Story, B. (2005) Synergistic modes of vocal tract articulation for American English vowels, *Journal of the Acoustical Society of America*, 118, 3834–59.

Story, B., Titze, I., & Hoffman, E. (1996) Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America*, 100, 537–54.

Subtelny, J., Li, W., Whitehead, R., & Subtelny, J. D. (1989) Cephalometric and cineradiographic study of deviant resonance in hearing impaired speakers. *Journal of Speech and Hearing Disorders*, 54, 249–65.

Suzuki, N. (1989) Clinical applications of EPG to Japanese cleft palate and glossectomy patients. *Clinical Linguistics and Phonetics*, 3, 127–36.

Suzuki, N. & Michi, K. (1986) Dynamic velography. *Proceedings of the 20th Congress of the International Association of Logopedics and Phoniatrics*, Tokyo, 172–3.

Suzuki, N., Sakuma, T., Michi, K. I., & Ueno, T. (1981) The articulatory characteristics of the tongue in anterior openbite: Observation by use of dynamic palatography. *International Journal of Oral Surgery*, 10, 299–303.

Tabain, M. (1998) Coarticulation in CV syllables: A locus equation and EPG perspective. *Journal of the Acoustical Society of America*,103(5), 2980.

Tameem, H. & Mehta, B. (2004) Solid modeling of human vocal tract using magnetic resonance imaging and acoustic pharyngometer. *Proceedings of the 26th International Conference of the IEEE: Engineering in Medicine and Biology*, San Francisco, 2, 5115–8.

Tasko, S. M., Kent, R. D., & Westbury, J. R. (2002) Variability in tongue movement kinematics during normal liquid swallowing. *Dysphagia*, 17, 126–38.

Tasko, S. & Westbury, J. (2002) Defining and measuring speech movement events. *Journal of Speech, Language and Hearing Research*, 45, 127–42.

Tasko, S. & Westbury, J. (2004) Speed-curvature relations for speech related articulatory movement, *Journal of Phonetics*, 32, 65–80.

Tom, K., Titze, I. R., Hoffman, E. A., & Story, B. H. (2001) 3-D vocal tract imaging and formant structure: Varying vocal register, pitch, and loudness, *Journal of the Acoustical Society of America*, 109, 742–7.

Tsuga, K., Hayashi, R., Sato, Y., & Akagawa, Y. (2003) Handy measurement for tongue motion and coordination with laryngeal elevation at swallowing. *Journal of Oral Rehabilitation*, 30, 985–9.

Tye-Murray, N. (1991) The establishment of open articulatory postures by deaf and hearing talkers. *Journal of Speech and Hearing Research*, 34, 453–9.

Ueda, D., Yano, K., & Okuno, A. (1993) Ultrasound imaging of the tongue, mouth, and vocal cords in normal children: Establishment of basic scanning positions. *Journal of Clinical Ultrasound*, 21, 431–9.

Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W., & Patterson, R. D. (2006) Locating the initial stages of speech–sound processing in human temporal cortex. *Neuroimage*, 31, 1284–96. Epub February 28, 2006.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005) Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, 117, 338–50.

Wakumoto, M., Masaki, S., Honda, K., & Ohue, T. (1998) A pressure sensitive palatography: Application of new pressure sensitive sheet for measuring tongue–palatal contact pressure. *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, 7, 3151–4. (Available online at: http://andosl.anu.edu.au.icslp98/main/html)

Watkin, K. L. (1999) Ultrasound and swallowing. *Folia Phoniatrica et Logopaedia*, 51, 183–98.

Watkin, K. L. & Rubin, J. M. (1989) Pseudo-three-dimensional reconstruction of ultrasonic images of the tongue. *Journal of the Acoustical Society of America*, 85, 496–9.

Wedeen, V. J., Reese, T. G., Napadow, V. J., & Gilbert, R. J. (2001) Demonstration of primary and secondary muscle fiber architecture of the bovine tongue by diffusion tensor magnetic resonance imaging. *Biophysics Journal*, 80, 1024–8.

Weismer, G., Yunusova, Y., & Westbury, J. R. (2003) Interarticulator coordination in dysarthria: An X-ray microbeam study. *Journal of Speech, Language, and Hearing Research*, 46, 1247–61.

Westbury, J. (1994) *X-ray Microbeam Speech Production Database User's Handbook*. Waisman Center, Madison: University of Wisconsin at Madison, 8–33.

Westbury, J., Lindstrom, M., & McClean, M. (2002) Tongues and lips without jaws: A comparison of methods for decoupling speech movements. *Journal of Speech, Language, and Hearing Research*, 45, 651–62.

Wood, S. (1979) A radiographic analysis of constriction locations for vowels. *Journal of Phonetics*, 7, 25–43.

Wrench, A. A. (2007) Advances in EPG palate design. *Advances in Speech-Language Pathology*, 9, 3–12.

Xue, S. A. & Hao, G. J. (2003) Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study. *Journal of Speech, Language, and Hearing Research*, 46, 689–701.

Yang, C. S. & Stone, M. (2002) Dynamic programming method for temporal registration of three-dimensional tongue surface motion from multiple utterances. *Speech Communication*, 38, 201–9.

Yoshioka, F., Ozawa, S., Sumita, Y. I., Mukohyama, H., & Taniguchi, H. (2004) The pattern of tongue pressure against the palate during articulating glossal sounds in normal subjects and glossectomy patients. *Journal of Medical and Dental Sciences*, 51, 19–25.

Zerhouni, E. A., Parish, D. M., Rogers, W. J., Yang, A., & Shapiro, E. P. (1988) Human heart: Tagging with MR imaging – a method for noninvasive assessment of myocardial motion. *Radiology*, 169, 59–63.

Zierdt, A., Hoole, P., Honda, M., Kaburagi, T., & Tillman, H. (2000) Extracting tongues from moving heads. *Proceedings of the 5th Speech Production Seminar: Models and Data*, Munich, 313–16).

Zierdt, A., Hoole, P., & Tillmann, H. G. (1999) Development of a System for Three-Dimensional Fleshpoint Measurement of Speech Movements. *Proceedings of the 14th International Conference of Phonetic Sciences (ICPhS '99)*, San Francisco, August.

2 The Aerodynamics of Speech

CHRISTINE H. SHADLE

1 Introduction

Aerodynamics is the study of the motion of air. It is a subset of fluid mechanics, since air is only one possible fluid; it is a subset in another sense because mechanics includes statics as well as dynamics, but one must understand something about fluid statics in order to consider dynamics. Acoustics is the study of sound, and sound involves a particular type of wave traveling through a medium. Acoustics in air is therefore a part of aerodynamics.

These distinctions may seem pedantic, but they are blurred often in speech research and result in some confusion. Aerodynamics in speech tends to be thought of as "everything the air is doing that isn't sound." In speech we ultimately care only about the sound that is radiated to the far field, well outside the vocal tract. Here, near the microphone or someone's ear, the air is essentially at rest except for the sound wave, and describing that wave is an acoustics-only problem. However, inside the vocal tract, the air is not at rest; we speak, for the most part, while exhaling, and the sound waves travel through that moving airstream. Further, most speech sounds are generated by that airstream: it sets the vocal folds vibrating which in turn chop up the steady airstream, and it can become turbulent and generate noise.

The chief difficulty in considering acoustics and aerodynamics of the vocal tract together is that they operate at different time and spatial scales. Convection velocities – how fast air moves from glottis to lips – are very slow compared to the velocity of sound. Conversely, the spatial resolution needed to model turbulence and its sound-generating mechanisms is much greater than that needed to model sound propagation. The usual approach is thus to consider the larger picture – i.e., the nonacoustic aerodynamics – in order to define the acoustic sources that are operating, include these sources in a model that considers only acoustic waves, and thereafter ignore the moving stream of air. However, our understanding of the various types of sources is not very far advanced in some cases, and the limitations of these definitions need to be understood. Further, some sources continue to interact with the moving air, and thus are less suited to such a separation of "acoustic" and "aerodynamic" function.

We also need to consider the wider aspects of aerodynamics when we measure speech or any aspect of speech production. There can be obvious effects, like the need to avoid breath noise on a microphone, or more subtle effects, like the limitations of inverse filtering. One can devise certain methods of recording various parameters in speech that avoid pitfalls, but new measurement techniques are developed all the time. It is important to be aware of the issues involved.

Aerodynamics texts are rarely written with speech applications in mind, and tend also to be highly mathematical. In spite of the high level of mathematics required, there are topics that currently resist any analytical solution, and must be dealt with empirically. In this chapter mathematics is not avoided altogether, but the chief aim is to convey an appreciation for the physical mechanisms involved, provide a pointer to more detailed treatments of each subject, and describe some of the limitations in our current understanding of the aerodynamics of speech.

In section 2 we describe some basic aerodynamic concepts and define the variables and nondimensional parameters needed. In section 3 we use these basic concepts to consider mechanisms of speech production, grouped in terms of the aerodynamic behavior(s) present. In section 4 we consider measurement methods and their limitations, including methods in general use and those adapted for speech research. Finally, in section 5 we discuss some models of speech production that incorporate aerodynamics.

2 Basic Considerations

We will first consider fluid statics, that is, the behavior of a fluid at rest, and the properties of a sound wave moving through it. Then we consider fluid dynamics. The motion of a fluid can alter sound passing through it; it can also generate sound, with the properties of that sound depending on the fluid motion and its interaction with its boundaries.

Air has a mass and a springiness, or compressibility. It takes energy to move air or to compress it, and the air imparts energy to an object that stops it from moving or confines it to its container when it expands. In a static situation – a set number of air molecules sealed in a container – the behavior of the air is described by its pressure, volume, and temperature, by the relation

$$PV = nRT \tag{1}$$

where P = pressure, V = volume, T = temperature, R = the universal gas constant, and n = mass of gas in moles (Halliday & Resnick, 1966). So, for this sealed-up gas where n cannot vary, if the temperature increases, the pressure or the volume or both must also increase; if the temperature stays the same, any increase in pressure must be offset by a corresponding decrease in volume, and vice versa.

The temperature, T, affects the density, ρ, viscosity, v, and speed of sound, c, in a gas. Equation (1) can be used to derive the equations relating T to ρ, v, and c. Values of these parameters for humid air at body temperature have been computed and are listed in the Appendix.

We are treating the enclosed mass of gas as though the pressure everywhere within it were constant. This is not strictly true: the gas at the bottom of the container has the weight of the gas above pressing on it, so its pressure is slightly greater. Because the density of air is low, it takes a very tall container for this effect to be noticeable: an increase in altitude of 1 km decreases atmospheric pressure by only 10 percent, for instance (Halliday & Resnick, 1966). But in a liquid, which is more dense, the effect is more noticeable, and this is exploited in the operation of the manometer, a basic instrument for measuring static pressure. In the manometer, a U-tube of constant inner diameter contains a liquid of known density ρ'. One end of the tube is attached to the gas with the pressure P to be measured; the other end is attached to a gas at a reference pressure P_0 (if that end is left open, the reference pressure is atmospheric pressure). The difference in the height of the liquid in each arm of the tube, h, is proportional to the difference in pressure:

$$P - P_0 = \rho' g h \tag{2}$$

where g is the gravitational acceleration. A denser liquid, with higher ρ', will show a smaller difference in height for the same pressure difference. Thus atmospheric pressure at sea level is 76.0 cm of mercury and 1,033 cm of water. The subglottal pressure during speech can range from 3 to 30 cm H_2O above atmospheric pressure; for such a relatively small value, the pressure can be measured more accurately by using water.

A sound wave traveling through a fluid that is otherwise at rest consists of a longitudinal pressure-rarefaction wave. This means that particles of the fluid are alternately pressed together more tightly than normal and pulled apart further than normal. As the wave travels through the fluid, individual particles oscillate about their original positions, but do not have a net movement. The molecules in the compressed regions tend to move towards the rarefied regions, so that particles in the rarefied regions have higher velocity. This tendency towards re-establishing equilibrium moves the high- and low-pressure regions along at a speed regulated by the properties of the fluid: the speed of sound.

The ideal gas law given in equation (1) can be simplified when we are talking about the pressure and volume changes induced by a sound wave traveling through air. In this case the gas is undergoing an adiabatic process, which means that no heat flows into or out of the system. Then

$$PV^{\gamma} = \text{constant} \tag{3}$$

where V = volume and $\gamma = 1.4$ for air. Note that this is not the same as saying that the temperature remains constant; instead, it says that if the temperature changes, it must change back again quickly before any heat exchange can take place. When a sound wave travels through air, the pressure at a given location increases and then decreases. The temperature locally rises and falls, but the sound wave passes through so quickly that it behaves adiabatically.

Pressure and particle or volume velocity of the fluid as a function of time and location in space are the basic quantities used to describe a sound wave. They can also be used to describe a fluid in motion without a sound wave traveling through it. As the name indicates, particle velocity, v, is the velocity at a specific point in a fluid, and is expressed in units of distance per unit time; a particle at that location will have that velocity. The volume velocity, U, instead describes the rate of volume flow per unit time past a particular cross-sectional area. Any differences in particle velocity across that area will be averaged out by the description in terms of volume velocity.

There are many different types of fluid flow; recognizing which type occurs in a certain situation allows one to simplify the equations describing the fluid motion accordingly. One of the simplest types of flow to describe is steady, incompressible flow. Steady flow means that the flow does not change in time: if we measure pressure and particle velocity at a particular point, the values will remain the same even as the flow continues past our measurement point. This means that the flow cannot be turbulent, since turbulence implies that pressure and velocity will vary randomly in space and time. But nonrandom changes over time are excluded as well: if the overall flowrate is very slowly increased and then decreased without producing turbulence, it is still not a steady flow.

Liquids are very nearly incompressible; gases, with their lower density, are compressible. Sound waves cannot exist unless a fluid is compressible. However, describing a fluid flow as incompressible does not mean we are restricting ourselves to liquids: it means that we are ignoring the compressible effects in our model. So, assuming steady, incompressible flow in a duct allows one to derive a form of Bernoulli's Equation relating the pressure and velocity at two places along the flow, assuming no work, heat transfer, or change of elevation occurs between those two places:

$$-gH_L = \frac{p_2 - p_1}{\rho} + \frac{v_2^2 - v_1^2}{2} \tag{4}$$

where g = the gravitational acceleration, H_L = head loss (or energy per unit weight lost to friction) from point 1 to point 2, p_1, p_2 = static pressure at points 1 and 2, v_1, v_2 = particle velocity at points 1 and 2, and ρ = density. We can use the relation of volume velocity to particle velocity $U = vA$ and the fact that the volume velocity will be the same at any point along the duct to rearrange the equation. The head loss is related to the internal energy of the fluid; because the fluid has friction, some energy is converted to heat. If we assume that the flow is frictionless, $H_L = 0$, and do some rearranging, we get:

$$U = \frac{A_2}{\sqrt{1-(A_2/A_1)^2}}\sqrt{\frac{2(p_1 - p_2)}{\rho}} \tag{5}$$

where U = volume flowrate (m^3/s), A_1, A_2 = cross-sectional flow areas at points 1 and 2 (m^2), p_1, p_2 = static pressure at points 1 and 2 (Pa), and ρ = density of the fluid (kg/m^3). Although this equation strictly applies only to frictionless, incompressible, steady flow, it is used in practice where these restrictions are violated to measure volume flowrates. The calibration procedures and empirical coefficients that can render such practice more accurate are discussed briefly in section 4, and more extensively in Doebelin (1983).

All fluids are viscous; as a result, the head loss can become significant for flow along a length of pipe. It is proportional to the length of pipe and to the flow velocity squared, but the constant of proportionality is an empirically determined friction factor that depends on the nondimensional parameters of wall roughness and Reynolds number. The Reynolds number is defined as the ratio of inertial to viscous forces, and can be determined by:

$$Re = VD/\nu \tag{6}$$

where V = a characteristic velocity, D = a characteristic dimension, and ν = the kinematic viscosity. For pipe flow, the V normally used is the average particle velocity in the center of the pipe (and, because of the averaging, is therefore typically capitalized in the literature, confusing it with volume) and D is the pipe diameter (Massey, 1984). Although we are not often called upon to compute the head loss in the vocal tract, the Reynolds number is used in models of speech production, and it is therefore important to understand what it means.

All fluid motion can be broadly classified into three regimes: laminar, unstable, and turbulent flow. For a particular geometry – take, for example, a constricted region in a duct – the flow progresses from one regime to the next as the Reynolds number is increased. For a particular size of that geometry, this could be observed simply by increasing the flow velocity. In laminar flow, at the lowest velocity range, individual particles follow paths that do not cross paths of other fluid particles. The particles nearest the walls of the duct will move the slowest, constrained by friction to stick to the non-moving walls. In the center of the duct the particles will move the fastest. In going through a constriction the flow will hug the walls of the duct, and the velocity gradient and therefore the velocity in the center of the duct will increase as the area decreases. Laminar flow is dominated by friction forces, and the empirical friction factor is highest for lowest Reynolds numbers.

As the flow velocity increases, inertial forces begin to dominate over friction forces. As the fluid enters the constriction, it overshoots a bit, and the moving flow separates from the walls. The vena contracta, thus formed, effectively reduces the area of the constriction. The region of transition from the fast-moving flow to the still flow near the walls is known as a boundary layer, and it can itself become unstable. In an unstable regime, any perturbations will tend to increase in amplitude.

If the Reynolds number is increased still further, the flow may pass through a sequence of unstable states, but eventually it becomes fully turbulent. Here inertial effects dominate. Paths of fluid particles cross each other unpredictably, so the flow as a whole has a random fluctuating component superimposed on the mean flow. This is very effective at mixing the flow.

For a particular geometry the characteristic velocity and dimension can be defined, and then a critical Reynolds number Re_{crit} can be found that marks the change from laminar to unstable flow regimes. This means that the flow regime can be predicted for any velocity in any size of that geometry. The value of Re_{crit} may differ though for a square instead of circular pipe, for instance, or a rectangular instead of circular constriction. The behavior above Re_{crit} may also depend on geometry: for fully turbulent flow in smooth pipes the friction factor decreases with increasing *Re*, but for rough pipes it remains relatively constant (Massey, 1984).

Sound waves traveling through a fluid can be affected by the flow regime. First, turbulence can diffract and absorb sound waves, though it is questionable whether this is a significant effect for speech (see discussion in Davies et al., 1993). Second, the sound wave traveling through a moving medium will travel faster downstream than upstream relative to an observer at rest. We can gauge the strength of this effect by computing the average Mach number $M = V/c$, where V = the average particle velocity of the fluid. In a vowel, where average volume velocity $U = 200$ cm^3/s and the most constricted region has an area of approximately $A_c = 1$ cm^2, the Mach number in the constricted region will be $M = U/(A_c c) = 200/(1 \cdot 35{,}000) = 0.0057$. Since $M \ll 1$, this effect is not significant. However, for fricatives, a typical $U \approx 600$ cm^3/s and $A_c \approx 0.1$ cm^2, so $M = U/(A_c c) = 600/(0.1 \cdot 35{,}000) = 0.17$. Here the value of M relative to 1 indicates that the convection velocity is significant with respect to the speed of sound, and may have to be taken into account.

In addition to these effects that flow can have on sound traveling through it, flow can also generate sound, with different characteristics according to the type of flow that produced it.

An unstable flow regime can lead to a self-sustaining aerodynamic oscillation. One or more positive feedback paths must exist. The sound that can result is characteristically high-amplitude, narrow-bandwidth: a whistle. Its frequency and the parameters that control it are related to the underlying instability.

We spoke earlier of the boundary layer that can detach from the walls of an orifice. In fact, a boundary layer exists between any two regions with significantly different flow parameters: they may have different velocities (as with the fluid moving in the center of a duct and the still fluid clinging to the walls of the duct), different densities (as with the Gulf Stream, which is warmer and saltier than the surrounding water), or actually be two different as yet unmixed fluids (cream just poured into coffee). The boundary itself is unstable for certain ranges of the difference of the two parameter values. In this unstable range, any small perturbation of the boundary will tend to grow. At first this will appear as ripples on the boundary; the ripples grow larger and curl up into vortices, which continue to rotate while being convected downstream.

The length of time required to traverse the feedback path tends to determine the spacing between vortices, because the initial perturbations are reinforced at that interval. In general an integral number of vortices will be found between abrupt discontinuities such as the two ends of a sharp-edged orifice, or the distance between an orifice exit and an edge. These patterns, and the sound generated, will couple into the resonances of a surrounding cavity. Increasing the flow velocity will tend to increase the frequency of the sound produced, but not uniformly; it will remain steadily coupled into one resonance, then jump abruptly to the next higher one, with the jumps exhibiting hysteresis (Chanaud & Powell, 1965; Holger et al., 1977).

Turbulence also generates sound, but since the motion is more random than an unstable state reinforced by feedback, the sound that results is noise with a relatively flat spectrum. Such noise cannot be predicted precisely from moment to moment, but can only be characterized statistically, and modeled by a collection of idealized flow-generated noise sources: the flow monopole, dipole, or quadrupole. These are analogous to idealized acoustic sources. The *acoustic* monopole can be thought of as produced by a pulsing sphere, which generates spherical sound waves. The acoustic dipole consists of two adjacent out-of-phase monopoles, which generate sound waves that interfere with each other; the result is a characteristic figure-eight directivity pattern. Solid objects such as a piston or a loudspeaker cone that act on the air can be modeled using these acoustic sources: in the far field, the directivity pattern observed is the same as that which would be produced by the idealized sources used to model it. In a *flow* source, the flow of air itself acts upon the surrounding air so that the far-field sound exhibits monopole, dipole, or quadrupole properties. Theoretically, the noise generated by turbulence away from any solid boundaries appears in the far field as if it were produced by flow quadrupoles; the noise generated by turbulence that results in a fluctuating force being applied to a solid object, by flow dipoles. As with acoustic sources, these can be thought of as collections of four and two flow monopoles, respectively, pulsing out of phase. In each case, the source strength depends upon the flow velocity. The total sound power of a flow quadrupole is proportional to V^8; that of a flow dipole, to V^6, and a flow monopole, to V^4. However, the flow quadrupole is much less efficient than a flow dipole, which in turn is less efficient than a flow monopole. It can be shown that the ratio of the total sound powers of the flow quadrupole to the flow dipole, or of the flow dipole to the flow monopole, is proportional to the Mach number squared (Goldstein, 1976). Thus, for $M < 1$, if a flow generates both dipole and quadrupole sources, the dipole sources will have higher sound power even though the sound power of the quadrupole sources increases faster with an increased flow velocity.

If the far-field sound pressure of a jet is recorded for a variety of jet sizes and mean velocities, the results can best be compared by plotting a normalized spectrum. A spectrum typically shows a measure of amplitude, such as sound pressure level, versus frequency. Every variation of V and D would result in a different curve. If we compared two circular jets with the same velocity V but different diameters, the larger jet will produce higher-amplitude noise with the peak at a lower frequency than that produced by the smaller jet. However, we can normalize the sound pressure level by dividing it by V^8D^2, which reflects the theoretically

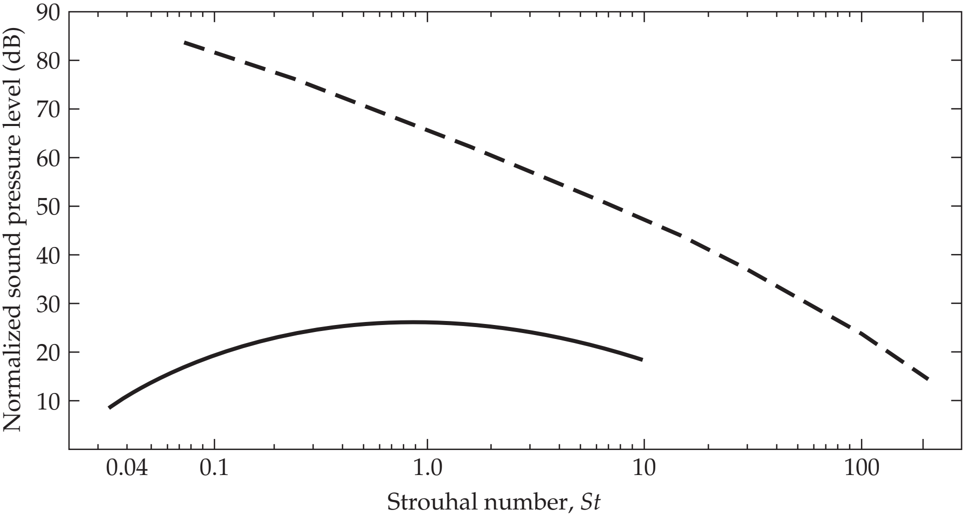


Figure 2.1 Normalized spectra of the noise generated by (a) free, subsonic jet noise (solid line) and (b) flow past a spoiler in a duct (dashed line), for various sizes of jets and spoilers. One-third octave sound pressure level is normalized by V^8D^2 for (a), by V^6 or V^4 for (b). Levels of the two curves relative to each other are arbitrary. (After Goldstein, 1976, and Nelson & Morfey, 1981)

predicted variation with V and D. We can also normalize the frequency axis by plotting instead the Strouhal number, St, where

$$St = \frac{fD}{V} \tag{7}$$

This will cause the peak frequencies to be aligned. As a result of these normalizations, all jet spectra for any size and velocity (as long as $V < c$) fit the same curve, as shown by the solid curve in Figure 2.1.

A similar collapse of data can be done for the noise produced by flow past a spoiler in a duct. The presence of the duct changes the dependence of source strength on V below the first cut-on frequency[1] (Nelson & Morfey, 1981). However, the principle of collapsing the data by using nondimensional parameters is the same. Here normalization by V^4 below cut-on and V^6 above cut-on frequency for the duct is used, and the resulting curve has a different shape from that of the free jet, as shown by the dashed curve in Figure 2.1.

The Strouhal number can be thought of in many ways. Equation (7) is derived by finding the ratio of the acceleration due to the unsteadiness of the flow to the convective acceleration due to the nonuniformity of flow. So, for small St, the unsteady component is relatively small. If $St < 10^{-2}$, the flow is quasisteady to a first approximation (Pelorson et al., 1994). The Strouhal number can also be used to characterize the shedding frequency of vortices from a jet. The frequency will depend on the jet velocity and diameter, but similar jets (same shape and thus behavior, even though of different D and V) will have the same St corresponding to the shedding frequency (Sinder, 1999).

3 Aerodynamically Distinct Tract Behaviors

In this section we consider the different mechanisms of speech production, grouping them from an aerodynamics point of view and proceeding from the simplest to the most complex. In each section we describe the physical events, and give parameter values typical for speech.

3.1 Breathing

Respiration is the simplest tract behavior aerodynamically because sound generation is not essential to the process and the time scales are relatively long.

The trachea extends about 11 cm below the larynx and then branches into the bronchial tubes. The bronchi continue to branch until the small, elastic-walled alveolar sacs are reached. The entire spongy mass is encased within the pleural sacs, which are suspended in the rib cage and surrounded on all sides by muscles. There are two sets of muscles that decrease lung volume when tensed: the internal intercostals, attached to the ribs, and the abdominal muscles. There are two sets that increase lung volume when tensed: the external intercostals, and the diaphragm, suspended across the bottom of the rib cage. By tensing and relaxing these sets of muscles in turn we can actively breathe in and out (as described by Hixon et al., 1973, and Hixon et al., 1976). But we can also use the elastic recoil force of the lung tissue itself as a passive mechanism for exhalation: if we cease to actively hold the rib cage expanded, we will passively exhale until the lung volume is small enough that the elastic recoil force no longer operates (see Figure 2.2). The lungs will not be empty at this point; the volume of air still in them is termed the functional residual capacity (FRC). To empty our lungs further we must actively tense muscles, and even doing so, we cannot empty them below a residual volume (RV).

The total lung capacity (TLC) in an adult male is approximately 7 liters of air. The RV is approximately 2 liters. The FRC varies with posture, but is typically 4 liters. The vital capacity is the maximum amount of air that can be exchanged in one breath, and is the difference between total lung capacity and residual volume, or about 5 liters (Ohala, 1990).

Typical respiration involves actively expanding lung volume (and therefore inhaling) to about 0.5 liters above FRC, and passively letting elastic recoil deflate the lungs (and therefore exhaling) back to the FRC. A typical respiration rate is 15 to 20 breaths per minute (Thomas, 1973). We hold the vocal folds as far apart as possible during inspiration (maximum area is 52 percent of tracheal area according to Negus, 1949; tracheal area ranges from 3.0–4.9 cm^2 according to Catford, 1977) and keep the tongue relaxed and velum down to provide a relatively

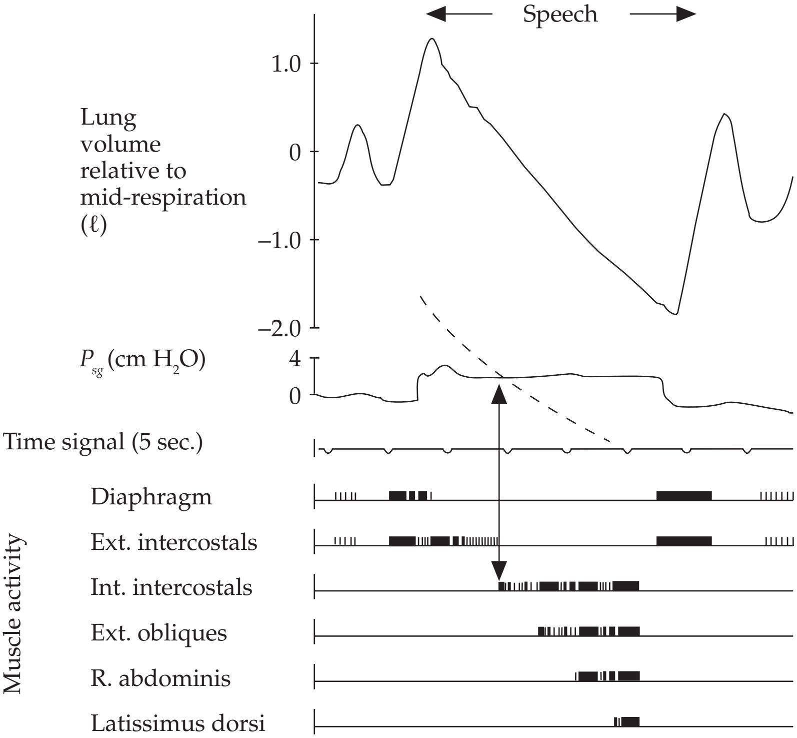


Figure 2.2 Lung volume versus time during speech and respiration, showing measured lung volume and subglottal pressure, and diagrammatic representation of the muscle activity. The dashed line indicates the relaxation pressure. (From Draper et al. (1959). Reprinted with permission from "Respiratory muscles in speech" by M. H. Draper, P. Ladefoged, and D. Whiteridge, *Journal of Speech and Hearing Research*, 7, 20. Copyright 1959 by American Speech-Language-Hearing Association. All rights reserved.)

unimpeded path for the air. During expiration the glottal area is smaller, but still of the order of 1 cm^2 (Sawashima, 1977); this is wide open compared to phonation, with an average glottal area of 0.05–0.1 cm^2.

For short utterances of speech at normal level, normal expiration is sufficient. For louder and/or longer speech, we need to use muscles actively to inhale more deeply, to offset the greater relaxation pressure, and to expel air below the FRC. During speech our goal appears to be to hold the subglottal pressure P_{sg} approximately constant, at a level corresponding to the loudness level of speech. It ranges from 3–30 cm H_2O (with normal speech typically 5–10 cm H_2O), as deduced by measuring esophageal pressure (Draper et al., 1959; Slifka, 2003) or by using tracheal puncture to measure the pressure directly (Isshiki, 1964). The lung volume then decrements fairly steadily; during stops the rate of decrement decreases momentarily, and during fricatives it increases (see Figure 2.3).

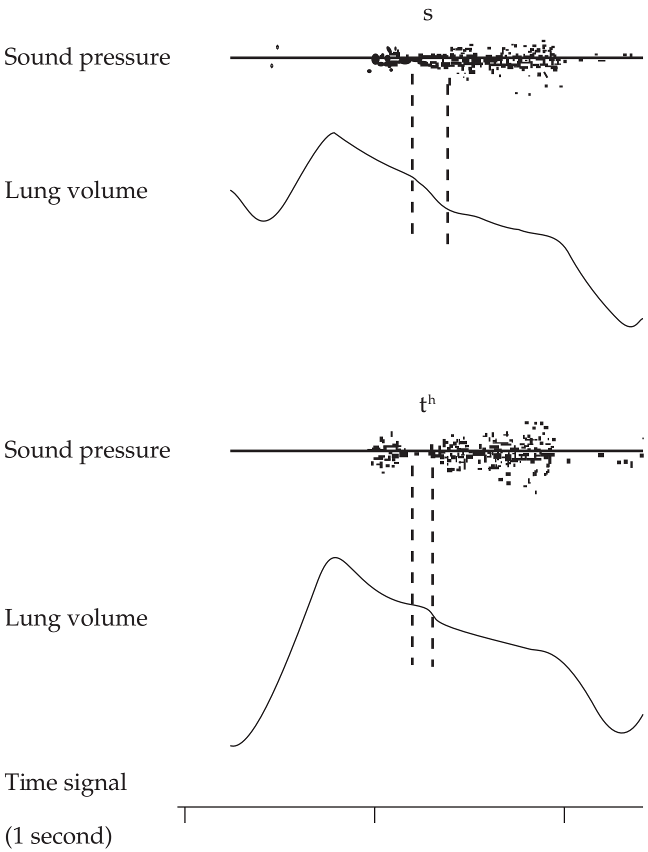


Figure 2.3 Lung volume versus time during the phrase "Deem–oon real," where the blank was filled in by [s] (top) and [t^h] (bottom). (From Ohala (1990). Reprinted from J. Ohala, "Respiratory Activity in Speech," in *Speech Production and Speech Modelling*, eds. W. J. Hardcastle and A. Marchal, p. 36, copyright 1990, Kluwer Academic Publishers, with kind permission of Springer Science and Business Media.)

The respiratory system can be modeled uncontroversially as a simple mechanical system, as described by Draper et al. (1959): a set of bellows, with one active force (external intercostals and diaphragm) pulling outwards on the handles, one active (internal intercostals and abdominal muscles) and one passive force (elastic recoil) pulling inwards, and a variable-resistance opening in the bellows. What remains controversial, however, is the control mechanism for such a model. Ohala (1990) asserts that we either aim for a constant pressure to be applied to the lungs, or a long-term constant lung-volume decrement, and provides evidence to support the former. In particular, he argues that observed variations in subglottal pressure and in lung-volume decrement are due to variations in the downstream flow resistance and to the inertia of the system, i.e., the time it takes to re-establish equilibrium. It is also true, however, that stressed syllables during an utterance are correlated with bursts of activity in the internal intercostal muscles (Draper et al., 1959). Both passive and active factors, then, may account for variations in the rate of lung-volume decrement.

The issue of passive versus active control mechanisms is also important in explanations of f_0 declination across the duration of an utterance. First, it is not clear whether declination is intended or a byproduct. Second, both respiratory and laryngeal muscles can affect f_0; it is not clear which produces declination. Variations in P_{sg} correlated with variations in f_0 are sometimes taken as evidence that respiratory activity is controlling f_0, but this is not necessarily the case since the tract impedance, including laryngeal posture, can affect P_{sg}. P_{sg} is a measurable quantity and is constant enough to seem to be a controlling parameter, but it is a result, not a pure source parameter. A fuller discussion of declination, which concludes that its cause remains unresolved, can be found in Ohala (1990).

3.2 Frication

Fricatives are produced by making a tight constriction, with area of the order of 0.1 cm^2, somewhere in the vocal tract. The air emerging from the constriction forms a turbulent jet, and this jet produces noise. For unvoiced fricatives the vocal folds are held apart, giving a typical glottal area of 1 cm^2. This means that most of the subglottal pressure is dropped across the supraglottal constriction, rather than across the glottis (the obvious exception to this is [h], where the glottal constriction can be the only constriction; as a result, the vowel context can make [h] into an approximant, as in /ihi/).

Although the area of the constriction is much smaller than any tract area used during a vowel, it is larger than the average glottal area during voicing and thus the volume flowrate is higher during unvoiced fricatives, ranging typically from 200 to 400 cm^3/s or more (for [h] it may be 1,000–1,200 cm^3/s). In a vowel–fricative transition usually the glottis opens before the supraglottal constriction is formed, resulting in a momentary maximum volume velocity (see Figure 2.4). Then, as $\hat{A}_c$ decreases, U decreases also and the pressure drop across the supraglottal constriction increases. At some point frication begins; it would be useful to be able to predict precisely when. For voiced fricatives, with a lower mean U, the situation is even more complicated: turbulence noise is usually generated more weakly than in the unvoiced equivalent, but it is also effectively modulated by the voicing. This was first described by Fant (1960); the changes to the noise source spectrum as a result of the modulation have been described more recently (Jackson & Shadle, 2000) and will be discussed further in section 3.4. Flanagan's model of fricatives (1972, pp. 248–59) incorporates modulation based on the Reynolds number, and is discussed further in section 5. For both voiced and unvoiced fricatives, the first question must be: for what dimensions and flowrate does a turbulent jet form? This can be rephrased as, what is the critical Reynolds number for vocal tract geometries?

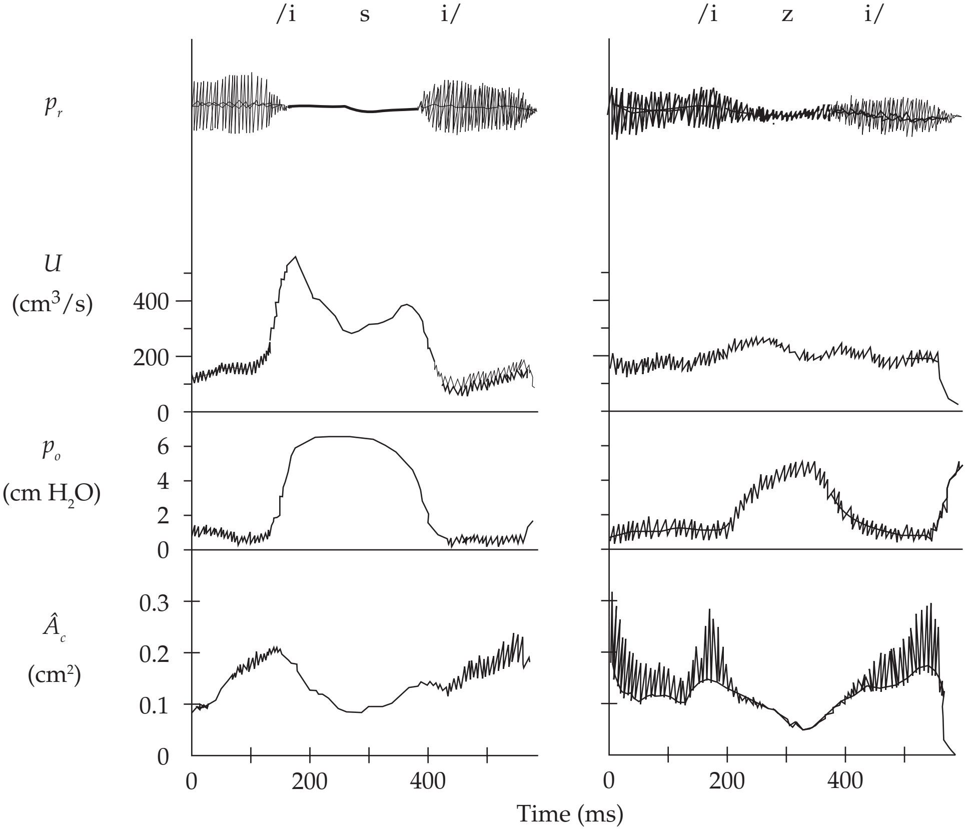


Figure 2.4 Time traces of measured radiated sound pressure p_r, volume velocity at the lips U, intraoral pressure p_o, estimated constriction area $\hat{A}_c$ for unvoiced and voiced fricatives. An adult male subject produced [pisi] (left) and [pizi] (right).

Meyer-Eppler (1953) conducted experiments to determine Re_{crit} for the fricatives [f, s, ʃ]. He measured radiated sound pressure (p_r) and oral pressure (p_o) for a speaker uttering the three fricatives and for air flowing through plastic tubes with three different elliptical constrictions. As shown in Figure 2.5, in each case a different minimum p_o was required to produce a measurable p_r; above this minimum, the rate of change of p_r with respect to p_o also varied. For the elliptical constrictions, he was able to arrive at a single line for p_r as a function of *Re* by using two different definitions of the effective width of the constriction for the three cases. For this line, he defined Re_{crit} to be the intercept where $p_r = 0$, and found $Re_{crit} = 1{,}800$. He then generalized this to speech, on the assumption that the same value of Re_{crit} would work for all fricatives provided the effective width was properly defined in each case.

This idea has gained wide acceptance. Studies using various ducts and orifices have led to a range of Re_{crit} values, from 1,700 to 2,300 (Ishizaka & Flanagan, 1972;

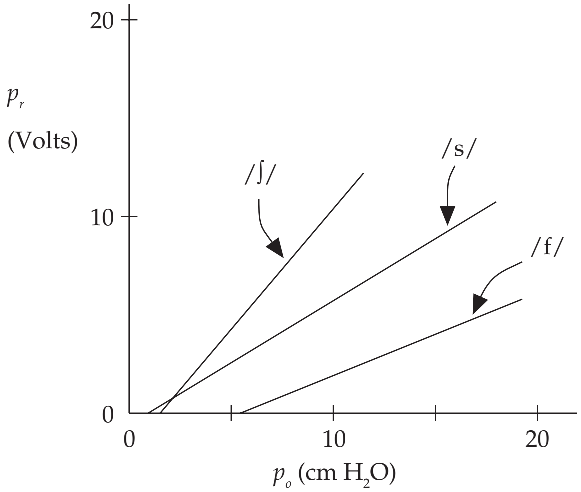


Figure 2.5 Radiated sound pressure p_r vs. intraoral pressure p_o for [f, s, ʃ]. (After Meyer-Eppler, 1953)

Catford, 1977). There are two problems, however. Since it is so difficult to measure the cross-sectional shape of the constriction, there is no independent check of *Re*. We do not know what the effective width should be for a particular constriction shape. Second, using the Reynolds number to collapse data carries with it the assumption that the geometries and therefore the source mechanism are the same, and thus allows comparison for different sizes and flowrates. But constriction shape is definitely not the same for different fricatives. Are we then losing or gaining by collapsing them together?

There is evidence that there are different source types operating to produce different fricatives. The noise produced by the jet alone, generated by relatively inefficient flow quadrupoles, is quite weak for the jet sizes encountered in the vocal tract. Anything solid in the path of the jet, however, produces a much more efficient noise-generation mechanism. Stevens (1971) recognized this difference, and adapted the work of Heller and Widnall (1970) on flow spoilers to frication. By treating the tongue-constriction as a spoiler, he found an equation giving source strength in terms of the pressure drop across the constriction. Although he acknowledged that the location of the constriction in the tract could affect the power-law relationship of the radiated sound power to the pressure drop, this was seen to be due to changes in the proximity of tract resonances to the source spectrum peak rather than an effect on the source mechanism.

Based on more recent analysis of speech and work with mechanical models, it appears that a flow dipole mechanism is operating, but not necessarily at the tongue constriction (Shadle, 1990, 1991). There are at least two distinctly different fricative geometries that result in different sources. The obstacle case has an obstacle such as the teeth at approximately right angles to the jet axis. The source is localized at the upstream face of the obstacle. [s, ʃ] fall into this category. The wall case has an "obstacle" such as the hard palate at a more oblique angle. The jet generates noise all along the wall, resulting in a much more distributed source. The fricatives [ç, x] and presumably all pharyngeal fricatives fall into this category. The weak front fricatives [f, θ] should also possibly be grouped in this category, since noise is clearly generated along the lips (Shadle, 1990). The "wall" does not continue on very far, however, and so it may be that these sounds should be considered as a third category.

The geometry affects not only where noise is generated, but how much, that is, the spectral characteristics of the noise and the way they change with flow velocity and area of the constriction. Rather than absorb these differences by means of effective width formulae, it would seem useful to express the acoustic properties of the noise in terms of the aerodynamic and articulatory parameters for each category. Some work has been done on this, e.g., source curves as a function of volume velocity have been measured for models of [ʃ, ç, x], and power laws have been determined for human speakers (Badin, 1989). Much remains to be done. For instance, it seems clear that ΔP across a constriction depends principally on the volume velocity through it and the constriction's shape and area. The amount of noise generated by it can be related to ΔP, but the particulars of the relationship will depend very much on what is present downstream of the constriction exit.

Sinder, Krane, and Flanagan took a more theoretical approach, developing a jet model based on Howe's work showing that sound generation occurs when jet vorticity crosses streamlines, as can occur with a change in duct area. Their jet model depends on the location of flow separation, and the geometry and flow speed at that location (Krane et al., 1998; Sinder, 1999). Some comparisons to experimental measurements of mechanical models were made (Sinder, 1999). Krane (2005) further elaborated the model, showing that the jet could be modeled for aeroacoustic sound generation purposes as either a train of vortex rings or a train of inclined vortex pairs. The source spectrum can be considered to be the convolution of a harmonic and a broadband function; the arrival time of the vortices determines which function dominates.

Howe and McGowan (2005) and McGowan and Howe (2007) also took a theoretical approach. For [s], they noted that the upper and lower teeth overlap, creating a channel in which turbulence could diffract sound and thus increase it, and obtained predictions of the radiated sound that matched experimental data. Their use of the compact Green's function explains the interaction of the source and the sound field and thus can accurately predict the level and shape of the source spectrum.

All of these fricative models (Shadle's, Krane & Sinder's, and Howe & McGowan's) have used drastic simplifications of the vocal tract shape during fricative production. Although the resulting source models differ somewhat, all agree that the geometry, not only the area function, of the constriction and downstream of the constriction have an important effect on sound generation by turbulence. It may also be that for some fricatives the articulatory charateristics most important for sound production have not all been identified. The ways in which the parameters of each model affect the sound generation need to be investigated further.

3.3 Transient excitation: Stops

Stops are intrinsically transient. Complete closure is effected somewhere in the vocal tract, from glottis to the lips. As shown in Figure 2.6, for a supraglottal unvoiced stop the pressure upstream of the closure typically builds up rapidly for a short time, and then continues to increase more slowly, possibly reaching a plateau. The neck and cheeks expand slightly in response to this pressure, and

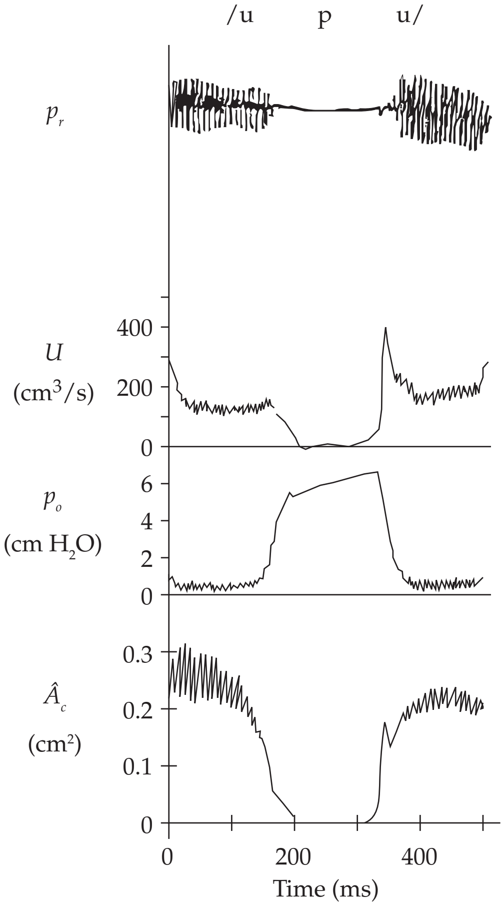


Figure 2.6 Time traces of measured radiated sound pressure p_r, volume velocity at the lips U, intraoral pressure p_o, and estimated constriction area $\hat{A}_c$ for the stop [p] in the context [upu]. An adult male was the subject.

the rate of decrease in lung volume eases slightly (Ohala, 1990). When the stop is released, either at the place of closure or at the velum, the oral pressure drops suddenly, lung volume suddenly begins to decrease more rapidly, and air is pushed out of the vocal tract explosively. The expelled air may become turbulent, and the patch of turbulence travels downstream, gradually dissipating. Depending on the position of the vocal folds, this brief period of high airflow may result in aspiration noise being generated.

The closure must be held for a perceptible amount of time, from a minimum of 20–30 ms to 100 ms or more. The release burst and ensuing frication last for a short time, of the order of 5 ms, and the aspiration, if it occurs, may last 50 ms or more before voicing begins. Indeed, the voice onset time will be longest if aspiration is present, and this is not a coincidence. Glottal area during stops is largest for unaspirated voiceless stops; for other cases, the glottal area depends somewhat on position of the stop within the word (Sawashima, 1977). Differences in voice onset times thus appear to be largely related to the time it takes to adduct the vocal folds (Catford, 1977). The wide-open glottis allows a high glottal volume velocity once the stop has been released, thus producing audible turbulence noise.

For a voiced stop, a pressure drop of at least 200 Pa must be maintained across the glottis for voicing to occur (Westbury, 1983). As a result the oral pressure does not increase as much as during an unvoiced stop. Fundamental frequency decreases as the pressure increases, and if closure is held long enough, vocal fold vibration may cease altogether. However, it appears that voicing during stops is extended by a combination of passive and active vocal tract expansion. The passive expansion occurs when cheek and neck tissues yield, puffing out slightly. We can control the degree of expansion somewhat by tensing or relaxing our cheek muscles; relaxed tissue yields more. The active expansion occurs by moving articulators: the larynx tends to move down, the soft palate up, and the tongue dorsum and blade down more during voiced than unvoiced stop closure. Both kinds of expansion serve to lower the pressure in the vocal tract, and therefore increase the transglottal pressure difference. Without such means, voicing can theoretically continue for approximately 60 ms after closure. With such means, voiced closure can extend theoretically to 200 ms or more – and in practice, voiced intervals of 100 ms or more are not uncommon (Westbury, 1983).

Sound production during and after the release has been modeled by Maeda (1987), and electrical analogs incorporating this developed by Stevens (1993, 1998). Maeda proposed a simple dynamic model that generates two different kinds of sources: an initial brief coherent source, followed by a longer frication source. The coherent source is caused by the assumption that when the closure is first opened, there is actually reverse flow into the sudden expansion, which causes a negative impulse of pressure. Although this flow monopole is predicted to last no more than 0.1 ms, it should be a very efficient sound source. The subsequent frication source is predicted to last from 1 to 5 or more milliseconds depending on the model parameter settings. Its strength would, of course, depend on the place of the constriction and the changing parameters.

Maeda demonstrated the coherent source with data from the utterance [mi]. The end of the [m] is released without a pressure buildup and therefore without the frication or aspiration sources, and shows a release bar on the spectrogram and extra negative-going radiated pressure. Flow visualization was done by seeding the flow with smoke particles. For a human subject, this is relatively easily accomplished by asking the subject to inhale cigarette smoke and using a high-speed camera to photograph the smoky jet leaving the lips as the subject says [mi]. Such a film showed a noticeable delay between opening the lips and the emergence of smoke, supporting Maeda's model (X. Pelorson, personal communication, 1994). More recently, Pelorson et al. (1997) studied bilabial plosives using flow visualization on human and mechanical models, numerical simulation, and theory. As Maeda found with [m], it takes approximately 20 ms for a jet to emerge once the lips are opened, and another 10–20 ms for vortex formation along the jet. The jet is essentially symmetric, in spite of the asymmetric lip horn, which might lead one to expect the jet to separate at different points from upper and lower lip. The frication noise that has been described to occur after the initial release (Stevens, 1993, 1998) appears likely to be produced by small-scale turbulence. Pressure–flow relationships are insufficient to describe this progression; the constriction shape as a function of time is needed. In the first few milliseconds after opening, viscous and boundary layer effects are important. After that, a low Reynolds number prevails, and boundary layer effects are not important.

3.4 *Mechanical oscillation: Trills and voicing*

Since the walls of the tract and the articulators are for the most part not rigid, it is possible for the airstream to set up a mechanical oscillation. This has been thought to be due to the Bernoulli force operating in the narrowed region such as the true or false vocal folds, the uvula, tongue tip or lips: here the air flows with a higher particle velocity and therefore the pressure drops. An inwards force is applied to the surrounding structure, and if that structure is flexible enough and the force strong enough, it may be pulled closed (see, for example, Catford, 1977). The closure of the "valve" formed by the vocal folds, tongue tip, etc. interrupts the airflow and allows pressure to be built up behind the closure, so it blows open and the process can repeat. The frequency of repetition is determined by both aerodynamic variables around the "valve," such as the original upstream pressure, the velocity through and area of the opening, and mechanical variables: the mass, compliance, and damping factors of the tissues making up the valve.

This simple model is no longer considered adequate. The quasistationary assumption on which the Bernoulli effect is based is good through much of the glottal cycle but does not hold when the glottis is very small, i.e., around closure. And the Bernoulli equation does not predict pressure–velocity relations well where the flow has separated (Titze, 2006).

The situation is somewhat similar to that of reed instruments such as the clarinet, in which the reed vibrates enough to close off the flow of air periodically, and those vibrations couple into and excite the resonances of the clarinet tube.

However, in the clarinet the natural frequency of the reed is well above the resonances of the tube, and so the pitch of the resulting sound is that of the lowest resonance (Benade, 1976). In the vocal tract, the natural frequency of the vocal folds is usually below that of the lowest formant, and so the pitch that results is that of the vocal fold vibration, ranging from 40 Hz (for creaky voice) to 1,000 Hz or more (for sopranos and children). For uvular and tongue-tip trills the mechanical oscillation is slower, in the range 20–35 Hz (Recasens, 1991; McGowan, 1992).

Vocal fold vibration has been extensively studied in both humans (see Hirose, this volume) and using excised canine larynges. There are many sets of muscles both in and around the vocal folds that can be adjusted to provide a very fine degree of control. The initial separation of the vocal folds, their length, and the tension of the three layers of the folds can all be separately controlled, in some cases by more than one mechanism. By these means the mode of vibration of the folds can be selected, and the frequency of vibration controlled within each mode.

The different modes of phonation are distinguished both by the pattern of movement of the vocal folds and by the resulting sound quality. The modes range from falsetto, in which the bulk of the folds are still and the margins vibrate, resulting in a relatively high-frequency sound with weak harmonics and a nearly sinusoidal glottal area function $A_g(t)$, to chest voice, in which a wave travels through the mucosa (the vocal fold cover) in the direction of the vocal tract's longitudinal axis, thus adding an extra component to the simple lateral motion of the folds (for more information see Gobl & Ní Chasaide, this volume). The closed phase for chest voice is a significant proportion of the total cycle, and upper harmonics of the fundamental carry a significant proportion of the total energy (Gauffin & Sundberg, 1989).

Within a mode, frequency of oscillation is primarily controlled by the length and tension of the folds and the subglottal pressure. The subglottal pressure is not an independent parameter in the way that the mechanical settings of the folds are: for instance, the minimum pressure required to achieve phonation appears to increase with f_0, and that relationship differs for singers and nonsingers (Titze, 1992, 2000).

There are numerous models of the vocal folds. Of the self-oscillating models, the best known are the one-mass and two-mass models (Flanagan & Landgraf, 1968; Ishizaka & Flanagan, 1972). Variations on the mechanical structure of the folds have included increasing the number of masses (Titze, 1973, 1974), using a distributed rather than lumped model (Titze & Talkin, 1979), a collapsible tube model (Conrad, 1985), and a translating and rotating one-mass model (Liljencrants, 1991a). In all of these, sufficient degrees of freedom are included to allow different modes of vibration. The different parts of each fold are coupled, either directly (e.g., via a spring) or indirectly (e.g., controlled by the same aerodynamic parameter). The effect on the flow of the current shape of the folds is handled generally by computing the point of flow separation within the glottis, and allowing pressure, velocity, and effective glottal area to vary accordingly. Pelorson et al. (1994) improved the two-mass model's performance by systematically testing different ways of computing the separation point, and incorporating the best model.

More recently, finite-element models (FEM) have been used, which are more computationally expensive but have the power to represent the internal mechanical properties of the vocal folds and, potentially, pathological structures as well. Gunter (2003) highlights the differences among some of them: Alipour et al.'s model (2000) is self-oscillating and is not restricted to unrealistic geometries like earlier continuum mechanics models, but lacks the fine spatial resolution and mechanical stress distribution calculations needed to investigate vocal fold pathologies. Jiang et al.'s model (1998) has a finer spatial resolution but does not represent collision forces, which are important for some vocal fold pathologies and studies of voice quality. Gunter's model (2003), intended for use in studying vocal fold pathologies, includes fine temporal and spatial resolution and represents vocal fold collisions, but is not self-oscillating.

Two very recent papers indicate still more progress. Tao et al. (2006) discuss a self-oscillating finite-element model with which they studied vocal fold impact pressure, relating that pressure to lung pressure and glottal width. Unlike Gunter's model, they modeled the air as well as the vocal fold tissue in order to have not only the interaction of the folds with each other, but the folds with the fluid, represented. Aerodynamic properties but not acoustic wave propagation were included. This model was then set up with a stiffness asymmetry, which they showed resulted in biphonation (Tao & Jiang, 2006). This particular structural asymmetry may not be the, or the only, cause of such biphonation, as they point out, but it does demonstrate the potential uses of the model and thus justifies the model's complexity.

It is difficult to test such models since it is impossible to compare "output" for a human phonating with the same parameter "settings." It is accepted, however, that source–tract interactions occur in humans (Rothenberg, 1981; Guérin, 1983; Titze, 2000), and evidence of such interactions is sought for each model. For instance, one of the advantages of the two-mass over the one-mass model is that the two-mass model shows more realistic behavior when F_0 approaches and exceeds the frequency of the first formant. More recent FEM models have been tested by, for instance, comparing pressures and predicted impact forces of one fold on the other (Story & Titze, 1995) with such pressures and forces measured in a canine larynx (Jiang & Titze, 1994). Even though the model is not tailored precisely to the particular subject, when the collision forces match but predicted peak pressure is five times smaller than measured, it is clear that the model is not yet realistic in that regard. The more complex a model is, and the more input variables it has, the more difficult it is to validate it, as described clearly by Gunter (2003).

There have been numerous studies of the detailed aerodynamics of the glottis using mechanical models. First, static models were used to measure the pressure–flow relationships (Scherer, Titze, & Curtis, 1983; Scherer & Titze, 1983; Scherer & Guo, 1991). Three glottal profiles have been used – convergent, uniform, and divergent – to capture various stages in a single glottal cycle; the findings were then assembled under a quasistationary assumption (Pelorson et al., 1994; Shinwari et al., 2003). Flow visualization of such models revealed the Coanda effect, in which a jet forms at the glottal exit and veers off to one side or the other, and remains attached to that side. However, an obstruction downstream, similar to the shape and position of the false vocal folds, could straighten the flow and prevent the Coanda effect (Shadle et al., 1991).

The static constraint on models has been lifted in a few different ways. One way is to use static models of the vocal folds, but start the flow impulsively (Hirschberg et al., 1996). This helped to establish where the separation point was at the glottal exit, how that varied according to the glottal profile, and the amount of time it took for a jet to appear and roll up into vortices. Hofmans et al. (2003) measured the time needed to establish the Coanda effect, which was much longer than a typical glottal cycle, and predicted therefore that it would not be able to be established in the more realistic situation when the vocal folds are moving. However, Erath and Plezniak (2006) did observe a Coanda effect for their static divergent glottal model with pulsatile flow.

The other class of mechanical model experiments involves a steady mean flow, but moving folds. These are usually driven, not self-oscillating, and have a fixed shape, but that shape can be varied between experiments in some setups. These have been used for various purposes, such as to visualize the flow downstream of the vibrating folds, with the results that the Coanda effect has been observed for a dynamic driven model with uniform glottis (Shadle et al., 1991). The quasistationary assumption was shown to hold apart from the early stages of the glottal cycle, when it departs significantly (Mongeau et al., 1997; Z. Zhang et al., 2002). Particle velocities and pressures have been measured in the tract in order to model the sound generation process (Barney et al., 1999).

The combination of all of these models with different constraints relaxed has reshaped our thinking about phonation. The separation point tends to be fixed when the glottal outlet is abrupt, as with a convergent or uniform glottis. With a divergent glottis, the separation point occurs before the glottal exit, at a point depending on the dimensions and angle of the glottis (if static). If the vocal folds are moving through all of these profiles, the separation point moves. A Coanda effect can then be observed within the glottis, with the jet attaching to one of the folds. The transition to turbulence is then asymmetric within the glottis, which changes the pressure–flow relationship significantly.

It has long been assumed that sound generation occurred at or near the glottal exit; this has been classically modeled as a monopole source, capturing the periodic appearance of the glottal jet. McGowan (1988) predicted theoretically that a downstream dipole source due to the vorticity–velocity interaction force, as well as a monopole source at the glottal exit, was necessary to characterize the phonation source. Since then, two experimental studies using driven folds have demonstrated other sources: Barney et al. (1999) showed that the glottal jet in their model developed a vortex street, and the vortices generated sound when they exited the tract. Z. Zhang et al. (2002) showed that the type of flow source varies during the glottal cycle; however, dipole sources dominate the tonal sound below 2 kHz. This result was supported by numerical simulations of Suh and Frankel (2007).

Although the quadrupole source generated by the glottal jet has been shown to contribute insignificantly to the radiated sound for driven mechanical models (e.g., Z. Zhang et al., 2002), turbulence noise generation at the glottis can become significant in breathy and hoarse phonation. In both cases, the vocal folds oscillate but do not completely close. In breathy voicing a chink is left open near the arytenoid cartilages (Fritzell et al., 1986; Södersten & Lindestad, 1990; Södersten et al., 1991). The dc offset measured in an inverse filtered glottal waveform is used in many studies as evidence of such a chink, and is observable in both men and women subjects "almost universally," though women's voice qualities, on average, are breathier than men's (Holmberg et al., 1988). Karlsson (1986), however, did not always observe a dc offset, even in her women subjects; when it occurred, it was mainly at weak effort levels. She noted large subject variation, and also cited several methodological aspects that could explain the differences among studies (Karlsson, 1992). The inverse filtering method itself does not take account of the difference in the velocities, and therefore travel time from glottis to lips, of convection and sound, which may further confound such studies. (See the further discussion in section 5.)

Hoarseness is more variable; it may be caused by swollen folds resulting in slow oscillation, a node on one fold preventing a clean closure, or a paralyzed fold allowing a more significant gap (Hammarberg et al., 1984). In all of these cases there is a relatively inefficient conversion of the energy from the steady airstream into sound. Some work has been done to model hoarse phonation by, for instance, modifying the two-mass vocal fold model to generate a pathological model (Koizumi & Taniguchi, 1990). The finite-element models discussed earlier can do this in more detail but so far do not predict acoustic output.

In breathy or hoarse phonation the turbulence noise fluctuates with the glottal cycle. This occurs in voiced fricatives as well, though the turbulence in that case is generated well downstream of the glottis. It has long been recognized that the frication noise is modulated by the voicing source, but the mechanism was not clear: does the sound generated at the glottis affect the turbulent jet downstream, or does the unsteady flow field generated by the oscillating vocal folds convect downstream and result in a pulsing jet at the constriction? It was observed that the harmonic and inharmonic components of the radiated sound were out of phase with each other during voiced fricatives, but not during vowels. The mechanism appears to be that sound generated at the glottis travels to the constriction, and there influences jet formation; the phase difference is related to the travel time from glottis to constriction at the speed of sound, and through the constriction and front cavity to the main noise source location at the slower convection velocity (Jackson & Shadle, 2000, 2001).

For tongue-tip trills, the vibrating structure is not so finely controlled as the vocal folds, and partly as a consequence has a smaller range of frequency of vibration. Both unvoiced and voiced trills can be produced. In either case, the tongue blade and dorsum are held steadily in position and the tongue tip vibrates against the hard palate at a rate of between 20 and 35 Hz. Closure is seldom complete, judging from electropalatography data of Catalan speakers and Rothenberg mask data (showing a nonzero minimum flow) of English speakers (Recasens, 1991; McGowan, 1992).

McGowan simulated the tongue-tip trill by modeling the tongue tip as a hinged trap door in the spirit of the one-mass vocal fold model. Wall compliance was included for the tract upstream of the tongue tip, and proved to be an essential part of the model. The oscillation of the tongue tip is only self-sustaining if net energy is transfered from the airflow to the motion of the tip during each cycle; this is accomplished if the pressure is greater during the opening phase than during the closing phase. This asymmetry occurs in the model because of the compliance of the walls. When the tongue-constriction is closed and the oral pressure rises, the walls expand. When the tip is released, they deflate, but they do so relatively slowly, thus maintaining a higher pressure for a time as the constriction opens. The wall effect is apparently more important for a smaller glottal area, since that limits the extent of variation in glottal volume velocity.

McGowan did not attempt to model the details of the flow near the tongue tip, and suggested that this might be important for two reasons. First, the simulated traces were much smoother than the measured ones. Second, flow separation in the constriction could also result in energy exchange tending to sustain the oscillation. Finally, although he included an adducted-glottis condition to approximate the average glottal opening during voicing, he did not actually allow the glottal area to vary, whether under direct control or via a self-oscillating vocal fold model.

3.5 Aerodynamic oscillation: Whistling

Whistling in speech occurs primarily in whistle languages (Busnel & Classe, 1976; Meyer & Gautheron, 2006), but may also occur in whistly fricatives, both deliberately as in Shona (Ladefoged & Maddieson, 1996, p. 171) and accidentally in languages that do not use a whistle for linguistic purposes (Shadle & Scully, 1995). Whistle languages can be used over distances of up to a few kilometres, and consist basically of a loud whistle that follows the $F2$ pattern of the whistler's ordinary language, or duplicates f_0 patterns of lexical tone. Whistly fricatives have whistles and frication noise occurring together; the whistle peak occurs generally in the high-amplitude region of the frequency spectrum, in the fricatives [s, ʃ, z]. Both kinds of whistling are best understood by considering "recreational" human whistling. Here there tends to be very little frication noise. The whistle may occur at $F2$ or $F3$, giving a frequency range of from 500 to 4,000 Hz (Shadle, 1983).

As described earlier, in order to produce a whistle sound there must be an unstable boundary layer and feedback that reinforces the instability. We would like to know when a whistle will occur and at what frequency, and therefore we need to know where the boundary layer forms and under what conditions it becomes unstable.

Because whistles are so geometry-dependent, the controlling parameters of a few classic geometries have been thoroughly investigated. Those that seem most applicable to the vocal tract are the orifice tone, the edge tone, and the hole tone. The orifice tone, however, depends on sharp edges at the inlet causing the boundary layer to separate from the walls of the orifice. This is inconsistent with the shape of the lips, and the controlling parameter – length of the orifice – predicts too high a whistle frequency (Shadle, 1985).

The hole tone can be produced without sharp-edged inlets. It results from two orifices in a row. The first produces an unstable jet, which curls up into vortices in the region between the orifices. In the absence of surrounding walls, the distance between the orifices determines the feedback path length; with surrounding walls, the whistle couples into one of the resonances of that cavity (Chanaud & Powell, 1965). If the constriction formed by the tongue is the first orifice, and the rounded lips form the second orifice, the resonances of the cavity in between should control the whistle frequency; the lowest of these is in fact *F*2 (Shadle, 1985), consistent with whistle languages.

For whistly fricatives, the edge tone appears to be a more appropriate model. In this geometry, the unstable jet formed by an orifice strikes a solid object: a sharp edge of varying angle, or a cylinder. With laminar flow, the jet will divide smoothly around the object. When the jet becomes unstable, it tends to go to one side or the other of the object, alternating periodically and shedding vortices alternately. Here the orifice diameter and the distance to the edge are critical parameters (Powell, 1961, 1962; Holger et al., 1977). Elder et al. (1982) describe how a combination of tones and broadband noise can be produced simultaneously. Using a mechanical model consisting of a long pipe with a side cavity, and edges protruding over the cavity opening, they measured sound produced as flow velocity was gradually increased, and identified the parameters controlling the various tones and turbulence produced. In addition to the well-known whistle phenomena of high-amplitude narrow-bandwidth peaks coupling into resonances for which the phase relationships reinforce the instability, they demonstrated how whistles combine with turbulence excitation of the pipe resonances. It appears that this mechanism could be at work with the whistly fricatives, with the tongue again forming the jet-producing constriction and the teeth serving as the edge. This is consistent with the role of the teeth in noise production, and simply indicates that some structure can exist in a turbulent flow (Shadle & Scully, 1995).

Because whistles are so sensitive to small changes in the geometry or flowrate, it is difficult to model them for the vocal tract where dimensions are difficult to determine and easily varied. They are also difficult to model for another reason: the whistle mechanism exhibits a complete interaction of “source” and “filter.”

4 Measurement Methods

4.1 Basic methods

A steady-state or slowly varying pressure can be measured by use of the manometer, which was described earlier. The tap can be placed in a sealed tank of gas, or at a particular place of interest along a duct. In the latter situation, where there is a relatively steady flow along the duct, the tap must be designed so as to measure the desired static pressure without altering the flow by its presence. In general, having the tap flush with the wall, of a diameter much smaller than the duct diameter, and the edge of the tap abrupt rather than beveled, is sufficient. One must also pay attention to local variations in the pressure. For instance, there is a net loss in pressure across an orifice, and it is often of interest to measure this difference. However, in and near the orifice the pressure may show the opposite tendency, rising just upstream of the constriction, dropping significantly just downstream, then gradually recovering somewhat. To measure the pressure drop reliably, then, one must space the taps away from the orifice by an amount that depends on the orifice shape; for instance, for a thin orifice plate, the taps should be located at 2½ diameters upstream and 8 diameters downstream of the orifice (Doebelin, 1983).

Pressure drop across a known orifice is often used to deduce flowrate. In some cases an existing orifice is measured and calibrated; in others, an orifice of known shape and area is inserted into a duct. In either case, the flowrate U is derived from the pressure drop measured at two taps for an incompressible fluid by:

$$U = \frac{C_d A_2}{\sqrt{1-(A_2/A_1)^2}} \sqrt{\frac{2(p_1 - p_2)}{\rho}} \tag{8}$$

where U is flowrate in m^3/s, A_1 = pipe cross-section area (m^2), A_2 = orifice cross-section area (m^2), p_1, p_2 = the pressure measured at the two taps (Pa), ρ = the density of the fluid (kg/m^3), and C_d is a dimensionless discharge coefficient that depends on Reynolds number and the ratio of orifice to pipe diameter, as shown in Figure 2.7. Including C_d, an empirically-determined coefficient that varies with orifice shape and the locations of the pressure taps, allows actual areas to be used rather than flow areas as in equation (5), and includes frictional losses. Calibration to determine C_d for every new setup can be avoided by using standard dimensions for the orifice meter and relying on the extensive experimental data available (Doebelin, 1983).

Equation (8) can be modified for compressible fluids to give the weight flowrate rather than the volume flowrate. For a small pressure drop ($p_2/p_1 > 0.99$), this is sufficient. For isentropic (i.e., frictionless and adiabatic) flows with larger pressure drop, an equation for weight flowrate can be derived whose only empirical coefficient is C_d. For a sharp-edged orifice plate, however, enough turbulence is generated that the isentropic assumption is not a good one. In this case, an experimental compressibility factor Y must be incorporated in the equation; Y depends on the pressure drop and orifice diameter in a different way for different placement of the pressure taps. For a known and stable configuration the final equation, though complicated, can be quite accurate (Doebelin, 1983). If the configuration is not known or is known to change, however, it may be more practical to use equation (8) for incompressible flow and determine or estimate an empirical

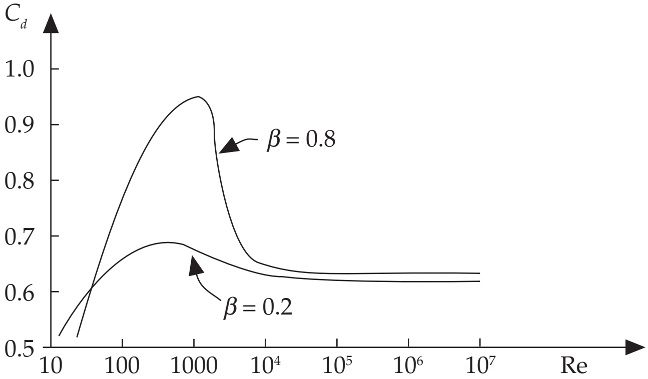


Figure 2.7 The dependence of discharge coefficient C_d on Reynolds number, Re, and on β, the ratio of orifice to pipe diameter. (After Doebelin, 1983, p. 531. Reproduced with permission of the McGraw-Hill Companies from E. Doebelin, *Measurement Systems*, 3rd ed., copyright 1983, McGrawHill)

coefficient for every change in geometry or significant change in flow (Massey, 1984).

Volume velocity can also be measured by a rotameter, which consists of a float in a vertical tube of varying cross-sectional area. The flow enters at the bottom of the tube and blows the float up to the point where the vertical forces of differential pressure, gravity, viscosity, and buoyancy are balanced. The same equations for flowrate as a function of area apply, but since the float position rather than pressure drop is the output measured, and flowrate is linearly related to float position (for the typical tube tapering) but related to the square root of pressure drop, the rotameter has a greater accurate range than orifice flowmeters (approximately 10:1 rather than 3:1 maximum:minimum flowrate, respectively) (Doebelin, 1983).

Particle velocity can be measured by a number of methods. The pitot tube is a probe that is placed directly into the flow, pointing upstream. It measures two pressures: the stagnation pressure, by a tap at its upstream end, and the static pressure, via taps along its sides. The difference between these two pressures can be used to derive the particle velocity at the location of the upstream tap.

Although the pitot tube is quite accurate, it cannot measure very low flow velocities, nor will it register quickly fluctuating velocities, as in turbulence. Higher frequency variations in particle velocity can be measured by using a hot-wire anemometer, which consists of a very fine wire with current passing through it. When held in a moving fluid, the flow cools it and changes the resistance slightly. In the constant-temperature form of the instrument, the current is adjusted to keep the wire temperature constant, as measured by its resistance. The square of the current is then related to the flow velocity. Because the wire is so fine, it responds quickly, and fluctuating flow velocities (up to as much as 100 kHz, depending on the compensating circuit) can be measured. Also, the wire and its support can be made small enough to provide minimal disturbance to the flow. The difficulties with the technique are that the wires are very fragile; they will not register flow direction, but only its magnitude; and each hot-wire must be calibrated with known velocities in the fluid in which it is to be used (Doebelin, 1983).

High-frequency pressure fluctuations can be measured with a microphone, but in some situations this includes more than the sound wave when only the sound wave is wanted. The most familiar example is breath noise; the microphone can be moved further away or out of the breath stream, or a foam windscreen can be used that absorbs the mean flow before it deflects the microphone's diaphragm. Inside a duct with both sound waves and a nonzero mean flow, similar problems occur. Pressure transducers can be flush-mounted on the walls, or effectively extended into the flow by use of probes. A probe in a moving fluid can, however, in itself become a location for sound generation. A different way to measure fluctuating pressures is to use two hot-wires a known distance apart. Their two velocities can be used to compute a velocity gradient proportional to pressure. The cross-correlation of the two signals can be used to compute the time delay between the two sensors, and therefore the speed of propagation of a particular signal. By this means hydrodynamic and acoustic pressure disturbances can be separated out: the former travel at approximately the mean flow velocity, the latter at the speed of sound.

Fluctuating pressures can also be measured in terms of the force they exert on an object. Heller and Widnall (1970) used force transducers to deduce the source strength from the force applied by the flow to spoilers in a duct. Accelerometers can also be quite useful for measuring the effect of flow on solid bodies, provided they have a mass much less than that of the object they are attached to.

Flow visualization can be accomplished by many techniques. The flow can be seeded with visible particles such as smoke, and pictures taken of the patterns thus revealed (e.g., Shadle et al., 1991, or Pelorson et al., 1997). Alternatively, the difference in refractive index caused by differences in density can be made visible by three different optical techniques. The shadowgraph technique is the simplest, but registers only large density gradients such as in shock waves. The Schlieren technique is more sensitive, but cannot reliably be used for absolute measurements of density (Pelorson et al., 1994; Pelorson et al., 1995). Interferometry can be used for quantitative density measurements, but is quite complex to set up. All three methods depend on passing light through the flow (Massey, 1984).

If the time between successive photographs is known, the time of travel of vortices and rate of their growth can be computed. In general, flow visualization works best with flows that are essentially two-dimensional, for example, with rectangular rather than circular jets. Obviously, internal flows (i.e., flow in ducts) cannot be visualized unless at least one wall of the duct is clear and the flow is "lit" by a means appropriate to the visualization method.

4.2 Speech-adapted methods

Ideally, in speech as in any other system, aerodynamic parameters should be measured without disturbing the flow producing them. Likewise, parameters not directly measured should be derived with due regard for the type of flow. However, the difficulties of accessing the vocal tract mean that many parameters cannot be measured directly, and a certain degree of pragmatism is therefore essential.

The aerodynamic parameters needed to model respiration tend to be more slowly varying than those for the larynx and supraglottal system. The lung volume cannot be measured directly; it is inferred by measuring changes in body volume. Total body volume can be measured with a plethysmograph, in which the body is sealed in an airtight container. Changes in volume are deduced either by measuring changes in pressure within the container, or by measuring the flowrate through a single port into the container. Alternatively, the motion of the thorax and abdomen can be monitored by use of multiple position sensors, and the lung volume then deduced (Draper et al., 1959; Hixon et al., 1973, 1976; Ohala, 1990; Slifka, 2003).

Subglottal pressure can be measured directly by tracheal puncture (Isshiki, 1964) or by pressure transducers lowered through the glottis (Cranen & Boves, 1985, 1988). It can be inferred from esophageal pressure (Draper et al., 1959; Slifka, 2003). All of these methods are invasive medical procedures requiring the presence of a physician, and thus cannot be done routinely. They can be invaluable to validate and evaluate other less invasive procedures, however. For instance, Cranen and Boves placed two pressure transducers above and two below the glottis. This allowed not only measure of the subglottal pressure, but use of the pressure gradient to deduce flow through the glottis, which could be compared to the glottal flow derived from simultaneous laryngograph, photoglottograph, and inverse filtering.

Supraglottal pressure can be measured directly much more easily than subglottal pressure by introducing a thin plastic tube at the side of the mouth and bending it behind the rear molars so that its open end is midsagittal and perpendicular to the longitudinal axis of the vocal tract. The pressure measured, P_o, should thus be the static pressure upstream of all labial, dental, and alveolar constrictions. The tube is typically attached to a pressure transducer sensitive to 1–2 kHz, and referenced to atmospheric pressure (Scully, 1986). It can then be used as an estimate of the pressure drop across the constriction, ΔP_c, with the proviso that $P_o \geq \Delta P_c$. During the stop [p], P_o increases quickly as pressure in the tract behind the constriction equalizes with the lung pressure. The maximum value of P_o measured during [p] can thus be used to estimate subglottal pressure, and the estimate can be extrapolated to the surrounding speech sounds. The exact value used depends on the respiratory model accepted, however: an assumption of constant pressure applied to lungs, or of constant lung-volume decrement, gives slightly different results (C. Scully, personal communication, 1994).

Volume velocity at the lips can be measured by a variety of masks containing flow or pressure transducers, some with nasal and oral airflow separately measured.

The Rothenberg mask provides the least acoustic distortion: it measures the pressure drop across screens of known flow resistance (Rothenberg, 1973). Its frequency range is limited to 0–1.8 kHz partly by that of the transducers used, but also by acoustic resonances of the mask itself (Hertegård & Gauffin, 1992). Although it is relatively nondistorting acoustically within this range, the screening very likely massively disrupts any vortex pattern emerging from the mouth. Whether or not this is significant for the far-field sound is unknown at present.

The volume flow from the mouth measured by the Rothenberg mask is also commonly used to estimate the volume flow from the glottis, U_g, by inverse filtering; U_g is then related to activity of the vocal folds and the voicing source. Because it is not invasive and provides an essential source function, it has been used extensively. Possible limitations of the method are related to the source–filter model of speech production on which it is based, and are therefore considered in the next section.

Particle velocity has been measured within the vocal tract by using shrouded hot-wire anemometers during production of open vowels. Open vowels were necessary so that the hot-wire holder could be inserted and traversed across the tract (Teager, 1980; Teager & Teager, 1983). The shrouding was used to enable detection of flow reversal; whether it does that without undue distortion of the flow is a matter of some debate. The hot-wires can be expected to have a short life in such an environment, but the difficulties of calibration for low flow velocities and the inherent inability of a single hot-wire to detect flow reversal are more significant problems (see, for example, the extensive discussion printed as part of Teager & Teager 1983, pp. 394–401). It can be quite useful, however, to use hot-wires in human subjects for validating more extensive hot-wire measurements done on mechanical models (see, e.g., Shadle et al., 1999).

The technique used by Heller and Widnall (1970) of mounting the flow spoilers on force transducers in order to measure the force generated by the flow directly is clearly not possible with an articulator like the tongue. However, accelerometers and other motion-sensing devices have been used in the vocal tract to measure motion of the velum, jaw, vocal folds, and tongue. It is beyond the scope of this chapter to review such methods. However, when aerodynamic parameters must be inaccurately measured, or deduced from indirect measurement, or outright estimated, the presence of independently obtained articulatory data can help put such estimates on a firmer footing. We describe such a process below.

In fricative consonants, the area of the constriction is a key parameter that is clearly related to the properties of the noise generated, although perhaps not so simply as has been proposed by Stevens (1971). It is difficult to derive this area from vocal tract imaging methods because it is so small. However, we can use an oral pressure tube and a Rothenberg mask simultaneously, and measure P_o and U_m as a function of time during, say, a vowel-fricative-vowel transition. We can then estimate the area, A_c, by rearranging equation (8):

$$\hat{A}_c = KU_m\sqrt{\frac{\rho}{2P_o}} \tag{9}$$

where K is an empirical constant, nominally a shape factor, taken equal to 1 by, for example, Scully (1986), and equal to 1/(0.65) by others. Since the constriction area has been assumed to be much less than the tract area ($A_c = A_2 \ll A_1$), K should correspond approximately to $1/C_d$.

An obvious limitation of this estimate is that P_o does not measure the pressure drop across the constriction only: lip rounding will increase it while not affecting constriction area or, presumably, frication noise. A less obvious problem is that this form of the equation is based on steady, incompressible, frictionless flow, which we clearly do not have. Although equation (9) is used for flow measurement in cases that also violate these assumptions, that is done for particular geometries for which extensive empirical data exist. Not only are such data nonexistent for the vocal tract, but the geometry is continually changing. The little we do know indicates that if we use Reynolds numbers and area ratios appropriate for the transition to and from a fricative, Figure 2.7 predicts that the discharge coefficient C_d in equation (8) will traverse a range of values, from approximately 0.9 to 0.6, yet K is typically held constant.

Pelorson (2001) tested this approximation and three variations on it by using mechanical models with three different constriction shapes, all possible shapes for speech. He showed that the best estimate of the area is found when the flow separation point can be estimated from a knowledge of the constriction shape. Since this is not always possible in speech, equation (9) using $K = 1$ is within 20 percent of the real area. The equation was also tested on unsteady flow, and is a reasonable approximation except near closure. It should be used with care for fricatives when turbulence is likely to occur within the constriction, as the losses will then be higher.

A related problem occurs in estimating the flow resistance of constrictions, which is relevant for the glottis as well as for fricatives. A typical procedure is to use the average volume velocity through and pressure drop across a constriction to define an operating point on an essentially parabolic function. The incremental resistance is then defined as the slope of the tangent to the curve at the operating point (Heinz, 1956). Pressure fluctuations due to sound waves are assumed to be small excursions about that point which can be modeled linearly; for small sound pressure amplitudes, this assumption is borne out by the measurements of Ingard and Ising (1967). However, the flow resistance in practice is often deduced from constriction area and volume velocity alone (Badin & Fant, 1984), whereas constriction shape can influence the pressure drop and, therefore, the operating point (Shadle, 1985).

5 Models Incorporating Aerodynamics

The classical acoustic theory of speech production models the acoustic properties of the vocal tract as an analogous electrical network. In so doing, several assumptions are made: sources and filter are independent, the filter is composed of passive elements and constitutes a linear system, sound propagation is one-dimensional, and in the most restrictive models, there is no mean flow. In this type of model, all aerodynamic effects are essentially confined to the source functions. Because source and filter are independent, whistles or whistly fricatives cannot be generated, but this lack would not of itself be of undue significance for speech models for most languages. Unfortunately, problems of greater consequence do arise; it is instructive to consider the ways in which some existing models have approached greater physical realism by relaxing some of the assumptions.

All models of phonation must include mean flow as an input, and, classically, fluctuating volume velocity is generated as an output. However, tract models do not always include mean flow. How can fricatives then be generated? Scully (1990) includes mean flow in her synthesizer by having separate acoustics and aerodynamics blocks. The aerodynamics block computes static pressure and mean flow throughout the tract, including the lungs, and generates frication sources with strength related to the pressure drop across the constriction. These sources are then fed forward to the final source–filter model. The sources and filter cannot interact extensively, but some influence is possible via numerous interconnecting paths.

A somewhat different approach is taken by Flanagan and Ishizaka (1976), who derive the fluctuating glottal flow from the two-mass model and a mean flow from a dc atmospheric-pressure source. This arrangement allows respiration, as well as frication. Frication is then modeled by providing each transmission-line section with a noise pressure source parameterized by Reynolds number. A particular source would generate noise only if the area and volume velocity in that section resulted in $\mathrm{Re} > \mathrm{Re}_{\mathrm{crit}}$. The amplitude of the noise source is modulated by a function proportional to Re^2, making the modulation observed in voiced fricatives possible (as demonstrated in earlier work based on a similar model, Flanagan, 1972). The noise source spectrum is flat, a reasonable simplification given the frequency range of the simulation (0–4 kHz).

A later synthesizer modified this scheme by using only one Re^2-dependent noise source per constriction (Sondhi & Schroeter, 1987). Location of the source was problematic, however: the internal impedance of the source was high enough that it restricted the volume flow unnaturally when placed at the constriction exit. Locating it downstream got around that problem, but each consonant required a different source location. This indicated greater physical realism, consistent with mechanical model studies (Shadle, 1985), but was "very inconvenient" in the context of an automatic text-to-speech synthesizer. The solution adopted was to place the source one section downstream of the narrowest part of the constriction, and represent it in parallel form, as a volume velocity source.

Narayanan and Alwan (2000) adopted a similar framework but worked to specify more physically realistic noise sources for a parametric synthesizer. They combined three-dimensional data derived from magnetic resonance imaging and source characteristics derived from mechanical model studies with an analysis-by-synthesis approach. All fricatives had a dipole source, corresponding to an obstacle downstream of the constriction, of either the teeth or the lips. All fricatives also had a monopole source, based on Pastel's findings (1987) from her experimental work modeling noise sources near the glottis. In addition, the palatoalveolar fricatives had another dipole source modeling wall noise. The monopole source was found to be unimportant, and the dipole source locations could be chosen uniformly within each fricative class, thus agreeing with mechanical model results. By adjusting the source strengths and spectral characteristics of the different sources, best fits were found for each fricative; in general, the stridents were matched more successfully than nonstridents.

There are no experimentally-derived source models for interdentals, and the position of the noise sources so near the lip opening is likely both to change source characteristics and make radiation and other loss models more critically important. For sibilants, the known changes in source characteristics above and below the cut-on frequency of the duct have not been modeled by Narayanan and Alwan, perhaps because this would be inconsistent with their assumption of a plane-wave model. These simplifications, as well as allowing the source strengths to be adjusted relative to each other, decrease the physical realism of the parametric source models. However, this work represents the best effort to date at modeling fricatives within the classical framework.

The inverse filtering procedures using the Rothenberg mask are based on a similar model of an independent source and a filter that is linear, time-invariant, and composed of passive elements only (Rothenberg, 1973). The model allows for the glottal volume velocity, including a mean flow component, to be estimated from the pressure drop measured across the mask; if intraoral pressure is simultaneously measured, the source strength of a stop or fricative can also be estimated.

To estimate the voicing source, the vocal tract transfer function must be calculated; the difficulties of doing so are discussed elsewhere in this volume (see Gobl & Ní Chasaide, this volume). The glottal volume velocity so derived is hard to reconcile with what we now understand of the physics at the glottis. A dipole as well as a monopole source are needed; which is dominant varies during the glottal cycle, as discussed in section 3. Travel time from glottis to lips is much slower at convection velocity than at the speed of sound (e.g., 170 ms and 0.5 ms, respectively), so that flow passing through a glottal chink arrives at the mask as a dc component much later than does the sound that was generated at the glottis at the same time. This disparity will vary with the tract area function and subglottal pressure, and so cannot be easily estimated and compensated for. While inverse filtering is undoubtedly useful, it appears that the waveform it generates has a more complex relationship to actual velocities existing near the glottis than was originally appreciated.

A more recent model of sound propagation in the vocal tract (Davies et al., 1993) retains a separation of source and filter while relaxing many of the traditional assumptions. Sound propagation is not always one-dimensional, and it need not be isentropic near junctions; mean flow is allowed, and the speed of sound is adjusted accordingly, but flow sources are not generated by the model. Because the tract is not modeled as an electrical analog, but instead is divided up into different duct elements that affect sound propagation differently, more physical realism is possible while still remaining powerful conceptually.

Teager sought to relax the assumption of independent source and filter. His hot-wire data showed evidence of nonuniform velocity across the vocal tract during vowel production (Teager, 1980). He suggested that source–filter interaction was therefore essential to a speech production model, and described a jet-cavity interaction paradigm (Teager & Teager, 1983). However, he did not propose a quantitative model, as discussed by Hirschberg et al. (1996).

Recently, quantitative models of aeroacoustic processes have been proposed. Pelorson et al.'s (1994) model predicting flow separation within the glottis has been discussed in section 3.4. Hirschberg et al. (1996) place this in a more general context. It is shown that viscosity, the friction of the fluid, cannot be neglected; including it predicts not only a pressure drop across the glottis as is observed, but a boundary layer next to the walls. Where the flow separates from the walls, forming a shear layer, the strong gradient across that layer generates vorticity, which causes the edges of the jet to roll up into vortices; the vorticity itself is the source of sound in the jet. Predicting boundary layer behavior precisely is difficult, especially for the complex glottal geometry; Pelorson's model is simplified, but works well, and is crucial in predicting sound generation. The simplification also allows it to be used for speech synthesis.

Hirschberg et al. also note that, although the turbulence of the jet generates noise, sound generation is significantly increased if there is a constriction downstream of the region of jet formation. The false folds result in dipole sources which are more efficient than the jet's quadrupole sources; similarly, teeth and/or lips downstream of a supraglottal constriction can generate efficient dipole sources in stop and fricative production. Sinder's model (Sinder, 1999), discussed in section 3.2, likewise uses a vorticity model to generate sound sources. This has been used as the basis of a synthesizer that generates its own noise sources (Krane et al., 1998).

McGowan and Howe (2007) describe the use of the Green's function, a general transfer function that includes but is not limited to the case of plane-wave propagation of sound, to model the exchange of energy between the hydrodynamic and acoustic modes of motion. With highly simplified geometry they are nevertheless able to explain why the dipole sources that so many studies have shown to be produced by a jet interacting with a downstream solid boundary, whether the glottal jet at the false folds or a supraglottal jet at the teeth, vary in their contribution to the far-field sound, especially at higher frequencies. They note that predicting the hydrodynamic field still must be done either experimentally or by numeric simulations, but their theoretical framework constitutes a model that is conceptually very powerful.

Full continuum simulations of the entire flow field represent another approach to including aerodynamics in a model of the vocal tract. Numerical simulations divide the fluid up into small volume elements, across each of which the conservation laws must hold. The equations of fluid motion are then solved, as well as the interactions of the fluid with the solid boundaries. Two critical decisions are the spatial resolution, i.e., the size of the volume elements, and the time resolution, the time steps for which pressure and velocity distributions will be computed. The spatial resolution determines the fluid structures, such as size of vortices, that can be simulated; the time resolution affects the stability of the solution, and determines the bandwidth for which the results are valid. Thus, simulating turbulence requires shorter time steps than simulating laminar flow (Blazek, 2005).

As a result, early studies (e.g., Thomas, 1986; Iijima et al., 1990; Liljencrants, 1991b) were limited to laminar flow through simplified geometries, simulating, for instance, pressure–flow relationships in simple glottal models. As computing power has increased, so has the usefulness of numerical simulations. In the mid-1980s Navier-Stokes equations could be solved numerically, allowing viscous flows to be simulated. This led in turn to methods developed for simulating turbulence: Direct Numerical Simulation (DNS), which is computationally the most intensive, Reynolds-averaged Navier-Stokes (RANS), which predicts mean flows only, and Large-Eddy Simulation (LES), which predicts instantaneous flow but of the large-scale motion only. Many other ways of reducing the computational load have been developed, including use of nonuniform grids, so that smaller elements can be used in the boundary layer than in the main flow channel. Ways of dividing a long duct into short sections, and simulating the flow in each in succession, have been developed. When possible, simulation is done in two dimensions, though this does not work well for simulating turbulence. There are also different ways to predict the sound generated. The compressible form of the Navier-Stokes equations (NSE) can be solved, which predicts the sound field directly; however, this is inaccurate for low flow speeds. Alternatively, the incompressible Navier-Stokes equations can be solved, and then an acoustic analogy used on the predicted pressure and velocity fields (Suh & Frankel, 2007).

As examples of the ways these different constraints can be traded off, consider these recent studies. Zhao et al. (2002) used the compressible NSE on an axisymmetric model of the vocal folds. A moving grid was used for the walls, which allowed forced oscillation to be simulated. The model could not predict turbulence, but the results did show vortex formation downstream of the glottis, and predicted that dipole sound sources due to the unsteady motion on the walls of the glottis were dominant. The effect of false folds and subglottal pressure variations could be studied (C. Zhang et al., 2002).

Adachi and Honda (2003) used LES to model fricative sound production. Complex vocal tract shapes were derived from MRI; the airflow in only the most anterior 4 cm of each vocal tract was simulated to make the problem computationally feasible. Even so, approximately 15 million cells were required to simulate turbulence. They were able to generate 10 ms of far-field sound up to 16 kHz, which compared reasonably well to sound produced by mechanical models of the same two vocal tract shapes.

Suh and Frankel (2007) used three-dimensional LES and an unsteady, compressible formulation to study flow through a static glottis, and predict sound generation for convergent and divergent glottis shapes. Their predictions agreed well with experimental results, and they were able to explain the sound generation mechanisms at different frequencies in detail.

There are other examples in the recent literature, but these will suffice to demonstrate on the one hand, the severe constraints of numerical simulation, but on the other hand, that judicious choices can complement experimental results and aid in the development of simpler aeroacoustic models.

In summary, a variety of models exists that incorporate aerodynamics to a greater or lesser degree. Although it is difficult to model such effects as phonation (fluid–solid interaction) or frication (turbulence) because the underlying phenomena are incompletely understood and resist an analytical solution, the fact that aerodynamics underlies all aspects of speech production makes such efforts important.

NOTES

I would like to thank Dr. Khalil Iskarous for comments on an earlier version of this chapter. Thanks also to Dr. Celia Scully for the raw data incorporated in Figures 2.4 and 2.6, which were obtained as part of an EC SCIENCE award, CEC-SCI*0147C(EDB). NIH grant NIDCD 006705 provided partial support during the preparation of the version for the 2nd edition of the Handbook of Phonetic Sciences.

[1] At frequencies below the first cut-on frequency only plane waves propagate, which excite the longitudinal modes. The first cut-on frequency, above which transverse as well as longitudinal modes can propagate, depends on the duct's cross-sectional shape and inversely on its largest cross-dimension. A circular duct 4 cm in diameter has a cut-on frequency of approximately 5 kHz (Kinsler et al., 1982). This is why using only cavity lengths and area ratios to compute formant frequencies works well up to 5 kHz, and less well above that.

APPENDIX: CONSTANTS AND CONVERSION FACTORS

The following values hold for dry air at 37°C. Values for completely saturated air are given in parentheses where available. (From Batchelor, 1967, and Davies, 1991.)

c	= speed of sound	= 35,300 cm/s	(35,900)
γ	= ratio of specific heats	= 1.400	(1.396)
R	= gas constant	= 2.87×10^6 erg/g	(2.977×10^6)
ρ	= density	= 1.139×10^{-3} g/cm^3	(1.098×10^{-3})
μ	= absolute viscosity	= 1.89×10^{-4} g/cm-s	
ν	= μ/ρ = kinematic viscosity	= 0.166 cm^2/s	
P_0	= standard atmospheric pressure at sea level	= 760 mm Hg	(760)

The conversion table for units of pressure in Table 2.1 should be interpreted as follows: 1 of the unit chosen from the leftmost column equals x of the unit chosen from the topmost row, where x is the value found at the intersection of the chosen row and column. For example, 1 bar = 10^5 Pa.

The conversion table for units of volume velocity in Table 2.2 should be interpreted as follows: 1 of the unit chosen from the leftmost column equals x of the unit chosen from the topmost row, where x is the value found at the intersection of the chosen row and column. For example, 1 liter/sec = 60.0 liters/min.

Table 2.1 Pressure (force per unit area)

	dyn/cm²	*Pa*	*bar*	*atm*	*cm H_2O*	*in H_2O*	*mm Hg*	*lb/in²*
dyn/cm²	1	0.1	10^{-6}	9.869×10^{-7}	1.0197×10^{-3}	4.015×10^{-4}	7.501×10^{-4}	1.4503×10^{-5}
Pa	10	1	10^{-5}	9.869×10^{-6}	1.0197×10^{-2}	4.015×10^{-3}	7.501×10^{-3}	1.4503×10^{-4}
bar	10^{6}	10^{5}	1	9.869×10^{-1}	1.0197×10^{3}	4.015×10^{2}	7.501×10^{2}	1.4503×10^{1}
atm	1.013×10^{6}	1.013×10^{5}	1.013	1	1033.0	406.8	760.0	14.7
cm H_2O	980.71	98.071	9.8071×10^{-4}	9.865×10^{-4}	1	0.3937	0.7355	1.422×10^{-2}
in H_2O	2491.0	249.1	2.491×10^{-3}	2.458×10^{-3}	2.54	1	1.868	3.613×10^{-2}
mm Hg	1.333×10^{3}	133.3	1.333×10^{-3}	1.316×10^{-3}	1.3597	0.5353	1	1.934×10^{-2}
lb/in²	6.895×10^{4}	6.895×10^{3}	6.895×10^{-2}	6.805×10^{-2}	70.307	27.68	51.71	1

Note: 1 μbar = 1 dyn/cm²; 1 Nt/m² = 1 Pa.; 1 psi = 1 lb/in². The threshold of hearing, often used as a reference pressure in computing decibels, is 20 μPa = 2×10^{-5} Pa. Values in this table are derived from Halliday and Resnick (1966).

Table 2.2 Volume velocity (volume flow past a cross-section per unit time)

	cm³/s	*m³/s*	*l/s*	*l/min*	*ft³/min*	*in³/s*
1 cm³	1	10^{-6}	10^{-3}	6.0×10^{-2}	2.119×10^{-3}	6.102×10^{-2}
1 m³/s	10^{6}	1	10^{3}	6.0×10^{4}	2.119×10^{3}	6.102×10^{4}
1 l/s	1.000×10^{3}	1.000×10^{-3}	1	60.0	3.531×10^{-2}	61.02
1 l/min	16.67	1.667×10^{-5}	0.0167	1	3.531×10^{-2}	1.017
1 ft³/min	471.9	4.719×10^{-4}	0.4719	28.32	1	28.80
1 in³/s	16.39	1.639×10^{-5}	1.639×10^{-2}	0.9834	3.472×10^{-2}	1

Note: cfm = cubic feet per minute = ft³/min; scfm = standard cubic ft/min, the volumetric flowrate of a gas corrected to standardized conditions of temperature, pressure and relative humidity (but note that there is no universally accepted set of "standard" conditions); 1 US fluid gallon = 4 US fluid quarts = 8 US pints = 128 US fluid ounces = 231 in³; 1 British imperial gallon = 277.42 in³. Values in this table are derived from Halliday and Resnick (1966).

REFERENCES

Adachi, S. & Honda, K. (2003) CFD approach to fricative sound sources. In S. Palethorpe & M. Tabain (eds.), *Proceedings of the 6th International Seminar on Speech Production* (pp. 1–6). Sydney: Macquarie University.

Alipour, F., Berry, D. A., & Titze, I. R. (2000) A finite-element model of vocal-fold vibration. *Journal of the Acoustical Society of America*, 108, 3003–3012.

Badin, P. (1989) Acoustics of voiceless fricatives: production theory and data. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, 3, 33–55.

Badin, P. & Fant, G. (1984) Notes on vocal tract computation. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, 2–3, 53–108.

Barney, A., Shadle, C. H., & Davies, P. O. A. L. (1999) Fluid flow in a dynamic mechanical model of the vocal folds and tract, I: Measurements and theory. *Journal of the Acoustical Society of America*, 105, 444–55.

Batchelor, G. K. (1967) *An Introduction to Fluid Dynamics*. Cambridge: Cambridge University Press.

Benade, A. H. (1976) *Fundamentals of Musical Acoustics*. New York: Oxford University Press.

Blazek, J. (2005) *Computational Fluid Dynamics: Principles and Applications*, 2nd edn. Amsterdam: Elsevier Science.

Busnel, R. G. & Classe, A. (1976) *Whistled Languages*. New York: Springer-Verlag.

Catford, J. C. (1977) *Fundamental Problems in Phonetics*. Bloomington, IN: Indiana University Press.

Chanaud, R. C. & Powell, A. (1965) Some experiments concerning the hole and ring tone. *Journal of the Acoustical Society of America*, 37, 902–11.

Conrad, W. A. (1985) Collapsible tube model of the larynx. In I. R. Titze & R. C. Scherer (eds.), *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control* (pp. 328–48). Denver: Denver Center for the Performing Arts.

Cranen, B. & Boves, L. (1985) Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production. *Journal of the Acoustical Society of America*, 77, 1543–51.

Cranen, B. & Boves, L. (1988) On the measurement of glottal flow. *Journal of the Acoustical Society of America*, 84, 888–900.

Davies, P. O. A. L. (1991) Program suite VOAC. Unpublished program documentation, Institute of Sound and Vibration Research, University of Southampton.

Davies, P. O. A. L., McGowan, R. S., & Shadle, C. H. (1993) Practical flow duct acoustics applied to the vocal tract. In I. R. Titze (ed.), *Vocal Fold Physiology: Frontiers in Basic Science* (pp. 93–142). San Diego: Singular Publishing Group, Inc.

Doebelin, E. O. (1983) *Measurement Systems: Application and Design*, 3rd edn. London: McGraw-Hill.

Draper, M. H., Ladefoged, P., & Whitteridge, D. (1959) Respiratory muscles in speech. *Journal of Speech and Hearing Research*, 2, 16–27.

Elder, S. A., Farabee, T. M., & Demetz, F. C. (1982) Mechanisms of flow-excited cavity tones at low Mach number. *Journal of the Acoustical Society of America*, 72, 532–49.

Erath, B. D. & Plesniak, M. W. (2006) The occurrence of the Coanda effect in pulsatile flow through static models of the human vocal folds. *Journal of the Acoustical Society of America*, 120, 1000–11.

Fant, C. G. M. (1960) *Acoustic Theory of Speech Production*. The Hague: Mouton.

Flanagan, J. L. (1972) *Speech Analysis Synthesis and Perception*, 2nd edn. Berlin: Springer-Verlag.

Flanagan, J. L. & Ishizaka, K. (1976) Automatic generation of voiceless excitation in a vocal cord-vocal tract speech synthesizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24, 163–70.

Flanagan, J. L. & Landgraf, L. L. (1968) Self-oscillating source for vocal-tract synthesizers. *IEEE Transactions on Audio and Electroacoustics*, AU-16, 57–64.

Fritzell, B., Hammarberg, B., Gauffin, J., Karlsson, I., & Sundberg, J. (1986) Breathiness and insufficient vocal fold closure. *Journal of Phonetics*, 14, 549–53.

Gauffin, J. & Sundberg, J. (1989) Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, 32, 556–65.

Goldstein, M. (1976) *Aeroacoustics*. New York: McGraw-Hill.

Guérin, B. (1983) Effects of the source–tract interaction using vocal fold models. In I. R. Titze & R. C. Scherer (eds.), *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control* (pp. 482–99). Denver: Denver Center for the Performing Arts.

Gunter, H. (2003) A mechanical model of vocal-fold collision with high spatial and temporal resolution. *Journal of the Acoustical Society of America*, 113, 994–1000.

Halliday, D. & Resnick, R. (1966) *Physics*. New York: John Wiley.

Hammarberg, B., Fritzell, B., & Schiratzki, H. (1984) Teflon injection in 16 patients with paralytic dysphonia: Perceptual and acoustic evaluation. *Journal of Speech and Hearing Disorders*, 49, 72–82.

Heinz, J. M. (1956) Fricative consonants. *MIT Research Laboratory of Electronics Quarterly Report*, Oct–Dec., 57.

Heller, H. H. & Widnall, S. E. (1970) Sound radiation from rigid flow spoilers correlated with fluctuating forces. *Journal of the Acoustical Society of America*, 47, 924–36.

Hertegård, S. & Gauffin, J. (1992) Acoustic properties of the Rothenberg mask. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, 2–3, 9–18.

Hirschberg, A., Pelorson, X., Hofmans, G. C. J., Hassel, R. R. van, & Wijnands, A. P. J. (1996) Starting transient of the flow through an in-vitro model of the vocal folds. In P. J. Davis & N. H. Fletcher (eds.), *Vocal Fold Physiology: Controlling Complexity and Chaos* (pp. 31–46). San Diego: Singular Publishing Group, Inc.

Hixon, T., Goldman, M., & Mead, J. (1973) Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung. *Journal of Speech and Hearing Research*, 16, 78–115.

Hixon, T., Mead, J., & Goldman, M. (1976) Dynamics of the chest wall during speech production: Function of the thorax, rib cage, diaphragm, and abdomen. *Journal of Speech and Hearing Research*, 19, 297–356.

Hofmans, G. C. F., Groot, G., Ranucci, M., Graziani, G., & Hirschberg, A. (2003) Unsteady flow through in-vitro models of the glottis. *Journal of the Acoustical Society of America*, 113, 1658–75.

Holger, D. K., Wilson, T. A., & Beavers, G. S. (1977) Fluid mechanics of the edgetone. *Journal of the Acoustical Society of America*, 62, 1116–28.

Holmberg, E. B., Hillman, R. E., & Perkell, J. S. (1988) Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America*, 84, 511–29.

Howe, M. S. & McGowan, R. S. (2005) Aeroacoustics of [s]. *Proceedings of the Royal Society A*, 461: 2056, 1005–28.

Iijima, H., Miki, N., & Nagai, N. (1990) Finite-element analysis of a vocal cord model with muscle of nonhomogeneous elasticity. *Journal of the Acoustical Society of Japan*, (E)11, 53–6.

Ingard, K. U. & Ising, H. (1967) Acoustic nonlinearity of an orifice. *Journal of the Acoustical Society of America*, 42, 6–17.

Ishizaka, K. & Flanagan, J. L. (1972) Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal*, 51, 1233–68.

Isshiki, N. (1964) Regulating mechanisms of vocal intensity variation. *Journal of Speech and Hearing Research*, 7, 17–29.

Jackson, P. J. B. & Shadle, C. H. (2000) Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *Journal of the Acoustical Society of America*, 108, 1421–34.

Jackson, P. J. B. & Shadle, C. H. (2001) Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. *IEEE Transactions on Speech and Audio Processing*, 9, 713–26.

Jiang, J. J., Diaz, C. E., & Hanson, D. G. (1998) Finite element modeling of vocal fold vibration in normal phonation and hyperfunctional dysphonia: Implications for the pathogenesis of vocal nodules. *Annals of Otology, Rhinology and Laryngology*, 107, 603–10.

Jiang, J. J. & Titze, I. R. (1994) Measurement of vocal fold intraglottal pressure and impact stress, *Journal of Voice*, 8, 132–44.

Karlsson, I. (1986) Glottal wave forms for normal female speakers. *Journal of Phonetics*, 14, 415–19.

Karlsson, I. (1992) Analysis and Synthesis of Different Voices with an Emphasis on Female Speech (ISRN KTH/ToM/FR–92/3–SE TRITA-TOM 1992:3). Doctoral Dissertation, Royal Institute of Technology, KTH, Department of Speech Communication and Music Acoustics, Stockholm.

Kinsler, L. E., Frey, A. R., Coppens, A. B., & Sanders, J. V. (1982) *Fundamentals of Acoustics*, 3rd edn. New York: John Wiley.

Koizumi, T. & Taniguchi, S. (1990) A novel model of pathological vocal cords and its application to the diagnosis of vocal cord polyp. In *Proceedings of the International Conference on Speech and Language Processing* (pp. 73–6). Kobe: Acoustical Society of Japan.

Krane, M. H. (2005) Aeroacoustic production of low-frequency unvoiced speech sounds. *Journal of the Acoustical Society of America*, 118, 410–27.

Krane, M. H., Sinder, D., & Flanagan, J. (1998) Approximate computational model for sound generation due to unsteady flows in pipes. Proceedings of the 16th International Congress on Acoustics and 135th meeting of the Acoustical Society of America, Seattle, WA; *Journal of the Acoustical Society of America*, 103, 2795.

Ladefoged, P. & Maddieson, I. (1996) *The Sounds of the World's Languages*. Oxford: Blackwell.

Liljencrants, J. (1991a) A translating and rotating mass model of the vocal folds. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, 1, 1–18.

Liljencrants, J. (1991b) Numerical simulations of glottal flow. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 99–104). San Diego: Singular Publication Group Inc.

Maeda, S. (1987) On the generation of sound in stop consonants. *Speech Communication Group Working Papers*, Research Laboratory of Electronics, MIT, 5, 1–14.

Massey, B. S. (1984) *Mechanics of Fluids*, 5th edn. Wokingham, UK: Van Nostrand Reinhold.

McGowan, R. S. (1988) An aeroacoustic approach to phonation. *Journal of the Acoustical Society of America*, 83, 696–704.

McGowan, R. S. (1992) Tongue-tip trills and vocal-tract wall compliance. *Journal of the Acoustical Society of America*, 91, 2903–10.

McGowan, R. S. & Howe, M. S. (2007) Compact Green's functions extend the acoustic theory of speech production. *Journal of Phonetics*, 35, 259–70.

Meyer, J. & Gautheron, B. (2006) Whistled speech and whistled languages. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn. (vol. 13, pp. 573–6). Oxford: Elsevier.

Meyer-Eppler, W. (1953) Zum Erzeugungsmechanismus der Gerauschlaute [On the generating mechanism of noise sounds]. *Zeitschrift für Phonetik*, 7, 196–212.

Mongeau, L., Franchek, N., Coker, C. H., & Kubli, R. A. (1997) Characteristics of a pulsating jet through a small modulated orifice, with application to voice production. *Journal of the Acoustical Society of America*, 102, 1121–33.

Narayanan, S. & Alwan, A. (2000) Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing*, 8, 328–44.

Negus, V. E. (1949) *The Comparative Anatomy and Physiology of the Larynx*. London: Heinemann.

Nelson, P. A. & Morfey, C. L. (1981) Aerodynamic sound production in low speed flow ducts. *Journal of Sound and Vibration*, 79, 263–89.

Ohala, J. J. (1990) Respiratory activity in speech. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 23–54). Dordrecht: Kluwer.

Pastel, L. M. P. (1987) Turbulent Noise Sources in Vocal Tract Models. MS Thesis, MIT.

Pelorson, X. (2001) On the meaning and accuracy of the pressure-flow technique to determine constriction areas within the vocal tract. *Speech Communication*, 35, 179–90.

Pelorson, X., Hirschberg, A., Hassel, R. R. van, & Wijnands, A. P. J. (1994) Theoretical and experimental study of quasi-steady flow separation within the glottis during phonation: Application to a modified two-mass model. *Journal of the Acoustical Society of America*, 96, 3416–31.

Pelorson, X., Hirschberg, A., Wijnands, A. P. J., & Bailliet, H. (1995) Description of the flow through in-vitro models of the glottis during phonation. *Acta Acustica*, 3, 191–202.

Pelorson, X., Hofmans, G. C. J., Ranucci, M., & Bosch, R. C. M. (1997) On the fluid mechanics of bilabial plosives. *Speech Communication*, 22, 155–72.

Powell, A. (1961) On the edgetone. *Journal of the Acoustical Society of America*, 33, 395–409.

Powell, A. (1962) Vortex action in edgetones. *Journal of the Acoustical Society of America*, 34, 163–6.

Recasens, D. (1991) On the production characteristics of apicoalveolar taps and trills. *Journal of Phonetics*, 19, 267–80.

Rothenberg, M. T. (1973) A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America*, 53, 1632–45.

Rothenberg, M. T. (1981) Acoustic interaction between the glottal source and the vocal tract. In K. N. Stevens & M. Hirano (eds.), *Vocal Fold Physiology* (pp. 305–23). Tokyo: University of Tokyo Press.

Sawashima, M. (1977) Fiberoptic observation of the larynx and other speech organs. In M. Sawashima & F. S. Cooper (eds.), *Dynamic Aspects of Speech Production* (pp. 31–47). Tokyo: University of Tokyo Press.

Scherer, R. C. & Guo, C.-G. (1991) Generalized translaryngeal pressure coefficients for a wide range of laryngeal configurations. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 83–90). San Diego: Singular Publishing Group, Inc.

Scherer, R. C. & Titze, I. R. (1983) Pressure–flow relationships in a model of the laryngeal airway with diverging glottis. In D. M. Bless & J. M. Abbs (eds.), *Vocal Fold Physiology: Contemporary Research and Clinical Issues* (pp. 179–93). San Diego: College-Hill Press.

Scherer, R. C., Titze, I. R., & Curtis, J. F. (1983) Pressure–flow relationships in two models of the larynx having rectangular glottal shapes. *Journal of the Acoustical Society of America*, 73, 668–76.

Scully, C. (1986) Speech production simulated with a functional model of the larynx and the vocal tract. *Journal of Phonetics*, 14, 407–13.

Scully, C. (1990) Articulatory synthesis. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 151–86). Dordrecht: Kluwer.

Shadle, C. H. (1983) Experiments on the acoustics of whistling. *The Physics Teacher*, March, 148–54.

Shadle, C. H. (1985) The Acoustics of Fricative Consonants. Ph.D. thesis, MIT Research Laboratory of Electronics, Technology Report 506.

Shadle, C. H. (1990) Articulatory-acoustic relationships in fricative consonants. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 187–209). Dordrecht: Kluwer.

Shadle, C. H. (1991) The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics*, 19, 409–24.

Shadle, C. H., Barney, A. M., & Davies, P. O. A. L. (1999) Fluid flow in a dynamic mechanical model of the vocal folds and tract, II: Implications for speech production studies. *Journal of the Acoustical Society of America*, 105, 456–66.

Shadle, C. H., Barney, A. M., & Thomas, D. W. (1991) An investigation into the acoustics and aerodynamics of the larynx. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 73–82), San Diego: Singular Publishing Group, Inc.

Shadle, C. H. & Scully, C. (1995) An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. *Journal of Phonetics*, 23, 53–66.

Shinwari, D., Scherer, R. C., DeWitt, K. J., & Afjeh, A. A. (2003) Flow visualization and pressure distribution in a model of the glottis with a symmetric and oblique divergent angle of 10 degrees. *Journal of the Acoustical Society of America*, 113, 487–97.

Sinder, D. J. (1999) Speech Synthesis Using an Aeroacoustic Fricative Model. Ph.D. thesis, Rutgers State University of New Jersey.

Slifka, J. (2003) Respiratory constraints on speech production: Starting an utterance. *Journal of the Acoustical Society of America*, 114, 3343–53.

Södersten, M. & Lindestad, P.-Å. (1990) Glottal closure and perceived breathiness during phonation in normally speaking subjects. *Journal of Speech and Hearing Research*, 33, 601–11.

Södersten, M., Lindestad, P.-Å., & Hammarberg, B. (1991) Vocal fold closure, perceived breathiness, and acoustic characteristics in normal adult speakers. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 217–24), San Diego: Singular Publishing Group, Inc.

Sondhi, M. M. & Schroeter, J. (1987) A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing* (ASSP-35:7, July), 955–67.

Stevens, K. N. (1971) Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, 50, 1180–92.

Stevens, K. N. (1993) Models for the production and acoustics of stop consonants. *Speech Communication*, 13, 367–75.

Stevens, K. N. (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Story, B. H. & Titze, I. R. (1995) Voice simulation with a body-cover model of the vocal folds, *Journal of the Acoustical Society of America*, 97, 1249–60.

Suh, J. & Frankel, S. H. (2007) Numerical simulation of turbulence transition and sound radiation for flow through a rigid glottal model. *Journal of the Acoustical Society of America*, 121, 3728–39.

Tao, C. & Jiang, J. J. (2006) Anterior-posterior biphonation in a finite element model of vocal fold vibration. *Journal of the Acoustical Society of America*, 120, 1570–7.

Tao, C., Jiang, J. J., & Zhang, Y. (2006) Simulation of vocal fold impact pressures with a self-oscillating finite-element model. *Journal of the Acoustical Society of America*, 119, 3987–94.

Teager, H. M. (1980) Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28, 599–601.

Teager, H. M. & Teager, S. M. (1983) Active fluid dynamic voice production models, or there is a unicorn in the garden. In I. R. Titze & R. C. Scherer (eds.), *Vocal Fold Physiology* (pp. 387–401). Denver: The Denver Center for Performing Arts.

Thomas, C. (ed.) (1973) *Taber's Cyclopedic Medical Dictionary*, 12th edn. Philadelphia: F. A. Davis Co.

Thomas, T. J. (1986) A finite element model of fluid flow in the vocal tract. *Computer Speech and Language*, 1, 131–52.

Titze, I. R. (1973) The human vocal cords: A mathematical model, Part I. *Phonetica*, 28, 129–70.

Titze, I. R. (1974) The human vocal cords: A mathematical model, Part II. *Phonetica*, 29, 1–21.

Titze, I. R. (1992) Phonation threshold pressure: A missing link in glottal aerodynamics. *Journal of the Acoustical Society of America*, 91, 2926–35.

Titze, I. R. (2000) *Principles of Voice Production*, 2nd printing. Iowa City, IA: National Center for Voice and Speech.

Titze, I. R. (2006) *The Myoelastic Aerodynamic Theory of Phonation*. Denver: National Center for Voice and Speech.

Titze, I. R. & Talkin, D. T. (1979) A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *Journal of the Acoustical Society of America*, 66, 60–74.

Westbury, J. R. (1983) Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *Journal of the Acoustical Society of America*, 73, 1322–36.

Zhang, C., Zhao, W., Frankel, S. H., & Mongeau, L. (2002) Computational aeroacoustics of phonation, Part II: Effects of flow parameters and ventricular folds. *Journal of the Acoustical Society of America*, 112, 2147–54.

Zhang, Z., Mongeau, L., & Frankel, S. H. (2002) Experimental verification of the quasi-steady approximation for aerodynamic sound generation by pulsating jets in tubes. *Journal of the Acoustical Society of America*, 112, 1652–63.

Zhao, W., Zhang, C., Frankel, S. H., & Mongeau, L. (2002) Computational aeroacoustics of phonation, Part I: Computational methods and sound generation mechanisms. *Journal of the Acoustical Society of America*, 112, 2134–46.

FURTHER READING

Baken, R. J. (1987) *Clinical Measurement of Speech and Voice*. Boston: College-Hill Press.

Sundberg, J. (1987) *The Science of the Singing Voice*. DeKalb, IL: Northern Illinois University Press.

Tennekes, H. & Lumley, J. L. (1972) *A First Course in Turbulence*. Cambridge, MA: MIT Press.

Versteeg, H. K. & Malalasekera, W. (2007) *An Introduction to Computational Fluid Mechanics: The Finite Volume Method*, 2nd edn. Harlow, UK: Pearson Education Ltd., Prentice-Hall.

Wagner, C., Hüttl, T., & Sagaut, P. (eds.) (2007) *Large-Eddy Simulation for Acoustics*. New York: Cambridge University Press.

3 Acoustic Phonetics

JONATHAN HARRINGTON

1 Introduction

In the production of speech, an acoustic signal is formed when the vocal organs move, resulting in a pattern of disturbance to the air molecules in the airstream that is propagated outwards in all directions eventually reaching the ear of the listener. Acoustic phonetics is concerned with describing the different kinds of acoustic signal that the movement of the vocal organs gives rise to in the production of speech by male and female speakers across all age groups and in all languages, and under different speaking conditions and varieties of speaking style. Just about every field that is covered in this book needs to make use of some aspect of acoustic phonetics. With the ubiquity of PCs and the freely available software for making spectrograms, for processing speech signals, and for labeling speech data, it is also an area of experimental phonetics that is very readily accessible.

Our knowledge of acoustic phonetics is derived from various different kinds of inquiry that can be grouped loosely into three areas that derive primarily from the contact of phonetics with the disciplines of engineering/electronics, linguistics/phonology, and psychology/cognitive science respectively.

1 *The acoustic theory of speech production*. These studies assume an idealized model of the vocal tract in order to predict how different vocal tract shapes and actions contribute to the acoustic signal (Stevens & House, 1955; Fant, 1960). Acoustic theory proposes that the excitation signal of the source can be modeled as independent from the filter characteristics of the vocal tract, an idea that is fundamental to acoustic phonetics, to formant-based speech synthesis, and to linear predictive coding which allows formants to be tracked digitally. The discovery that vowel formants can be accurately predicted by reducing the complexities of the vocal tract to a three-parameter, four-tube model (Fant, 1960) was one of the most important scientific breakthroughs in phonetics of the last century. The idea that the relationship between speech production and

acoustics is nonlinear and that, as posited by the quantal theory of speech production (Stevens, 1972, 1989; Stevens & Hanson, this volume), such discontinuities are exploited by languages in building up their sound systems, is founded upon models that relate idealized vocal tracts to the acoustic signal.

2 *Linguistic phonetics* draws upon articulatory and acoustic phonetics in order to explain why the sounds of languages are shaped the way they are. The contact with acoustic phonetics is in various forms, one of which (quantal theory) has already been mentioned. Developing models of the distribution of the possible sounds in the world's languages based on acoustic principles, as in the ground-breaking theory of adapative dispersion in Liljencrants and Lindblom (1972), is another. Using the relationship between speech production and acoustics to explain sound change as misperception and misparsing of the speech signal (Ohala, 1993, this volume) could also be grouped in this area.

3 *Variability*. The acoustic speech signal carries not only the linguistic structure of the utterance, but also a wealth of information about the speaker (physiology, regional affiliation, attitude and emotional state). These are entwined in the acoustic signal in a complex way acoustically both with each other and with background noise that occurs in almost every natural dialogue. Moreover, speech is highly context-dependent. A time slice of an acoustic signal can contain information about context, both segmental (e.g., whether a vowel is surrounded by nasal or oral sounds) and prosodic (e.g., whether the vowel is in a stressed syllable, in an accented word at the beginning or near the end of a prosodic phrase). Obviously, listeners cope for the most part effortlessly with all these multiple strands of variability. Understanding how they do so (and how they fail to do so in situations of communication difficulty) is one of the main goals of speech perception and its relationship to speech production and the acoustic signal.

As in any science, the advances in acoustic phonetics can be linked to technological development. Present-day acoustic phonetics more or less began with the invention of the sound spectrograph in the 1940s (Koenig et al., 1946). In the 1950s, the advances in vocal tract modeling and speech synthesis (Dunn, 1950; Lawrence, 1953; Fant, 1960) and a range of innovative experiments at the Haskins Laboratories (Cooper et al., 1951) using synthesis from hand-painted spectrograms underpinned the technology for carrying out many types of investigation in speech perception. The advances in speech signal processing in the 1960s and 1970s resulted in techniques like cepstral analysis and the linear prediction of speech (Atal & Hanauer, 1971) for source-filter separation and formant tracking. As a result of the further development in computer technology in the last 20–30 years and above all with the need to provide extensive training and testing material for speech technology systems, there are now large-scale acoustic databases, many of them phonetically labeled, as well as tools for their analysis (Bird & Harrington, 2001).

A recording of the production of speech with a pressure-sensitive microphone shows that there are broadly a few basic kinds of acoustic speech signal that it will be convenient to consider in separate sections in this chapter.

- *Vowels and vowel-like sounds*. Included here are sounds that are produced with periodic vocal fold vibration and a raised velum so that the airstream exits only from the mouth cavity. In these sounds, the waveform is periodic, energy is concentrated in the lower half of the spectrum, and formants, due to the resonances of the vocal tract, are prominent.
- *Fricatives and fricated sounds*. These will include, for example, fricatives and the release of oral stops that are produced with a turbulent airstream. If there is no vocal fold vibration, then the waveform is aperiodic; otherwise there is combined aperiodicity and periodicity that stem respectively from two sources at or near the constriction and due to the vibrating vocal folds. I will also include the silence that is clearly visible in oral stop production in this section.
- *Nasals and nasalized vowels*. These are produced with a lowered velum and in most cases with periodic vocal fold vibration. The resulting waveform is, as for vowels, periodic but the lowered velum and excitation of a side-branching cavity causes a set of anti-resonances to be introduced into the signal. These are among the most complex sounds in acoustic phonetics.

My emphasis will be on describing the acoustic phonetic characteristics of speech sounds, drawing upon studies that fall into the three categories described earlier. Since prosody is covered elsewhere in two chapters in this book, my focus will be predominantly on the segmental aspects of speech. I will also not cover vowel or speaker normalization in any detail, since these have been extensively covered by Johnson (2005).

2 Vowels, Vowel-Like Sounds, and Formants

2.1 The F1 × F2 plane

The acoustic theory of speech production has shown how vowels can be modeled as a straight-sided tube closed at one end (to model the closure phase of vocal fold vibration) and open at the lip end. Vowels also have a point of greatest narrowing known as a constriction location (Stevens & House, 1955; Ladefoged, 1985) that is analogous to place of articulation in consonants and that divides the tube into a back cavity and a front cavity. As Fant's (1960) nomograms show, varying the constriction location from the front to the back of the tube causes changes predominantly to the first two resonant frequencies. The changes are *nonlinear* which means that there are regions where large changes in the place of articulation, or constriction location, have a negligible effect on the formants (e.g., in the region of the soft palate) and other regions such as between the hard and soft palate where a small articulatory change can have dramatic acoustic consequences. Since there are no side-branching resonators – that is, since there is only one exit at the mouth for the air expelled from the lungs – the acoustic structure of a vowel is determined by resonances that, when combined (convolved) with the source signal, give rise to *formants*. The formants are clearly visible in a spectrographic display and they occur on average at intervals of $c/2L$, where c is the speed of sound and L the length of the vocal tract (Fant, 1973) – that is, at about 1,000 Hz intervals for an adult male vocal tract of length 17.5 cm (and with the speed of sound at 35,000 cm/s). As far as the relationship between vocal tract shape and formants are concerned, some of the main findings are:

- All parts of the vocal cavities have some influence on all formants and each formant is dependent on the entire shape of the complete system (see, e.g., Fant, 1973).
- A maximally high F1 (the first, or lowest, formant frequency) requires the main constriction to be located just above the larynx and the mouth cavity to be wide open. An increasing constriction in the mouth cavity results in a drop in F1 (see also Lindblom & Sundberg, 1971).
- A maximally high F2 is associated with a tongue constriction in the palatal region. More forward constrictions produce an increase in F3 and F4 that is due to the shortening of the front tube (Ladefoged, 1985) so that there is a progressive increase first in F2, then in F3, then in F4 as the constriction location shifts forward of the palatal zone. F2 is maximally low when the tongue constriction is in the upper part of the pharynx.
- Either a decrease of lip-opening area or an increase of the length of the lip passage produces formant lowering. Lip-protrusion has a marked effect on F3 in front vowels and on F2 in back vowels – see, e.g., Lindblom and Sundberg (1971) and Ladefoged and Bladon (1982).

The acoustic theory of speech production shows that there is a relationship between phonetic height and F1 and phonetic backness and F2, from which it follows that if vowels are plotted in the plane of the first two formant frequencies with decreasing F1 on the *x*-axis and decreasing F2 on the *y*-axis, a shape resembling the articulatory vowel quadrilateral emerges. This was first demonstrated by Essner (1947) and Joos (1948), and since then the F1 × F2 plane has become one of the standard ways of comparing vowel quality in a whole range of studies in linguistic phonetics (Ladefoged, 1971), sociophonetics (Labov, 2001), and in many other fields.

Experiments with hand-painted spectrograms using the Pattern Playback system at the Haskins Laboratories showed that vowels of different quality could be accurately identified from synthetic speech that included only the first two or only the first three formant frequencies (Delattre et al., 1955). In the 1970s and 1980s, experimental evidence of a different kind, involving an analysis of the pattern of listeners' confusions between vowels (e.g., Klein et al., 1970; Shepard, 1972) showed that perceived judgments of vowel quality depend in some way on the F1 × F2 space. The nature of these experiments varied: in some, listeners were presented with a sequence of three vowels and asked to judge whether the third is more similar to the first or to the second; or listeners might be asked to judge vowel quality in background noise. The pattern of resulting listener vowel confusions can be transformed into a spatial representation using a technique known as *multidimensional scaling* (Shepard, 1972). Studies have shown that up to six dimensions may be necessary to explain adequately the listeners' pattern of confusion between vowels (e.g., Terbeek, 1977), but also that the two most important dimensions for explaining these confusions are closely correlated with the first two formant frequencies (see also Johnson, 2004, for a discussion of Terbeek's data). These studies are important in showing that the F1 × F2 space, or some auditorily transformed version of it, represents the principal dimensions in which listeners judge vowel quality. Moreover, if listener judgments of vowel quality are primarily dependent on the F1 × F2 space, then languages should maximize the distribution between vowels in this space in order that they will be perceptually distinctive and just this has been shown in the computer simulation studies of vowel distributions in Liljencrants and Lindblom (1972).

Even in citation-form speech, the formants of a vowel are not horizontal or "steady-state" but change as a function of time. As discussed in section 2.5, much of this change comes about because preceding and following segments cause deviations away from a so-called *vowel target* (Lindblom, 1963; Stevens & House, 1963). The vowel target can be thought of as a single time point that in monophthongs typically occurs nearest near the temporal midpoint, or a section of the vowel (again near the temporal midpoint) that shows the smallest degree of spectral change and which is the part of the vowel least influenced by these contextual effects. In speech research, there is no standard method for identifying where the vowel target occurs, partly because many monophthongal vowels often have no clearly identifiable steady-state or else the steady-state, or interval that changes the least, may be different for different formants. Some researchers (e.g., Broad & Wakita, 1977; Schouten & Pols, 1979a, 1979b) apply a Euclidean-distance metric to the vowel formants to find the least-changing section of the vowel, while others estimate targets from the time at which the formants reach their maximum or minimum values (Figure 3.1). For example, since a greater mouth opening causes F1 to rise, then when a nonhigh vowel is surrounded by consonants, F1 generally rises to a maximum near the midpoint (since there is greater vocal tract constriction at the vowel margins) and so the F1-maximum can be taken to be the vowel target (see van Son & Pols, 1990 for a detailed comparison of some of the different ways of finding a vowel target).

2.2 F3 and f_0

When listeners labeled front vowels from two-formant stimuli in the Pattern Playback experiments at the Haskins Laboratories, Delattre et al. (1952) found that they preferred F2 to be higher than the F2 typically found in the corresponding natural vowels and they reasoned that this was due to the effects of F3. This preferred upwards shift in F2 in synthesizing vowels with only two formants was subsequently quantified in a further set of synthesis and labeling experiments (e.g., Carlson et al., 1975) in which listeners heard the same vowel (a) synthesized with two formants and (b) synthesized with four formants, and were asked to

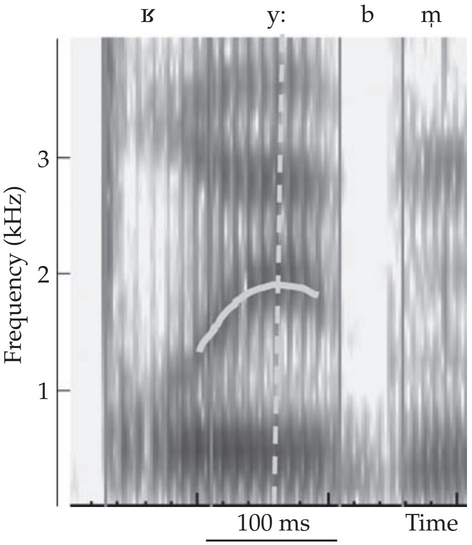


Figure 3.1 Spectrogram of the German word *drüben*, [dʁy:bm̩], produced by an adult male speaker of German. The intersection of the vertical dotted line with the hand-drawn F2 is the estimated acoustic vowel target of [y:] based on the time at which F2 reaches a maximum.

adjust F2 until (a) was perceptually as close to (b) as possible. The adjusted F2 is sometimes referred to as an *effective upper formant* or *F2-prime*.

As discussed in Strange (1999), the influence of F3 on the perception of vowels can be related to studies by Chistovich (1985) and Chistovich and Lublinskaya (1979) showing that listeners integrate auditorily two spectral peaks if their frequencies are within 3.0–3.5 Bark. Thus in front vowels, listeners tend to integrate F2 and F3 because they are within 3.5 Bark of each other, and this is why in two-formant synthesis an effective upper formant is preferred which is close to the F2 and F3 average.

Based on the experiments by Chistovich referred to above, Syrdal (1985) and Syrdal and Gopal (1986) proposed F3 – F2 in Bark as an alternative to F2 as the principal correlate of vowel backness. In their studies, a distinction between front and back vowels was based on the 3.5 Bark threshold (less for front vowels, greater for back vowels). When applied to the vowel data collected by Peterson and Barney (1952), this parameter also resulted in a good deal of speaker normalization. On the other hand, although Syrdal and Gopal (1986) show that the extent of separation between vowel categories was greater in a Bark than in a Hertz space, it has not, as far as I know, been demonstrated that F3 – F2 Bark provides a more effective distinction between vowels than F2 Bark on its own.

In the post-alveolar approximant [ɹ] and the "r-colored" vowels in American English (e.g., *bird*), F3 is very low. F3 also contributes to the unrounded/rounded distinction in front vowels in languages in which this contrast is phonemic (e.g., Vaissière, 2007). In such languages, [i] is often prepalatal, i.e., the tongue dorsum constriction is slightly forward of the hard palate and it is this difference that is responsible for the higher F3 in prepalatal French [i] compared with palatal English [i] (Wood, 1986). Moreover, this higher F3 sharpens the contrast to [y] in which F3 is low and close to F2 because of lip-rounding.

It has been known since studies by Taylor (1933) and House and Fairbanks (1953) that there is an intrinsic fundamental frequency association with vowel height: all things being equal, phonetically higher vowels tend to have higher f_0. Traunmüller (1981, 1984) has shown in a set of perception experiments that perceived vowel openness stays more or less constant if Bark-scaled f_0 and F1 increase or decrease together: his general conclusion is that perceived vowel openness depends on the difference between F1 and f_0 in Bark. In their reanalysis of the Peterson and Barney (1952) data, Syrdal and Gopal (1986) show that vowel height differences can be quite well represented on this parameter and they show that high vowels have an F1 $- f_0$ difference that is less than the critical distance of 3 Bark.

2.3 *Dynamic cues to vowels*

Many languages make a contrast between vowels that are spectrally quite similar but that differ in duration. On the other hand, there is both a length and a spectral difference in most English accents between the vowels of *heed* versus *hid* or *who'd* versus *hood*. These vowel pairs are often referred to as "tense" as opposed to "lax." Tense vowels generally occupy positions in the F1 × F2 space that are more peripheral, i.e., further away from the center than lax vowels. There is some evidence that tense–lax vowel pairs may be further distinguished based on the proportional time in the vowel at which the vowel target occurs (Lehiste & Peterson, 1961). Huang (1986, 1992) has shown in a perception experiment that the cross-over point from perception of lax [ɪ] to tense [i] was influenced by the relative position of the target (relative length of initial and final transitions) – see also Strange and Bohn (1998) for a study of the tense/lax distinction in North German. Differences in the proportional timing of vowel targets are not confined to the tense/lax distinction. For example, Australian English [i:] has a late target, i.e., long onglide (Cox, 1998) – compare for example the relative time at which the F2 peak occurs in the Australian English and Standard German [i:] in Figure 3.2.

Another more common way for targets to differ is in the contrast between monophthongs and diphthongs, i.e., between vowels with a single as opposed to two targets. Some of the earliest acoustic studies of (American English) diphthongs were by Holbrook and Fairbanks (1962) and Lehiste and Peterson (1961). Gay (1968, 1970) showed that the second diphthong target is much more likely to be undershot and reduced than the first. From this it follows that the first target and the direction of spectral change may be critical in identifying and distinguishing between diphthongs, rather than whether the second target is actually attained.

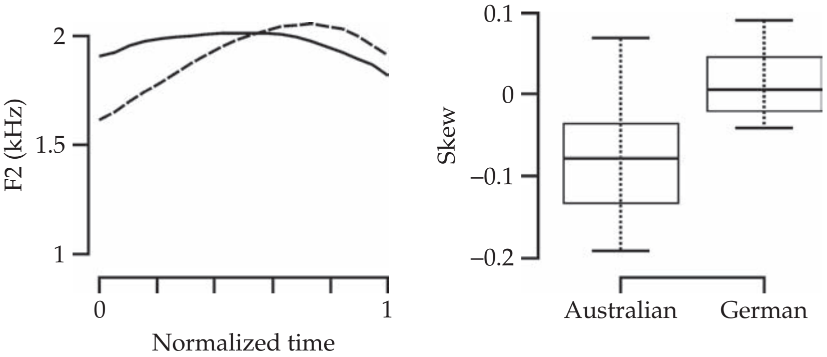


Figure 3.2 Left: Linearly time-normalized plots of F2 averaged across 57 [i:] vowels produced by a male speaker of Australian English (dotted) and across 38 [i:] vowels produced by a male speaker of Standard German (solid). All vowels were extracted from lexically stressed syllables in read sentences. Right: The distribution of these [i:] vowels on a parameter of the F2-skew for the Australian and German speakers separately, calculated with the third statistical moment (see equation (8) and section 3.1).

Gottfried et al. (1993) analyzed acoustically in an F1 × F2 logarithmic space three of the different hypotheses for diphthong identification discussed in Nearey and Assmann (1986). These were that (a) both targets, (b) the onset plus the rate of change of the spectrum, and (c) the onset plus the direction, are critical for diphthong identification. The results of an analysis of 768 diphthongs provided support for all three hypotheses, with the highest classification scores obtained from (a), the dual target hypothesis.

Many studies in the *Journal of the Acoustical Society of America* in the last 30 years have been devoted to the issue of whether vowels are sufficiently distinguished by information confined to the vowel target. It seems evident that the answer must be no (Harrington & Cassidy, 1994; Watson & Harrington, 1999), given that, as discussed above, vowels can vary in length, in the relative timing of the target, and in whether vowels are specified by one target or two. Nevertheless, the case for vowels being “dynamic” in general was made by Strange and colleagues based on two sets of data. In the first, Strange et al. (1976) found that listeners identified vowels more accurately from CVC than from isolated V syllables; and in the second, vowels were as well identified from so-called silent center syllables, in which the middle section of CVC syllable had been spliced out leaving only transitions, as from the original CVC syllables (Strange et al., 1983). Both sets of experiments led to the conclusion that there is at least as much information for vowel identification in the (dynamically changing) transitions as at the target. Compatibly, human listeners make more errors in identifying vowels from static (steady-state) synthetic vowels compared with synthetic vowels that include formant change (e.g., Hillenbrand & Nearey, 1999) and a number of acoustic experiments have shown that vowel classification is improved using information other than just at the vowel target (e.g., Hillenbrand et al., 2001; Huang, 1992; Zahorian & Jagharghi, 1993).

2.4 Whole-spectrum approaches to vowel identification

Although no one would dispute that the acoustic and perceptual identification of vowels is dependent on formant frequencies, many have also argued that there is much information in the spectrum for vowel identity apart from formant center frequencies. Bladon (1982) and Bladon and Lindblom (1981) have advocated a whole-spectrum approach and have argued that vowel identity is based on gross spectral properties such as auditory spectral density. More recently, Ito et al. (2001) showed that the tilt of the spectrum can cue vowel identity as effectively as F2. On the other hand, manipulation of formant amplitudes was shown to have little effect on listener identification of vowels in both Assmann (1991) and Klatt (1982); and Kiefte and Kluender's (2005) experiments show that, while spectral tilt may be important for identifying steady-state vowels, its contribution is less important in more natural speaking contexts. Most recently, in Hillenbrand et al. (2006), listeners identified vowels from two kinds of synthesized stimuli. In one, all the details of the spectrum were included while in the other, the fine spectral structure was removed preserving information only about the spectral peaks. They found that identification rates were higher from the first kind, but only marginally so (see also Molis, 2005). The general point that emerges from these studies is that formants undoubtedly provide the most salient information about vowel identity in both acoustic classification and perception experiments and that the rest of the shape of the spectrum may enhance these distinctions (and may provide additional information about the speaker which could, in turn, indirectly aid vowel identification).

Once again, the evidence that the primary information for vowel identification is contained in the formant frequencies emerges when data reduction techniques are applied to vowel spectra. In this kind of approach (e.g., Klein et al., 1970; Pols et al., 1973), energy values are summed in auditorily scaled bands. For example, the spectrum up to 10 kHz includes roughly 22 bands at intervals of 1 Bark, so if energy values are summed in each of these Bark bands, then each vowel's spectrum is reduced to 22 values, i.e., to a point in 22-dimensional space. The technique of *principal components analysis* (PCA) finds new axes through this space such that the first axis explains most of the variance in the original data, the second axis is orthogonal to the first, the third is orthogonal to the second, and so on. Vowels can be distinguished just as accurately from considerably fewer dimensions in a PCA-rotated space of these Bark-scaled filter bands as from the original high-dimensional space. But also, one of the important findings to emerge from this research is that the first two dimensions are often strongly correlated with the first two formant frequencies (Klein et al., 1970). (This technique has also been used in child speech in which formant tracking is difficult – see Palethorpe et al., 1996.)

This relationship between a PCA-transformed Bark space and the formant frequencies is evident in Figure 3.3 in which PCA was applied to Bark bands

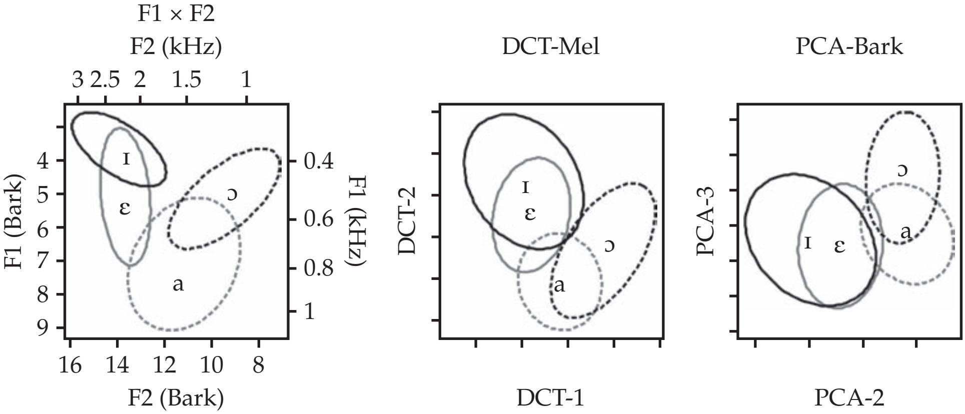


Figure 3.3 95% confidence ellipses for four lax vowels extracted from lexically stressed syllables in read sentences and produced by an adult female speaker of Standard German in the planes of F2 x F1 in Bark (left), the first two DCT coefficients (center), and two dimensions derived after applying PCA to Bark bands calculated in the 200–4,000 Hz range (right). The numbers of tokens in the categories [ɪ, ɛ, a, ɔ] were 85, 41, 63, and 16 respectively.

spanning the 200–4,000 Hz range in some German lax vowels [ɪ, ɛ, a, ɔ]. Spectra were calculated for these vowels with a 16-ms window at a sampling frequency of 16 kHz and energy values were calculated at one Bark intervals over the frequency range 200–4,000 Hz, thereby reducing each spectrum to a point in a 15-dimensional space. The data were then rotated using PCA. As Figure 3.3 shows, PCA-2 is similar to F1 in separating vowels in terms of phonetic height while [a] and [ɔ] are separated almost as well on PCA-3 as on F2. Indeed, if this PCA space were further rotated by about 45 degrees clockwise, then there would be quite a close correspondence to the distribution of vowels in the F1 × F2 plane, as Klein et al. (1970) had shown.

We arrive at a similar result in modeling vowel spectra with the discrete cosine transformation (DCT; Zahorian & Jagharghi, 1993; Watson & Harrington, 1999; Palethorpe et al., 2003). As discussed in more detail in section 3.1 below, the result of applying a DCT to a spectrum is a set of DCT coefficients that encode properties of the spectrum's shape. When a DCT analysis is applied to vowel spectra, then the first few DCT coefficients are often sufficient for distinguishing between vowels, or the distinction is about as accurate as from formant frequencies (Zahorian & Jagharghi, 1993). In Figure 3.3, a DCT analysis was applied to the same spectra in the 200–4,000 Hz range that were subjected to PCA analysis. Before applying the DCT analysis, the frequency axis of the spectra was converted to the auditory mel scale. Again, a shape that resembles the F1 × F2 space emerges when these vowels are plotted in the plane of DCT-1 × DCT-2. (It should be mentioned here that DCT coefficients derived from mel spectra are more or less the same as mel-frequency cepstral coefficients that are often used in automatic speech recognition – see, e.g., Nossair & Zahorian, 1991; and Milner & Shao, 2006.)

2.5 Vowel reduction

It is important from the outset to make a clear distinction between phonological and phonetic vowel reduction: the first is an obligatory process in which vowels become weak due to phonological and morphological factors, as shown by the alternation between /eɪ/ and /ə/ in *Canadian* and *Canada* in most varieties of English. In the second, vowels are phonetically modified because of the effects of segmental and prosodic context. Only the second is of concern here.

Vowel reduction is generally of two kinds: *centralization* and *coarticulation*, which together are sometimes also referred to as *vowel undershoot*. The first of these is a form of paradigmatic vowel reduction in which vowels become more schwa-like and the entire vowel space shrinks as vowels shift towards the center. Coarticulation is syntagmatic: here there are shifts in vowels that can be more directly attributed to the effects of preceding and following context.

The most complete account of segmental reduction is Lindblom's (1990, 1996) model of hyper- and hypoarticulation (H&H) in which the speaker plans to produce utterances that are sufficiently intelligible to the listener, i.e., a speaker economizes on articulatory effort but without sacrificing intelligibility. Moreover, the speaker makes a moment-by-moment estimate of the listener's need for signal information and adapts the utterance accordingly. When the listener's needs for information are high, then the talker tends to increase articulatory effort (hyperarticulate) in order to produce speech more clearly. Thus when words are excised from a context in which they are difficult to predict from context, listeners find them easier to identify than when words are spliced out of predictable contexts (Lieberman, 1963; Hunnicutt, 1985, 1987). Similarly, repeated words are shorter in duration and less intelligible when spliced out of context than the same words produced on the first occasion (Fowler & Housum, 1987).

As far as vowels are concerned, hyperarticulated speech is generally associated with less centralization and less coarticulation, i.e., an expansion of the vowel space and/or a decrease in coarticulatory overlap. There is evidence for both of these in speech that is produced with increased clarity (e.g., Picheny et al., 1986; Moon & Lindblom, 1994; Smiljanić & Bradlow, 2005). Additionally, Wright (2003) has demonstrated an H&H effect even when words are produced in isolation. He showed that the vowels of words that are "hard" have an expanded vowel space relative to "easy" words. The distinction between hard and easy takes account both of the statistical frequency with which words are used in the language and the lexical *neighborhood density*: if a word has a high value on neighborhood density, then there are very many other words which are phonemically identical to it based on substituting any one of the word's phonemes. Easy words are those which are high in frequency and low in neighborhood density. By contrast, hard words occur infrequently in the language and are confusable with other words, i.e., have high neighborhood density.

There have been several recent studies exploring the relationship between redundancy and hypoarticulation (van Son & Pols, 1999, 2003; Bybee, 2000; Bell et al., 2003; Jurafsky et al., 2003; Munson and Soloman, 2004; Aylett & Turk, 2006). The study by Aylett and Turk (2006) made use of a large corpus of citation-form speech including 50,000 words from each of three male and five female speakers. Their analysis of F1 and F2 at the vowel midpoint showed that vowels with high predictability were significantly centralized relative to vowels in less redundant words.

Many studies have shown an association between vowel reduction and various levels of the stress hierarchy (Fry, 1965; Edwards, Beckman, & Fletcher, 1991; Fourakis, 1991; Sluijter & van Heuven, 1996; Sluijter et al., 1997; Harrington et al., 2000; Hay et al., 2006) and with rate (e.g., Turner et al., 1995; Weismer et al., 2000). The rate effects on the vowel space are not all consistent (van Son & Pols, 1990, 1992; Stack et al., 2006; Tsao et al., 2006) not only because speakers do not all increase rate by the same factor, but also because there can be articulatory reorganization with rate changes.

As far as syntagmatic coarticulatory effects are concerned, Stevens and House (1963) found that consonantal context shifted vowel formants towards more central values, with the most dramatic influence being on F2 due to place of articulation. More recently, large shifts due to phonetic context have been reported in Hillenbrand et al. (2001) for an analysis of six men and six women producing eight vowels in CVC syllables. At the same time, studies by Pols (e.g., Schouten & Pols, 1979a, 1979b) show that the size of the influence of the consonant on vowel targets is considerably less than the displacement to vowel targets caused by speaker variation and in the study by Hillenbrand et al. (2001), consonant environment had a significant, although small, effect on vowel intelligibility. Although consonantal context can cause vowel centralization, Lindblom (1963), Moon and Lindblom (1994), and van Bergem (1993) emphasize that coarticulated vowels do not necessarily *centralize* but that the formants shift in the direction of the loci of the flanking segments.

Lindblom and Studdert-Kennedy (1967) showed that listeners compensate for the coarticulatory effects of consonants on vowels. In their study, listeners identified more tokens from an /ɪ–ʊ/ continuum as /ɪ/ in a /w_w/ context than in a /j_j/ context. This comes about because F2 lowering is a cue not only for /ʊ/ as opposed to /ɪ/, but also because F2 lowering is brought about by the coarticulatory effects of the low F2 of /w/. Thus, because of this dual association of F2 lowering, there is a greater probability of hearing the same token as /ɪ/ in a /w_w/ than in a /j_j/ context if listeners factor out the proportion of F2 lowering that they assume to be attributable to /w/-induced coarticulation.

Based on an analysis of the shift in the first three formants of vowels in /bVb, dVd, gVg/ contexts, Lindblom (1963) developed a mathematical model of vowel reduction in which the extent of vowel undershoot was exponentially related to vowel duration. The model was founded on the idea that the power, or articulatory effort, delivered to the articulators remained more or less constant, even if other factors – such as consonantal context, speech tempo, or a reduction of stress – caused vowel duration to decrease. The necessary outcome of the combination of a constant articulatory power with a decrease in vowel duration is, according to this model, vowel undershoot (since if the power to the articulators remains the same, there will be insufficient time for the vowel target to be produced).

The superposition model of Broad and Clermont (1987) is quite closely related to Lindblom's (1963) model, at least as far as the exponential relationship between undershoot and duration is concerned (see also van Son, 1993: ch. 1 for a very helpful discussion of the relationship between these two models). Their model is based on the findings of Broad and Fertig (1970), who showed that formant contours in a CVC syllable can be modeled as the sum of $f(t) + g(t) + V_T$ where $f(t)$ and $g(t)$ define as a function of time the CV and VC formant transitions respectively and V_T is the formant frequency at the vowel target. This superposition model is also related to Öhman's (1967) numerical model of coarticulation based on VCV sequences in which the shape of the tongue at a particular point in time was modeled as a linear combination of a vowel shape, a consonant shape, and a coarticulatory weighting factor.

In one version of Broad and Clermont (1987), the initial and final transition functions, $f(t)$ and $g(t)$, are defined as:

$$f(t) = K_i(T_v - L_i)e^{-\beta_i t} \qquad (1)$$

$$g(t) = K_f(T_v - L_f)e^{\beta_f(t-D)} \qquad (2)$$

where K (i.e., K_i and K_f for initial and final transitions respectively) is a consonant-specific scale-factor, $T_v - L_i$ and $T_v - L_f$ are the target–locus distances in CV and VC transitions respectively, β is a time-constant that defines the rate of transition, and D is the total duration of the CVC transition. Just as in Lindblom (1963), the essence of (1) and (2) is that the greater the duration, the more the transitions approach the vowel target.

Figure 3.4 shows an example of how an F2 transition in a syllable /dɪd/ could be put together with (1) and (2) (and using the parameters in Table VI of Broad & Clermont, 1987). The functions $f(t)$ and $g(t)$ define F2 of /dɪ/ and /ɪd/ as a function of time. To get the output for /dɪd/, $f(t)$ and $g(t)$ are summed at equal points in time and then these are added to the vowel target, which in this example is set to 2,276 Hz. Notice firstly that the initial and final transitions are negative and asymptote to zero, so that when they are added to the vowel target, their combined effect on the formant contour is least at the vowel target and progressively greater towards the syllable margins. Moreover, the model incorporates the idea from Broad and Fertig (1970) that initial and final transitions can influence each other at *all* time points, but that importantly the mutual influence of the initial on the final transitions progressively wanes for time points further away from the target.

In the first row of Figure 3.4, the duration of the CVC syllable is sufficient for the target to be almost attained. In row 2, the CVC has a duration that is 100 ms

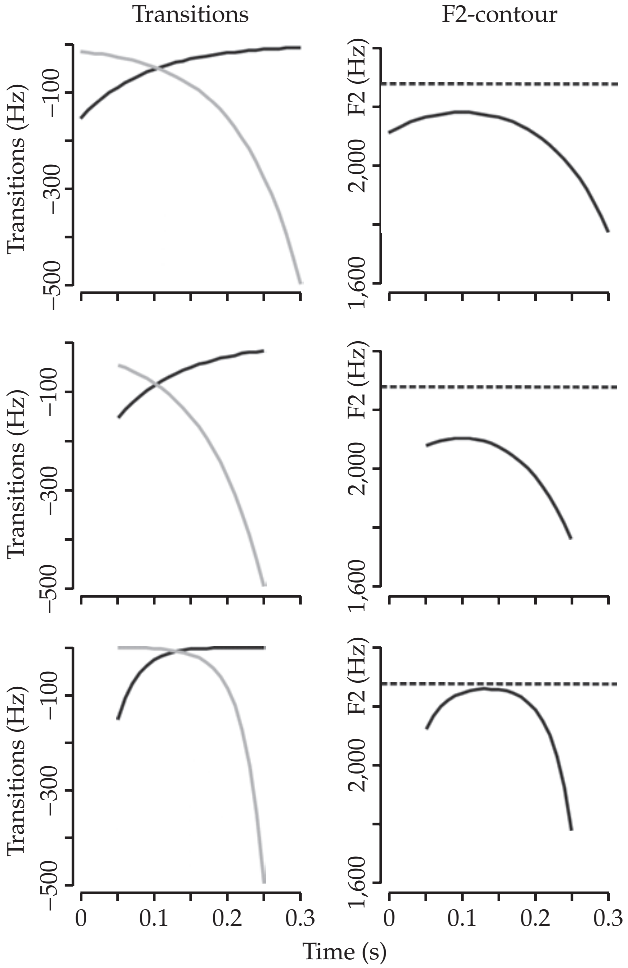


Figure 3.4 An implementation of the equations (1) and (2) for constructing an F2-contour appropriate for the context [dɪd] using the parameters given in Table VI of Broad and Clermont (1987). Left: the values of the initial [dɪ] (black) and final [ɪd] (gray) transitions. Right: the corresponding F2 contour that results when the transitions on the left are summed and added to the vowel target shown as horizontal dotted line. Row 1: vowel duration = 300 ms. Row 2: the same parameters are used as in row 1, but the duration is 100 ms less resulting in greater undershoot (shown as the extent by which the contour on the right falls short in frequency of the horizontal dotted line). Row 3: the same parameters as in row 2, except that the transition rates, defined by β in equations in (1) and (2), are faster.

less than in row 1. The transition functions are *exactly the same*, but now there is less time for the target to be attained and as a result there is greater undershoot – specifically, the vowel target is undershot by about another around 100 Hz. This is the sense of undershoot in Lindblom (1963): the parameters controlling the transitions do not change (because the force to the articulators is unchanged) and the extent of undershoot is predictable from the durational decrease.

However, studies of speech production have shown that speakers can and do increase articulatory velocity when vowel duration decreases (Kuehn & Moll, 1976; Kelso et al., 1985; Beckman et al., 1992). As far as formulae (1) and (2) are concerned, this implies that the time constants can change to speed up the transition (see also Moon & Lindblom, 1994). An example of changing the time constants and hence the rate of transition is shown in the third row of Figure 3.4: in this case, the increase in transition speed (decrease in the time constants) easily offsets the 100 ms shorter duration compared with row 1 and the target is very nearly attained.

2.6 F2 locus and consonant place of articulation

The idea that formant transitions provide cues to place of articulation can be traced back to Potter, Kopp, and Green (1947) and to the perception experiments carried out in the 1950s with hand-painted spectrograms using two-formant synthesis at the Haskins Laboratories (Liberman et al., 1954; Delattre et al., 1955). These perception experiments showed that place of articulation could be distinguished by making F2 point to a "locus" on the frequency axis close to the time of the stop release. The Haskins Laboratories experiments showed that /b/ and /d/ were optimally perceived with loci at 720 Hz and 1,800 Hz respectively. An acceptable /g/ could be synthesized with the F2 locus as high as 3,000 Hz before nonback vowels, but no acceptable locus could be found for /g/ before back vowels.

In the 1960s–1980s various acoustic studies (Lehiste & Peterson, 1961; Öhman, 1966; Fant, 1973; Kewley-Port, 1982) explored whether there was evidence for an F2 locus in natural speech data. In general, these studies did not support the idea of an invariant locus; they also showed the greatest convergence towards a locus frequency for /d/.

F3 transitions can also provide information about stop place and in particular for separating alveolars from velars (Öhman, 1966; Fant, 1973; Cassidy & Harrington, 1995). As the spectrographic study by Potter et al. (1947) had shown, F2 and F3 at the vowel onset seem to originate from a mid-frequency peak that is typical of velar bursts: for example, F2 and F3 are much closer together in frequency at vowel onset following a velar than an alveolar stop, as the spectrograms in Figure 3.5 show.

In the last 15 years or so, a number of studies in particular by Sussman and Colleagues (e.g., Sussman, 1994; Sussman et al., 1993, 1995; Modarresi et al., 2005) have used so-called *locus equations* as a metric for investigating the relationship between place of articulation and formant transitions. The basic form of the locus equation is given in (3) and it is derived from another observation in Lindblom (1963) that the formant values at the vowel onset (F_{ON}) and at the vowel target (F_T) are linearly related:

$$F_{ON} = \alpha F_T + c \tag{3}$$

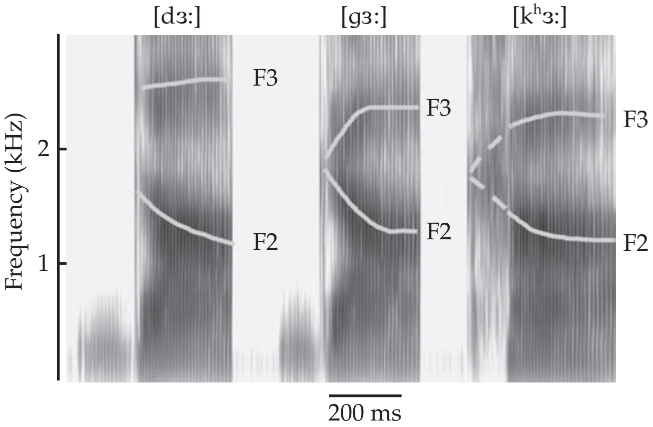


Figure 3.5 Spectrograms, male speaker of Australian English, extracted from isolated productions of the nonword *dird* and the words *gird* and *curd* (Australian English is non-rhotic). The F2 and F3 transitions were traced by hand from the onset of periodicity in the first two words, and from the burst release in *curd*.

Krull (1989) showed that the slope, α, could be used to measure the extent of V-on-C coarticulation. The theory behind this is as follows. The more that a consonant is influenced by a vowel, the less the formant transitions converge to a common locus and the greater the slope in the plane of vowel onset frequency by vowel target frequency. This is illustrated for two hypothetical cases of F2 transitions in the syllables [bɛ] and [bo] in Figure 3.6. On the left, the F2 transitions converge to a common locus: in this case, F2 onset is completely unaffected by the following vowel (the anticipatory V-on-C coarticulation at the vowel onset is zero). From another point of view, the vowel target could not be predicted from a knowledge of the vowel onset (since the vowel onsets are the same for [bɛ] and [bo]). On the right is the case of *maximum* coarticulation: in this case, the V-on-C coarticulation is so strong that there is no convergence to a common locus and the formant onset is the same as the formant target (i.e., the formant target is completely predictable for any known value of formant onset). In the panels on the right, these hypothetical data were plotted in the formant target by formant onset plane. The line that connects these points is the *locus equation*, and it is evident that the two cases of zero and maximal coarticulation differ in the lines' slopes which are 0 and 1 respectively.

It is possible to re-write (3) in terms of the locus frequency, L (Harrington & Cassidy, 1999):

$$F_{ON} = \alpha F_T + L(1 - \alpha) \qquad (4)$$

From (4), it becomes clear that when α is zero, $F_{ON} = L$ (i.e., the vowel onset equals the locus frequency as in Figure 3.6 left) and when α is 1, $F_O = F_T$ (i.e., the vowel

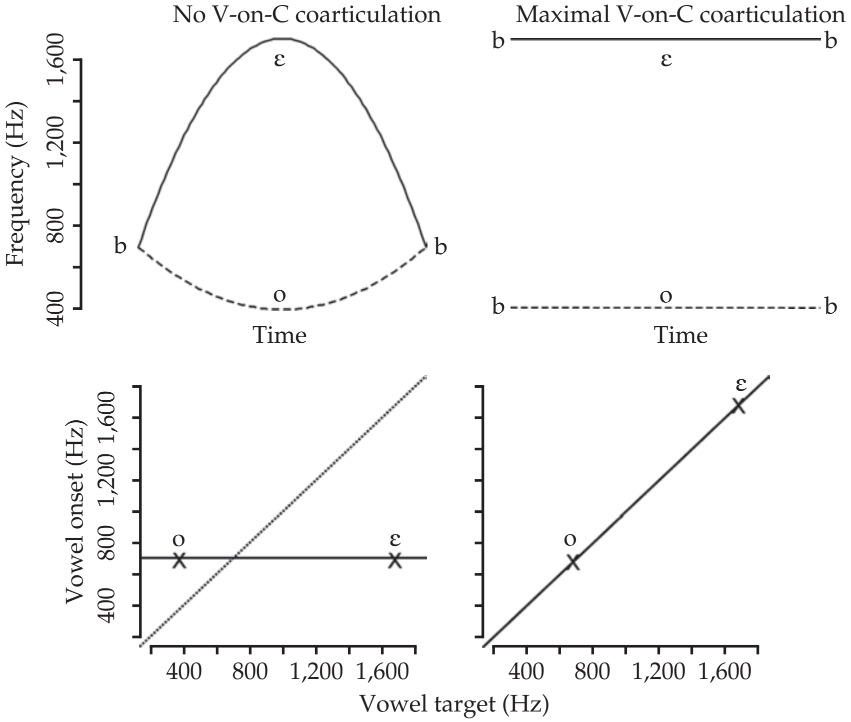


Figure 3.6 Hypothetical F2 trajectories of [bɛb] (solid) and [bob] (dashed) when there is no V-on-C coarticulation at the vowel onset/offset (left) and when V-on-C coarticulation is maximal (right). Row 1: the trajectories as a function of time. Row 2: a plot of the F2 values in the plane of the vowel target by vowel onset for the data in the first row. The solid line is analogous to the locus equation. The locus frequency can be obtained either from equation (5) or from the point at which the locus equation intersects the dotted line, $F2_{Target} = F2_{Onset}$ (this dotted line overlaps completely with the locus equation on the right meaning that for these data, there is no locus frequency).

onset equals the vowel target as in Figure 3.6 right). More importantly, the fact that the slope varies between 0 and 1 can be used to infer the magnitude of V-on-C coarticulation. This principle is illustrated for some /dVd/ syllables produced by an Australian English male speaker in Figure 3.7.

The V in this case varied over almost all the monophthongs of Australian English and the plots in the first row are F2 as a function of time, showing the same F2 data synchronized firstly at the vowel onset on the left and at the vowel offset on the right. These plots of F2 as a function of time in row 1 of Figure 3.7 show a greater convergence to a common F2 onset frequency for initial compared with final transitions. From this it can be inferred that the size of V-on-C coarticulation is less in initial /dV/ than in final /Vd/ sequences (i.e., /d/ *resists* coarticulatory influences from the vowel to a greater extent in syllable-initial than

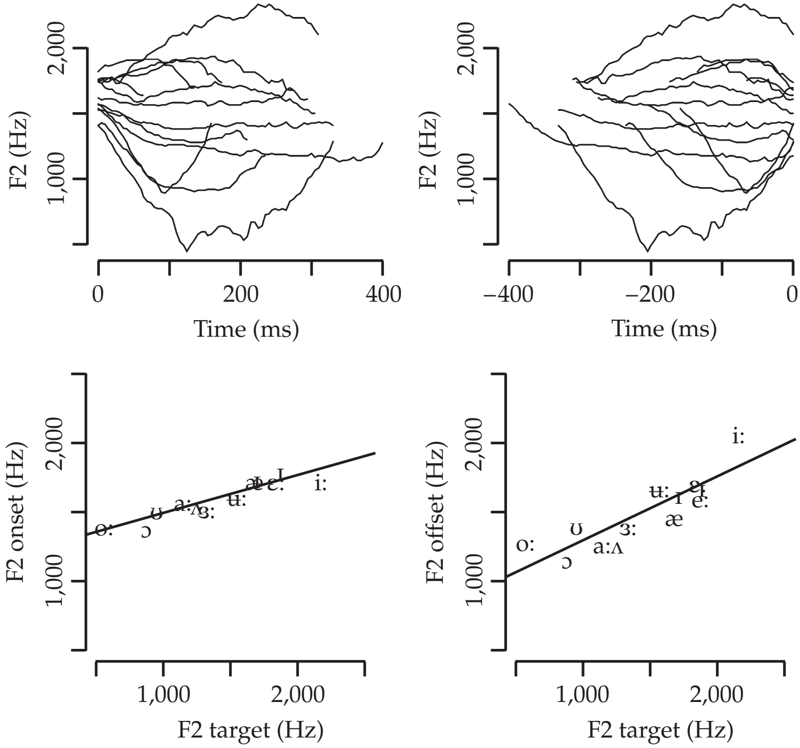


Figure 3.7 Row 1: F2 trajectories of isolated /dVd/ syllables produced by an adult male speaker of Australian English and synchronized (t = 0 ms) at the vowel onset (left) and at the vowel offset (right). There is one trajectory per monophthong (n = 14). Row 2: corresponding locus equations with the vowel labels marked at the F2 target × F2 onset positions. The slopes and intercepts of the locus equations are respectively 0.27, 1,220 Hz (initial transitions, left) and 0.46, 829 Hz (final transitions, right).

in syllable-final position). These positional differences are consistent with various other studies showing less coarticulation for initial /d/ compared to final /d/ (Krull, 1989; Sussman et al., 1993).

In Figure 3.7 row 2, F2 at the vowel target has been plotted as a function of the F2 onset and F2 offset respectively and locus equations were calculated by drawing a straight line through each of the two scatters separately. The slope of the regression line (i.e., of the locus equation) is higher for the final /Vd/ than for the initial /dV/ transitions, which is commensurate with the interpretation in this figure that there is greater accommodation of final /d/ than initial /d/ to the vowel.

A locus equation like any straight line in an x–y plane, has, of course, both a slope and an intercept and various studies (e.g., Fowler, 1994; Sussman, 1994; Chennoukh et al., 1997) have shown how different places of articulation have different values on slopes and intercepts together (the information from both the slope and intercept together is sometimes called a *second-order locus equation*). Whereas the slope says something about the extent of V-on-C coarticulation, the intercept encodes information about the best estimate of the locus frequency weighted by the slope. From (3) and (4) it is evident that the intercept, c, locus frequency, L, and slope, α, are related by $c = L(1 - \alpha)$. Thus the locus frequency can be estimated from the locus equation intercept and slope:

$$L = c/(1 - \alpha) \tag{5}$$

For the initial /dV/ data (Figure 3.7, row 1, left), the intercept and slope are given by 1,220.3 Hz and 0.27 so the best estimate of the F2 locus is 1,220.3/(1 – 0.27) = 1,671 Hz which is indeed close to the frequency towards which the F2 transitions in row 1 of Figure 3.7 seem to converge.

Some of the main findings to emerge from locus equation (LE) studies in recent years are:

- The data points in the plane of F2 onset × F2 target are tightly clustered about a locus equation and the locus equation parameters (intercept, slope) differ for different places of articulation (Krull, 1989; various studies by Sussman and colleagues referred to earlier).
- Alveolars have the lowest LE slopes which, as discussed earlier, implies that they are least affected by V-on-C coarticulation (e.g., Krull, 1989). They also usually have higher intercepts than bilabials, which is to be expected given the relationship in (5) and the other extensive evidence from perception experiments and acoustic analyses that the F2 locus of alveolars is higher than that of labials.
- It is usually necessary to calculate separate locus equations for velar stops before front and back vowels (Smits et al., 1996a, 1996b) because of the considerable variation in F2 onset frequencies of velars due to the following vowel (or, if velar consonants are pooled across vowels, then they tend to have the highest slopes, as the acoustic and electropalatographic (EPG) data in Tabain, 2000 have shown).
- Subtle place differences involving the same articulator cannot easily be distinguished using LE parameters (Krull et al., 1995; Tabain & Butcher, 1999; Tabain, 2000).
- There is controversy about whether LE parameters vary across manner of articulation (Fowler, 1994) and voicing (Engstrand & Lindblom, 1997; but see Modarresi et al., 2005). For example, Sussman (1994) reports roughly similar slopes for /d, z, n/; however, in an electropalatographic analysis of CV onsets, Tabain (2000) found that LE parameters distinguished poorly within fricatives.
- As already mentioned, Krull (1989) has shown that locus equations can be very useful for analyzing the effects of speaking style: in general, spontaneous speech is likely to have lower slopes because of the greater V-on-C coarticulation than citation-form speech. However in a more recent study, van Son and

Pols (1999) found no difference in intercepts and slopes comparing read with spontaneous speech in Dutch.

- While Chennoukh et al. (1997) relate locus equations to articulatory timing using the distinctive region model (DRM) of area functions (Carré & Mrayati, 1992), none of the temporal phasing measures in VCV sequences using movement data in Löfqvist (1999) showed any support for the assumption that the LE slope serves as an index of the degree of coarticulation between the consonant and the vowel.
- While Sussman et al. (1995) have claimed that "the locus equation metric is attractive as a possible context-independent phonemic class descriptor and a logical alternative to gestural-related invariance notions", the issue concerning the auditory or cognitive status of LEs has been disputed (e.g., Brancazio & Fowler, 1998; Fowler, 1994).

Finally, and this is particularly relevant to the last point above, the claim has been made that it is possible to obtain "perfect classification accuracy (100%) for place of articulation" (Sussman et al., 1991) from LE parameters. However, it is important to recognize that LE parameters themselves are generalizations across multiple data points (Fowler, 1994; Löfqvist, 1999). Therefore, the perfect classification accuracy in distinguishing between three places of articulation is analogous to finding no overlap between three vowel categories that had been averaged by category across each speaker (as in classifying 10 [i], 10 [u], and 10 [a] points in an F1 × F2 space, where each point is an average value per speaker). Seen from this point of view, it is not that entirely surprising that 100 percent classification accuracy could be obtained, especially for citation-form speech data.

2.7 Approximants

Voiced approximants are similar in acoustic structure to vowels and diphthongs and are periodic with F1–F3 occurring in the 0–4,000 Hz spectral range. As a class, approximants can often be distinguished from vowels by their lower amplitude and from each other by the values of their formant frequencies. Figure 3.8 shows that for the sonorant-rich sentence "Where were you while we were away?" there are usually dips in two energy bands that have been proposed by Espy-Wilson (1992, 1994) for identifying approximants.

Typical characteristics for approximant consonants that have been reported in the literature (and of which some are shown in the spectrogram in Figure 3.8) are as follows:

- [w] has F1 and F2 close together and both low in frequency. The ranges reported for American English are 300–400 Hz for F1 and 600–800 Hz for F2 (e.g., Lehiste, 1964; Mack & Blumstein, 1983). [w], like labials and labial-velars, has a low F2 and this is one of the factors that contributes to sound changes involving these segments (see Ohala & Lorentz, 1977, for further details).
- [j] like [i] has a low F1 and a high F2 – see Figure 3.8.

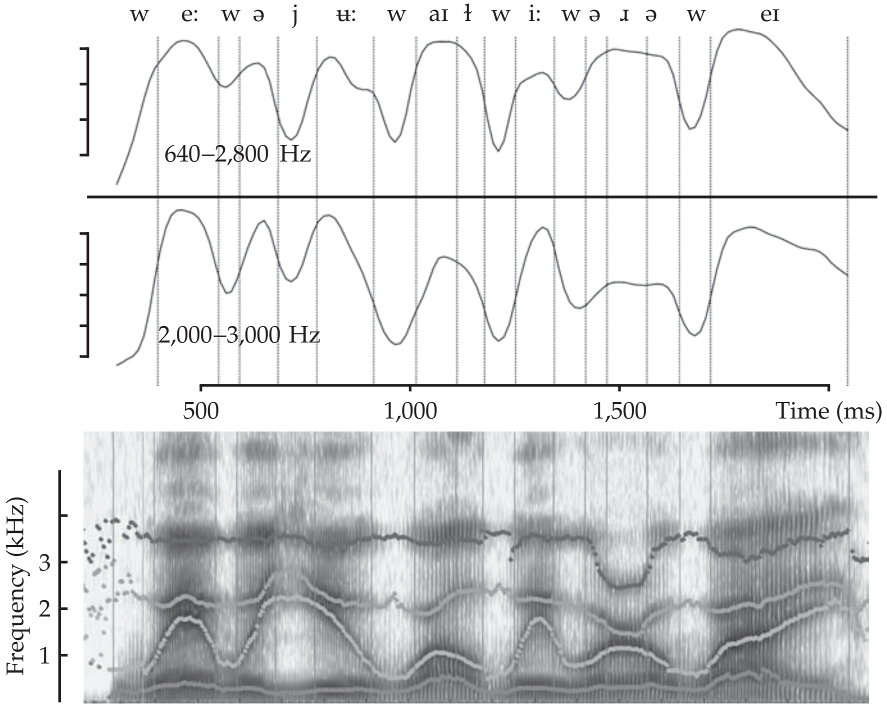


Figure 3.8 Summed energy values in two frequency bands and the first four formant frequencies superimposed on a spectrogram of the sonorant-rich sentence “Where were you while we were away?” produced by an adult male Australian English speaker. (Adapted from Harrington & Cassidy, 1999, p. 110, figure 4.33, with kind permission of Springer Science and Business Media)

- American English /r/ and the post-alveolar approximant that is typical in Southern British English have a low F3 typically in the 1,300–1,800 Hz range (Lehiste, 1964; Nolan, 1983), which is likely to be a front cavity resonance (Fant, 1960; Stevens, 1998; Espy-Wilson et al., 2000; Hashi et al., 2003).
- /l/ when realized as a so-called clear [l] in syllable-initial position in many English varieties has F1 in the 250–400 Hz range and a variable F2 that is strongly influenced by the following vowel (Nolan, 1983). F3 in “clear” realizations of /l/ may be completely canceled by an anti-resonance due to the shunting effects of the mouth cavity behind the tongue blade. The so-called dark velarized /l/ that occurs in syllable-final position in many English varieties has quite a different formant structure which, because it is produced with velarization and raising of the back of the tongue, resembles a high back round vowel in many respects: in this case, F2 can be as low as 600–900 Hz (Lehiste, 1964; see also the final /l/ in *while* in Figure 3.8). Bladon and Al-Bamerni (1976) showed that /l/ varies in clarity depending on various prosodic factors, including syllable position; and also that dark realizations of /l/ were

much less prone to coarticulatory influences from adjacent vowels compared with clear /l/.

- Compared with the other approximants, American English /l/ is reported as having longer and faster transition (Polka & Strange, 1985).
- /l/ sometimes has a greater spectral discontinuity with a following vowel that is caused by the complete alveolar closure: that is, there is often an abrupt F1-transition from an /l/ to a following vowel which is not in evidence for the other three approximants (O'Connor et al., 1957).
- In American English, [w] can sometimes be distinguished from [b] because of its slower transition rate into a following vowel (e.g., Mack & Blumstein, 1983).

3 Obstruents

Fricatives are produced with a turbulent airstream that is the result of a jet of air being channelled at high speed through a narrow constriction and hitting an obstacle (see Shadle, this volume). For [s] and [ʃ] the obstacles are the upper and lower teeth respectively; for [f] the obstacle is the upper lip and for [x] it is the wall of the vocal tract (Johnson, 2004). The acoustic consequence of the turbulent airstream is aperiodic energy. In Figure 3.9, the distinction between the fricatives and sonorants in the utterance "Is this seesaw safe?" can be seen quite easily from the aperiodic energy in fricatives that is typically above 1,000 Hz. Fricatives are produced with a noise source that is located at or near the place of maximum constriction and their spectral shape is strongly determined by the length of the cavity in front of the constriction – the back cavity makes scarcely any contribution to the spectrum since the coupling between the front and back cavities is weak (Stevens, 1989). Since [s] has a shorter front cavity than [ʃ], and also because [ʃ] but not [s] has a sublingual cavity which effectively lengthens the front cavity (Johnson, 2004), the spectral energy tends to be concentrated at a higher frequency for [s]. Since the length of the front cavity is negligible in [f, θ], their spectra are "diffuse," i.e., there are no major resonances and their overall energy is usually low. In addition, the sibilants [s, ʃ] have more energy at higher frequencies than [f, θ] not just because of the front cavity differences, but also because in the sibilants the airstream hits the teeth producing high-frequency turbulence (Stevens, 1971).

Voiced fricatives are produced with a simultaneous noise and voice sources. In the same spectrogram in Figure 3.9, there is both aperiodic energy in [z̥ð̥] of *is this* above 6,000 Hz and evidence of periodicity, as shown by the weak energy below roughly 500 Hz. The energy due to vocal fold vibration is often weak both in unstressed syllables such as these and more generally in voiced fricatives: this is because the high intraoral air pressure that is required for turbulence tends to cancel the subglottal pressure difference that is necessary to sustain vocal fold vibration. There is sometimes a noticeable continuity in the noise of fricatives with vowel formants (Soli, 1981). This is also apparent in Figure 3.9 as shown

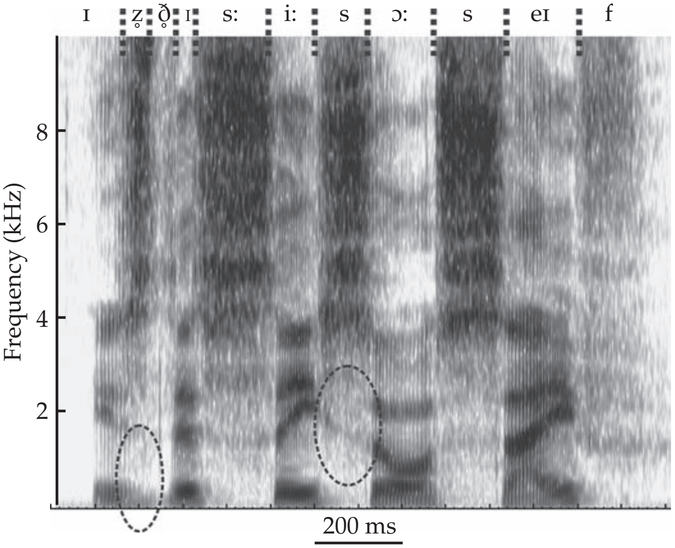


Figure 3.9 Spectrogram of the sentence "Is this seesaw safe?" produced by an adult male speaker of Australian English. There is evidence of weak periodicity in the devoiced [z̥ð̥] at the boundary of *is this* (ellipse, left) and of an F2 transition in the noise of the second [s] of *seesaw* (ellipse, right). (Adapted from Harrington & Cassidy, 1999, p. 58, figure 4.1, with kind permission of Springer Science and Business Media)

by the falling F2 transition across the noise in [iso] of *seesaw*. Fricatives especially [s, ʃ] are perceptually salient and they can mask a preceding nasal in vowel-nasal-fricative sequences: Ohala and Busà (1995) reason that this is one of the main factors that contributes to the common loss of nasals before fricatives diachronically (e.g., German *fünf*, but English *five*).

An oral stop is produced with a closure followed by a release which includes *transient*, *frication*, and sometimes *aspiration* stages (Repp & Lin, 1989; Fant, 1973). The transient corresponds to the moment of release and it shows up on a spectrogram as a vertical spike. The acoustics of the frication at stop release are very similar to the corresponding fricative produced at the same place of articulation. Aspiration, if it is present in the release of stops, is the result of a noise source at the glottis that may produce energy below 1 kHz (Figure 3.10). In the acoustic analysis of stops, the *burst* is usually taken to include a section of the oral stop extending for around 20 ms from the transient into the frication and possibly aspiration phases.

3.1 Place of articulation: Spectral shape

From considerations of the acoustic theory of speech production (Fant, 1960; Stevens, 1998), there are place-dependent differences in the spectral shape of stop bursts. Moreover, perception experiments have shown that the burst carries

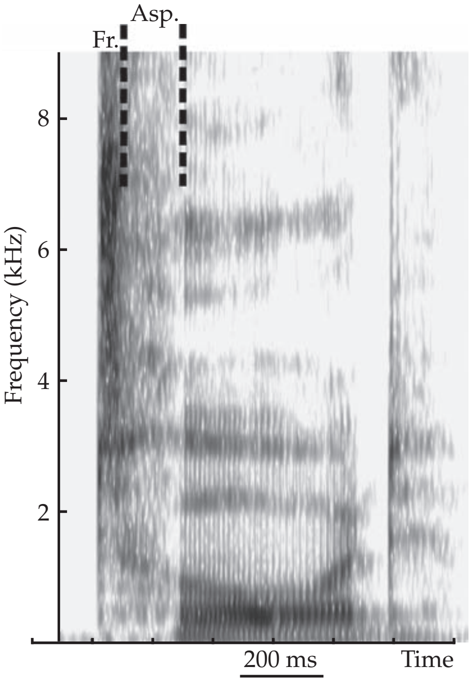


Figure 3.10 Spectrogram of an isolated production of the nonword [t^hɔ:d] (*tawed*) by a male speaker of Australian English showing the fricated and aspiration stages of the stop.

cues to stop place of articulation (Smits et al., 1996a; Fischer-Jørgensen, 1972). As studies by Blumstein and Stevens (1979, 1980) have shown, labial and alveolar spectra can often be distinguished from each other based on the slope of the spectrum which tends to fall for bilabials, but to rise with increasing frequency above roughly 3,000 Hz for alveolars. The separation of velars from other stops can be more problematic, partly because the vowel-dependent place of articulation variation in velars (fronted before front vowels and backed before back vowels) has such a marked effect on the spectrum. But a prediction from acoustic theory is that velars should have a mid-frequency spectral peak, i.e., a concentration of energy roughly in the 2,000–4,000 Hz range, whereas for the other two places of articulation, energy is more distributed over these frequencies compared with velars. This mid-frequency peak may well be the main factor that distinguishes velar from alveolar bursts before front vowels. Winitz et al. (1972) have shown that velar bursts are often misheard as alveolar before front vowels and this, as well a perceptual reinterpretation of the following aspiration, may be responsible for the diachronic change from /k/ to /tʃ/ in many languages (Chang et al., 2001).

A number of researchers have emphasied that burst cues to place of articulation may not depend on "static" information at a single spectral slice, but instead on the shape of the spectrum as it unfolds in time during the stop release and into the following vowel (e.g., Kewley-Port et al., 1983; Lahiri et al., 1984; Nossair & Zahorian, 1991). Since the burst spectrum of [b] falls with increasing frequency and since vowel spectra also fall with increasing frequency due to the falling glottal spectrum, then the change in spectral slope for [bV] from the burst to the vowel is in general small (Lahiri et al., 1984). As far as velar stops are concerned, these are sometimes distinguished from [b, d] by the presence of mid-frequency peaks that persist between the burst and the vowel onset (Kewley-Port et al., 1983).

Figure 3.11 shows spectra for Australian English [p^{h}a, t^{h}a, k^{h}a] between the burst and vowel onset as a function of normalized time. The displays are averages across five male Australian English speakers and are taken from syllable-initial stressed stops in read speech. The spectral displays were linearly time-normalized prior to averaging so that time point 0.5 is the temporal midpoint between the burst onset and the vowel's periodic onset. Once again, the falling, rising, and compact characteristics at the burst are visible for the labial, alveolar, and velar places of articulation respectively. The falling slope is maintained more or less into the vowel for [p^{h}a:], whereas for [t^{h}a:] the rising spectral slope that is evident at burst onset gives way to a falling slope towards the vowel onset producing a substantial change in energy in roughly the 3–5 kHz range. The same figure shows that the mid-frequency peak visible for [k^{h}a] as a concentration of energy at around 2.5 kHz at the burst onset persists through to the onset of the vowel (normalized time point 0.8).

The overall shape of the spectrum can be parameterized with *spectral moments* (e.g., Forrest et al., 1988) which are derived from statistical moments that are sometimes applied to the analysis of the shape of a histogram. Where x is a histogram class interval and f is the count of the number of tokens in a class interval, the ith statistical moment, m_i, ($i = 1, 2, 3, 4$) can be calculated as follows:

$$m_1 = \frac{\sum fx}{\sum f} \tag{6}$$

$$m_2 = \frac{\sum f(x - m_1)^2}{\sum f} \tag{7}$$

$$m_3 = \left(\frac{\sum f(x - m_1)^3}{\sum f} \right) m_2^{-1.5} \tag{8}$$

$$m_4 = \left[\left(\frac{\sum f(x - m_1)^4}{\sum f} \right) m_2^{-2} \right] - 3 \tag{9}$$

In *spectral* moments, a spectrum is treated as if it were a histogram so that x becomes the intervals of frequency and f is the dB value at a given frequency. If the

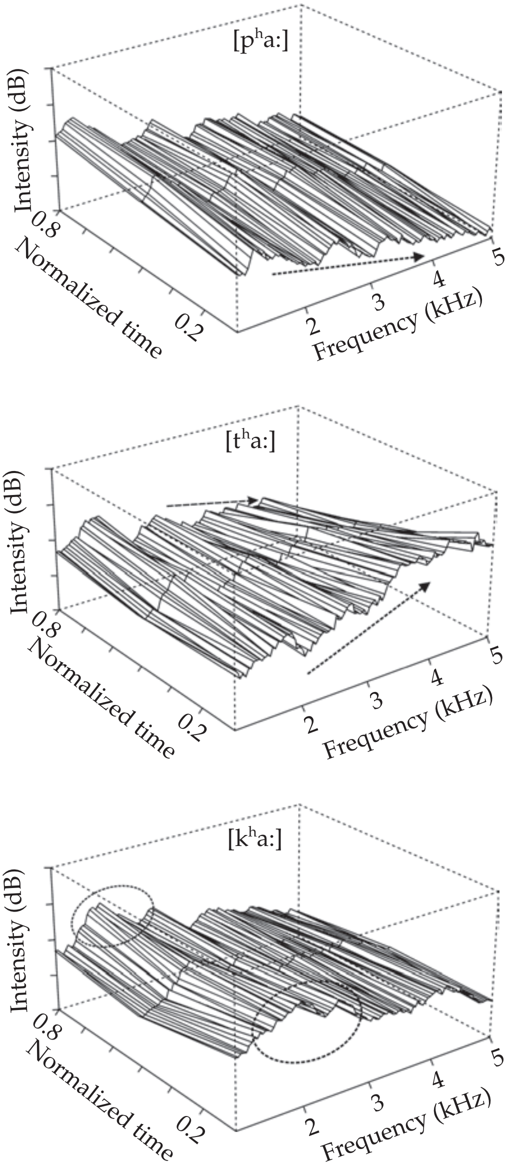


Figure 3.11 Spectra as a function of normalized time extending from the burst onset (time 0) to the acoustic onset of the vowel (time 0.8) for syllable-initial, stressed bilabial, alveolar, and velar stops preceding [a:] averaged across five male speakers of Australian English. The stops were taken from both isolated words and from read speech and there were roughly 100 tokens per category. The arrows mark the falling and rising slopes of the spectra at burst onset in [pʰa] and [tʰa] (and the arrow at time point 0.8 in [tʰa] marks the falling spectral slope at vowel onset). The ellipses show the mid-frequency peaks that persist in time in [kʰa]. (Adapted from Harrington & Cassidy, 1999, p. 86, figure 4.15, with kind permission of Springer Science and Business Media)

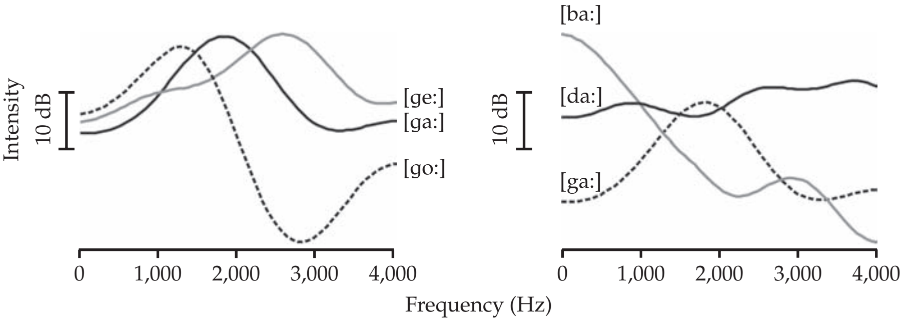


Figure 3.12 Cepstrally smoothed spectra calculated with a 16 ms window centered at the burst onset in word-initial [b, d, g] stops taken from isolated words produced by an adult male speaker of German. Left: spectra of [ge:, ga:, go:] bursts. Their m_1 (spectral center of gravity values) are 2,312 Hz, 1,863 Hz, and 1,429 Hz respectively. Right: spectra of the bursts of [ba:, da:, ga:]. Their $\sqrt{m_2}$ (spectral standard deviation) values are 1,007 Hz, 977 Hz, and 655 Hz respectively.

frequency axis is in Hz, then the units of m_1 and m_2 are Hz and Hz^2 respectively, while the third and fourth moments are dimensionless. It is usual in calculating moments to exclude the DC offset (frequency at 0 Hz) and to rescale the dB values so that the minimum dB value in the spectrum is set to 0 dB.

The first spectral moment m_1, gives the frequency at which the spectral energy is predominantly concentrated. Figure 3.12 shows cepstrally smoothed spectra calculated at the burst onset in stop-initial words produced in German. The figure in the left panel shows how m_1 decreases across [ge:, ga:, go:], commensurate with the progressive decrease in the frequency location of the energy peak in the spectrum that shifts due to the coarticulatory influence of the backness of the following vowel.

The second spectral moment, m_2, or its square root, the *spectral standard deviation*, is a measure of how distributed the energy is along the frequency axis. Thus in the right panel of Figure 3.12, m_2 is higher for [ba:, da:] than for [ga:] because, as discussed above, the spectra of the former are relatively more diffuse whereas [g] spectra tend to be more compact with energy concentrated around a particular frequency.

m_3 is a measure of asymmetry (see Figure 3.3 where this parameter was applied to F2 of [i:]). Given that spectra are always band-limited, the third spectral moment would seem to be necessarily correlated with m_1 (see for example the data in Jongman et al., 2000, table I): that is, m_3 is positive or negative if the energy is predominantly concentrated in low- and high-frequency ranges respectively. Finally m_4, kurtosis, is an expression of the extent to which the spectral energy is concentrated in a peak relative to the energy distribution in low and high frequencies. In general, m_4 is often correlated with m_2, although this need not be so (see, e.g., Wuensch, 2006 for some good examples).

Fricative place has been quantified with spectral moments in various studies (e.g., Forrest et al., 1988; Jongman et al., 2000; Tabain, 2001). Across these studies, two of the most important findings to emerge are:

- [s, z] have higher m_1 values than [ʃ, ʒ]. This is to be expected given the predictions from articulatory-to-acoustic mapping that the center frequency of the noise is higher for the former. When listeners label tokens from a synthetic /s–ʃ/ continuum, there is a greater probability that the same token is identified as /s/ before rounded compared with unrounded vowels (Mann & Repp, 1980). This comes about firstly because a lowered m_1 is a cue both for /ʃ/ and the result of anticipatory lip-rounding caused by rounded vowels; secondly, because listeners compensate for the effects of coarticulation, i.e., they factor out the proportion of m_1 lowering that is attributable to the effects of lip-rounding and so bias their responses towards /s/ when tokens are presented before rounded vowels.
- The second spectral moment tends to be higher for nonsibilants than sibilants, which is again predictable given their greater spectral diffuseness (e.g., Shadle & Mair, 1996).

Another way of parameterizing the shape of a spectrum is with the DCT (Nossair & Zahorian, 1991; Watson & Harrington, 1999). This transformation decomposes a signal into a set of half-cycle frequency cosine waves which, if summed, reconstruct the signal to which the DCT was applied. The amplitudes of these cosine waves are the *DCT coefficients* and when the DCT is applied to a spectrum, the DCT coefficients are equivalently *cepstral coefficients* (Nossair & Zahorian, 1991; Milner & Shao, 2006). For an N-point signal $x(n)$ extending in time from $n = 0$ to $N - 1$ points, the m^{th} DCT coefficient, C_m, ($m = 0, 1, \ldots N - 1$) can be calculated with:

$$C_m = \frac{2k_m}{N}\sum_{n=0}^{N-1} x(n)\cos\left(\frac{(2n+1)m\pi}{2N}\right) \tag{10}$$

$$k_m = \frac{1}{\sqrt{2}},\ m = 0;\ k_m = 1,\ m \neq 0$$

It can be shown that the first three DCT coefficients (C_0, C_1, C_2) are proportional to the *mean*, *linear slope*, and *curvature* of the signal respectively (Watson & Harington, 1999).

Figure 3.13 shows some spectral data of three German dorsal fricatives [ç, x, ʃ] taken from 100 read sentences of the Kiel corpus of read speech produced by a male speaker of Standard North German. The spectra were calculated at the fricatives' temporal midpoint with a 256-point discrete Fourier transform (DFT) at a sampling frequency of 16,000 Hz and the frequency axis was transformed to the Bark scale. DCT coefficients were calculated on these Bark spectra over the 500–7,500 Hz range. The fricatives were extracted irrespective of the segmental or prosodic contexts in which they occurred.

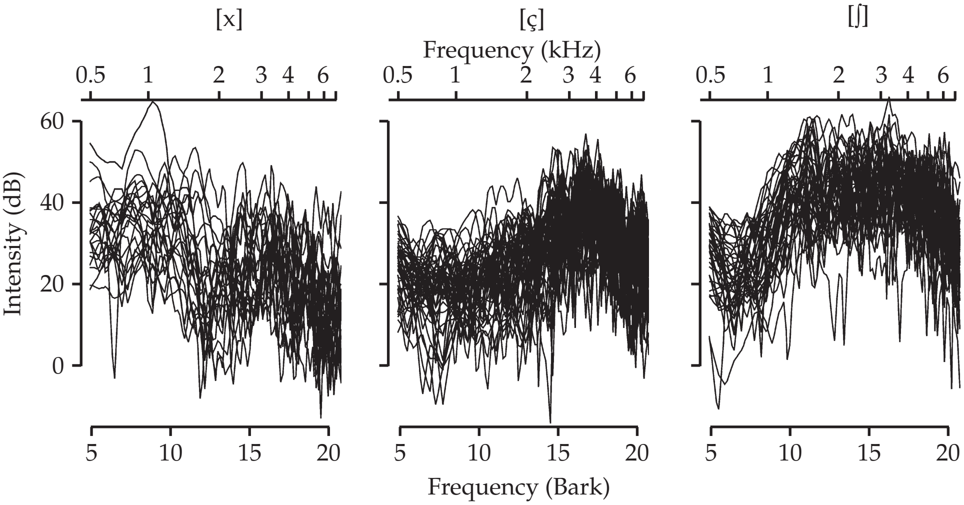


Figure 3.13 Spectra in the 0–8 kHz range calculated with a 16 ms DFT at the temporal midpoint of the German fricatives [x] (left, n = 25), [ç] (center, n = 50), and [ʃ] (right, n = 39) and plotted with the frequency axis proportional to the Bark scale. The data are from read sentences produced by one male speaker of Standard German.

As is well known, [ç] and [x] are allophones of one phoneme in German that are predictable from the frontness of the preceding vowel, but they also have very different spectral characteristics. As discussed in Johnson (2004), the energy in back fricatives like [x] tracks F2 of the following vowel, whereas in palatal fricatives like [ç], the energy is concentrated at a higher frequency and is continuous with the flanking vowel's F3. As Figure 3.13 shows, the palatal [ç] patterns more closely with [ʃ] because [x] has a predominantly falling spectrum whereas the spectra of [ç] and [ʃ], which show a concentration of energy in the 2–5 kHz range, are rising. The distinction between [ʃ] and [ç] could be based on curvature: [ʃ], there is a greater concentration of energy around 2–3 kHz so that the [ʃ] spectra have a greater resemblance to an inverted U-shape than those of [ç].

Figure 3.14 shows the distribution of the same spectra on the DCT coefficients C_1 and C_2. Compatibly with these predictions from Figure 3.13, [x] is separated from the other fricatives primarily on C_1 (spectral slope) whereas the [ʃ]–[ç] distinction depends on C_2 (spectral curvature). Thus together C_1 and C_2 provide quite an effective separation between these three dorsal fricative classes, at least for this single speaker.

3.2 Place of articulation in obstruents: Other cues

Beyond these considerations of gross spectral shape discussed in the preceding section and F2 locus cues in formant transitions discussed in 2.6, place of articulation within obstruents is cued by various other acoustic attributes, in particular:

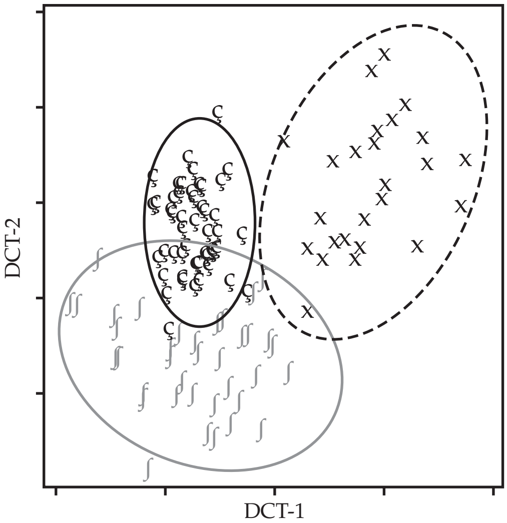


Figure 3.14 95 percent confidence ellipses for three fricatives in the plane of DCT-1 and DCT-2 obtained by applying a DCT to the Bark-scaled spectra in Figure 3.13.

- The bursts of labials tend to be weak in energy (Fischer-Jørgensen, 1954; Fant, 1973) since they lack a front cavity and perceptual studies have shown that this energy difference in the burst can be used by listeners for distinguishing labials from alveolars (e.g., Ohde & Stevens, 1983). The overall intensity of the burst relative to that of the vowel has also been used by Jongman et al. (1985) for place of articulation distinctions in voiceless coronal stops produced by three adult male talkers of Malayalam.
- The duration of the stop release up to the periodic onset of the vowel in CV syllables, i.e., voice onset time (VOT), can also provide information about the stop's place of articulation: in carefully controlled citation-form stimuli, within either voicing category, velar stops have longer VOTs than alveolar stops, whose VOTs are longer than those of bilabial stops (e.g., Kewley-Port, 1982).
- The *amplitude of the frication noise* has been shown to distinguish perceptually the sibilant fricatives [s, ʃ] from nonsibilants like [f, θ] (Heinz & Stevens, 1961). Ali et al. (2001) found an asymmetry in perception such that decreasing the amplitude of sibilants leads them to be perceived as nonsibilants (whereas increasing the amplitude of nonsibilants does not cause them to be perceived as sibilants).
- Studies by Harris (1958) and Heinz and Stevens (1961) showed that, whereas the noise carried more information for place distinctions than formant transitions, F2 and F3 may be important in distinguishing [f] from [θ] given that labiodentals and dentals have very similar noise spectra (see Tabain, 1998 for an analysis of spectral information above 10 kHz for the labiodental/dental

fricative distinction). More recently, Nittrouer (2002) found that in comparison with children, adults tended to be more reliant on noise cues than formant transition cues in distinguishing [f] from [θ]. F2 transitions in noise have been shown to be relevant for distinguishing [s] from [ʃ] acoustically and perceptually (Soli, 1981).

3.3 *Obstruent voicing*

VOT is the duration from the stop release to the acoustic periodic onset of the vowel and it is perhaps the most salient acoustic cue for distinguishing domain-initial voiced from voiceless stops in English and in many languages (Lisker & Abramson, 1964, 1967). If voicing begins during the closure (as in the example in Figure 3.5), then VOT is negative. The duration of the noise in fricatives is analogous to VOT in stops and it has been shown to be an important cue for the voicing distinction within syllable-initial fricatives (Cole & Cooper, 1975) although noise duration is not always consistently less in voiced than in voiceless fricatives (Jongman, 1989).

VOT differences can be related to differences in the onset frequency and transition of F1. When the vocal tract is completely occluded, F1 is at its theoretically lowest value. Then, with the release of the stop, F1 rises (Stevens & House, 1956; Fant, 1960). The F1-transition rises in both voiced and voiceless CV stops, but since periodiocity starts much earlier in voiced stops (in languages that use VOT for the voicing distinction), much more of the transition is periodic and the onset of voiced F1 is often considerably lower (Fischer-Jørgensen, 1954).

In a set of synthesis and perception experiments, Liberman et al. (1958) showed that delaying the onset of F1 relative to the burst and to F2 and F3 was a primary cue for the voiced/voiceless distinction (see also Darwin & Seton, 1983). Subsequent experiments in speech perception have shown that a rising periodic F1-transition (e.g., Stevens & Klatt, 1974) and a lower F1-onset frequency (e.g, Lisker, 1975) cue voiced stops and that there may be a trading relationship between VOT and F1-onset frequency (Summerfield & Haggard, 1977). Thus as is evident in comparing [k^h] with [g] in Figure 3.5, both F2 and F3 converge back towards a common onset frequency near the burst, but the first part of these transitions is aperiodic in the voiceless stop. Also, although F1 rises in both cases, the rising part of the transition is aperiodic in [k^h] resulting in a higher F1-onset frequency at the beginning of the voiced vowel.

In many languages, voiceless stops are produced with greater articulatory force and as a result the burst amplitude (Lisker & Abramson, 1964) and the rate at which the energy increases is sometimes greater in voiceless stops (Slis & Cohen, 1969). In various perception experiments, Repp (1979) showed that increasing the amplitude of aspiration relative to that of the following vowel led to greater voiceless stop percepts. The comparison of burst amplitude across stop voicing categories is one example in which *first-differencing* the signal can be important. When a signal is differenced, i.e., samples at time points n and $n - 1$ are subtracted from each other, there is just under a 6 dB rise per octave or doubling of frequency

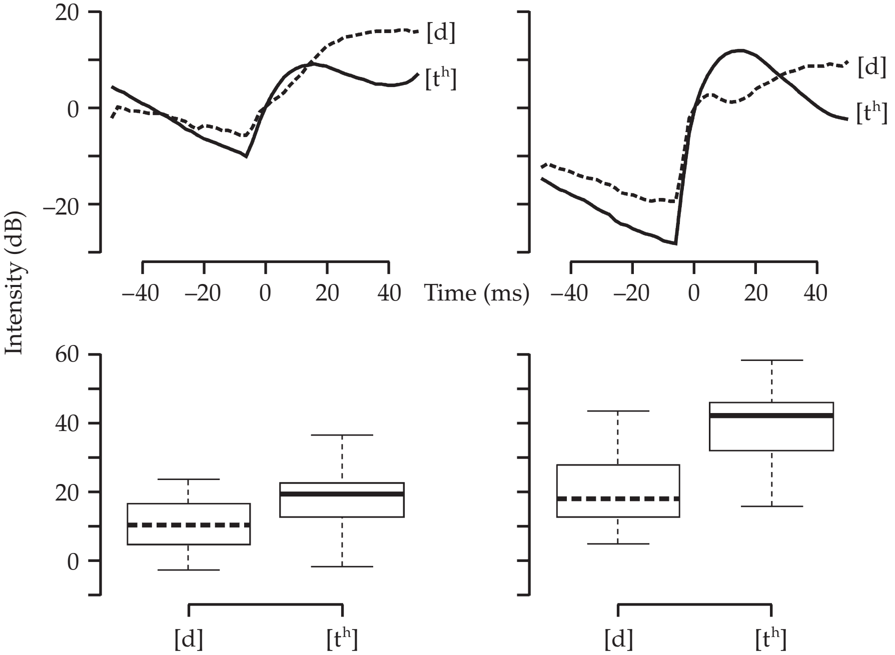


Figure 3.15 Row 1: averaged dB-RMS trajectories of [d] (*n* = 22) and [tʰ] (n = 69) calculated with a 10 ms rectangular window on sampled speech data without (left) and with (right) first-differencing. 0 ms marks the burst onset. The averaging was done after rescaling the amplitude of each token relative to 0 dB at the burst onset. The stops are from two male speakers of Australian English and were extracted from prevocalic stressed syllable-initial position from 100 read sentences per speaker irrespective of vowel context. Row 2: boxplots showing the corresponding distribution of [d, tʰ] on the parameter *b–a*, where *b* and *a* are respectively the dB values 10 ms after, and 10 ms before the burst onset. (The height of the rectangle marks the interquartile range).

in the spectrum, so that the energy at high frequencies is boosted. Given that at stop release there may well be greater energy in the upper part of the spectrum in voiceless stops, the effect of first-differencing is likely to magnify any energy differences across voiced and voiceless stops. In Figure 3.15, the root-mean-square (RMS) energy has been calculated in voiced and voiceless stops: in the left panels, there was no differencing of the sampled speech data, whereas in the right panels the speech waveform was first differenced before the RMS energy calculation was applied. As the boxplots show, there is only a negligible difference in burst amplitude across the voicing categories on the left; but with the application of first differencing, the rise in amplitude of the stop burst is much steeper and the difference in energy 10 ms before and after the release is noticeably greater in the voiceless stop.

The fundamental frequency is higher after voiceless than voiced obstruents (House & Fairbanks, 1953; Lehiste and Peterson, 1961; Hombert et al., 1979) and this has been shown to be a relevant cue for the voicing distinction both in stops (e.g., Whalen et al., 1993) and in fricatives (Massaro & Cohen, 1976). Löfqvist et al. (1989) have shown that these voicing-dependent differences in f_0 are the result of increased longitudinal tension in the vocal folds (but see Hombert et al., 1979, for an aerodynamic interpretation).

Several studies have concerned themselves with the acoustic and perceptual cues that underlie final (e.g., *duck/dug*) and intervocalic (*rapid/rabid*) voicing distinction. Denes (1955) showed that the distinction between /ju:s/ (*use*, noun) and /ju:z/ (*use*, verb) was based primarily on the vowel duration acoustically and perceptually. The acoustic cues that signal the final voicing in pairs have also been shown to include the F1-offset frequency and rate of F1-offset transition (e.g., Wardrip-Fruin, 1982).

Lisker (1978, 1986) showed that voicing during the closure is one of the main cues distinguishing *rapid* and *rabid* in English. Kohler (1979) demonstrated that the cues for the same phonological contrast have different perceptual rankings in different languages. He showed that, whereas voicing during the closure is a more salient cue than vowel : consonant duration ratios in French, it is the other way round in German. Another important variable in the post-vocalic voicing distinction in German can be the drop of the fundamental frequency contour in the vowel which is greater preceding voiced stops (Kohler, 1985).

4 Nasal Consonants and Nasalized Vowels

Nasal consonants are detectable on spectrograms by the presence of a *nasal murmur* corresponding to the phase of nasal consonant production in which the oral tract is closed and air passes through the nasal cavity. The overall amplitude of the nasal murmur is low and the energy is concentrated predominantly in a low frequency range. The beginning and end of the nasal murmur can often be quite easily detected by abrupt spectral discontinuities that are associated with the combined lowering/raising of the velum and closing/opening of the oral tract at the onset/offset of the nasal consonant. Such discontinuities are considered by Stevens (1985, 2002) to carry some of the main cues to the place of articulation in nasal consonants – this point is discussed again more fully below. In English and in many languages, this abruptness, i.e., syntagmatic distinction between the vowel and nasal, is a good deal more marked in syllable-initial nasal-vowel than in syllable-final vowel-nasal transitions (e.g., Repp & Svastikula, 1988; Redford and Diehl, 1999). These syllable-position differences can, in turn, be related to studies of sound change showing a greater propensity for the vowel and nasal to merge when the nasal is syllable-final (e.g., Hajek, 1997).

The spectrum of the nasal murmur is characterized by a set of nasal formants (N1, N2, . . .) that are the result of excitation of the combined nasal-pharyngeal tube. N1 has been calculated from vocal tract models to occur in the 300–400 Hz region and higher nasal formants occur for an adult male tract at intervals of about 800 Hz (Fant, 1960; Flanagan, 1972). Various studies also concur that that nasal formant bandwidths are broad and that N1 is high in amplitude (e.g., Fujimura, 1962).

In addition, the oral cavity acts as a side-branching resonator to the main nasal-pharyngeal tube and this results in *oral anti-formants* that absorb energy from the main nasal-pharyngeal tube. The presence of anti-formants is one of the reasons why the overall amplitude of nasal consonants is low. (Another is that, because the mouth cavity is sealed, the amount of acoustic energy leaving the vocal tract is much less than for vowels.) The spectral effect of introducing an anti-formant is both to produce a spectral dip at the anti-formant frequency and to alter the spectral balance or spectral tilt (Atal, 1985) and it is this change of spectral balance that may be as important a cue for nasalization as the frequency at which the anti-formant occurs.

The center frequency of the first oral anti-formant (FZ1) in nasal consonants is predicted to vary inversely with the length of the mouth cavity and is lowest for [m], higher for [n], highest for [ŋ] (Fant, 1960; Fujimura, 1962). A uvular nasal [N] has no anti-formants since, as the tongue constriction is so far back in the mouth, there is no oral side-branching resonator. Since FZ1 for [n] tends to coincide with N3 in roughly the 1,800 Hz range and since FZ1 for [m] occurs at a lower frequency, [n] nasal murmurs are expected to have less energy in the 1,500–2,000 Hz range than those of [m]. These FZ1-dependent spectral differences between [m] and [n] were incorporated into a metric for distinguishing between these two places of articulation in Kurowski and Blumstein (1987).

The F2 locus theory should also be applicable to nasal consonants and Liberman et al. (1954) were the first to show place of articulation in nasal consonants could be cued by formant transitions that pointed to different locus frequencies; more recently, locus equations have been applied to place of articulation distinctions in nasal consonants (Sussman, 1994). There have been several studies in which nasal murmurs and transitions have been cross-spliced, in which for example an [m]-murmur is combined with transitions appropriate for [n] (see Recasens, 1983 for a review). One of the first of these was by Malécot (1956), who showed that listeners' judgments were predominantly guided by the transitions and not the murmur. On the other hand, Kurowski & Blumstein (1984) found that nasal place was more accurately identified from a section of the waveform spanning the murmur–vowel boundary than from waveforms containing either only the murmur or only the vowel transition. This finding is consistent with studies in speech perception (Repp, 1986, 1987; Repp & Svastikula, 1988; Ohde et al., 2006) and with various acoustic studies (Kurowski & Blumstein, 1987; Seitz et al., 1990; Harrington, 1994) showing that the salient cues to nasal place of articulation are at the murmur–vowel boundary.

The spectrograms in Figure 3.16 of five phonemically contrastive nasal consonants in the Central Australian language Warlpiri recorded from one female speaker by Andrew Butcher in 2005 show evidence of differences in both the murmur and the transitions. In particular:

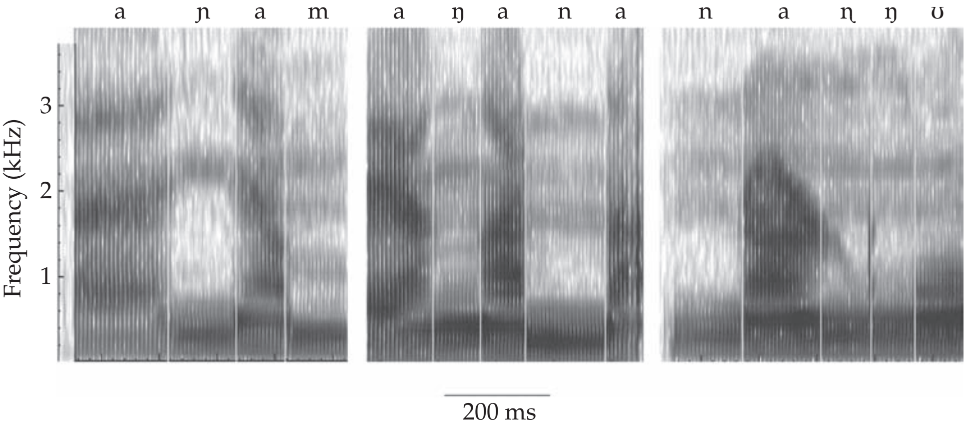


Figure 3.16. Spectrograms of /a#ɲampu/ (left), /a#ŋana/ (center), /naɳŋʊ/ (right) produced by a female speaker of the Central Australian language Warlpiri (# is a word boundary).

- Compatibly with some of the studies reviewed above, [n] lacks very much energy in the 1–1.5 kHz range because of the presence of an anti-formant in this frequency region.
- The lack of energy is also evident for [ɲ] but it occurs over a wider range from 1–2 kHz possibly because, since the mouth cavity is shorter for the palatal nasal, NZ1 for [ɲ] is at a higher frequency than for [n]. Also, [ɲ] has an intense formant at around 2,200 Hz that is reminiscent of F2 of the palatal vowels [i] or [ɪ].
- [ŋ] has quite a flat spectrum up to 3,000 Hz, i.e., no marked dips or peaks. The absence of any marked energy dips is expected given that the lowest anti-formant is predicted to occur above 3,000 Hz (Fant, 1960).
- It is evident that some of the distinguishing characteristics between these nasals are in the formant transitions. For example, F2 rises in the vowel of [aɲ], F2 falls in the vowel of [aŋ], and there is a steeply falling F2 (or possibly F3) in the vowel of [aɳ].

In the acoustic analysis of nasal consonants, some researchers have tried to parameterize place differences using formant variables (e.g., Chen, 1995, 1997). Although such an approach has the advantage of linking acoustic structure to vocal tract activity, it is, in practice, extremely difficult to identify with any certainty both whether a particular resonance is a nasal or an oral formant (and if so which formant number) and also whether a dip that can be seen in a spectrogram or spectral slice really is due to an anti-formant or else to a trough between formant peaks. Then there is the added complexity that vowels adjacent to nasal consonants are often nasalized which again makes the identification of vowel oral formant frequencies problematic. For this reason, a whole-spectrum approach is often favored in the acoustic (e.g., Qi & Fox, 1992) and perceptual (e.g., Kataoka et al., 2001) analysis of nasal place of articulation, which does nevertheless often allow relationships to be established to formant frequencies. For example, Kurowski and Blumstein (1987) reasoned that the energy change from the murmur into the following vowel should be greater for [n] than [m] in the 11–14 Bark (approx. 1,450–2,300 Hz) range because this is the frequency region in which [n], but not [m] has an anti-formant (see also Qi & Fox, 1992 for compatible evidence).

Energy change at the murmur–vowel boundary can be parameterized with a *difference spectrum* (i.e., by subtracting a spectrum close to the vowel onset from a spectrum close to the murmur offset) and both Kurowski and Blumstein (1987) and Seitz et al. (1990) showed high classification scores for the [m–n] distinction using metrics based on difference spectra. The analysis in Harrington (1994) also included information from the murmur and from the vowel, but the classification was based on *combined*, rather than differenced, spectral information across the murmur–nasal boundary. The idea that the combination of separately processed murmur and vowel spectra provides salient cues to place of articulation in nasals has also been shown in perception experiments of adult and child speech (Ohde et al., 2006).

Nasal vowels can occur phonetically due to the effects of context and in many languages they contrast phonemically with oral vowels. There is a correlation between phonetic vowel height and velum height: when high vowels are nasalized, the velum is not lowered to the same extent as when low vowels are nasalized. The reasons for this may be either physiologically determined by a muscular connection between the velum and the tongue (e.g., Moll, 1962; but see Lubker, 1968) or based on auditory factors that require a certain ratio of oral to nasal impedance for nasalization to be perceptible (House & Stevens, 1956; and see the discussion in Abramson et al., 1981).

When a vowel is nasalized, the mouth aperture is bigger than the nose aperture and as a result, the nasal cavity becomes a side-branching resonator to the oral-pharyngeal tube, which introduces an additional set of *nasal formants* and *nasal anti-formants* into the spectrum (House & Stevens, 1956; Fant, 1960; Fujimura, 1962). Some of the main acoustic consequences that result from coupling of the oral and nasal tubes in the production of nasalized vowels are as follows:

- There are changes to the oral formants. In particular, F1 moves up in frequency, is lowered in intensity, and has a broader bandwidth (House & Stevens, 1956).
- Compared with oral vowels, nasal vowels often have a greater density of formants in the 0–3,000 Hz range due to the presence of both oral and nasal formants. In the spectrogram on the left in Figure 3.16, the word-medial /a/ in /a#ɲampu/ is evidently more nasalized than the preboundary /a#/ in the same word: there are at least three resonance peaks for the former compared with two for the latter in the 500–2,500 Hz range. Similarly, the nasalized realization of /iː/ in *meaning* in Figure 3.17 has an additional nasal resonance at around 1,000 Hz compared with the oral production of /iː/ in *deeper* produced in the same prosodic phrase by the same speaker. An increase in

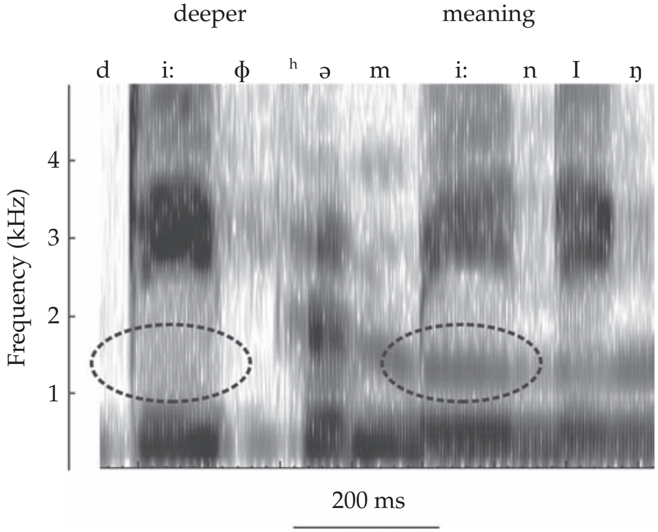


Figure 3.17 Spectrogram of *deeper meaning* from the 2004 Queen Elizabeth II Christmas broadcast data (Harrington, 2006). The ellipses extend over the interval of [i:] in the two words showing the absence and presence respectively of a nasal formant at just under 1.5 kHz. (/p/ in *deeper* has been transcribed with a bilabial fricative since, as the aperiodic energy over this interval shows, the closure of the stop is evidently not complete).

the amplitude between F1 and F2 is common when high vowels are nasalized and recently Kataoka et al. (2001) have shown that this amplitude increase is correlated with perceived hypernasality in children.

- In mid vowels, i.e., vowels that have F1 roughly in the 300–800 Hz region, F1 is replaced with a triplet of an oral formant, nasal formant, and nasal anti-formant, i.e. an F1–N1–NZ1 combination (e.g., Hawkins & Stevens, 1985). Often F1 and N1 are not distinct, so that the overall effect in comparing oral and nasal mid vowels is that the bandwidth of the first formant (merged F1 and N1) is considerably broader than F1 of the corresponding oral vowel (Hattori et al., 1958). The peak amplitude of the first peak in nasalized mid vowels is also likely to be lower, both because of the broader bandwidth, and because of the presence of NZ1.

There is a loss of perceptual contrast especially along the height dimension when vowels are nasalized i.e., high vowels tend to be perceived to be lower and low vowels are perceived to be phonetically higher (Wright, 1975, 1986). The acoustic basis for this loss of distinction is likely to be that in high vowels, the mouth-cavity dependent F1 is raised, while in low vowels, the entire spectral center of gravity in the region of F1 is lowered due to the presence of N1 that is lower in frequency than F1 (Krakow et al., 1988). The perceptual lowering effect of nasalization was demonstrated by Beddor et al. (1986). They synthesized two continua from *bad* to *bed* by lowering F1. In one continuum, the vowels were oral, /bæd–bɛd/, and in the other they were nasal, /bæ̃d–bɛ̃d/. They found more tokens from the nasal continuum were labelled as *bad* than from the oral continuum (see Ohala, 1993 for a number of sound changes that are consistent with this effect). In a related experiment, Krakow et al. (1988) additionally synthesized a continuum from *bend* to *band* using nasal vowels (thus /bæ̃nd–bɛ̃nd/). They found that the responses from /bæ̃nd–bɛ̃nd/ patterned with the *oral* continuum from *bad–bed* /bæd–bɛd/ rather than with nasal /bæ̃d–bɛ̃d/. They reasoned that this is because listeners attribute the acoustic nasalization in /bæ̃nd–bɛ̃nd/ not to the vowel (as they necessarily must do in /bæ̃d–bɛ̃d/ since vowel nasalization has no coarticulatory raison d'être in this case) but to the contextual effects of the following /n/ consonant.

5 Concluding Comment

The three areas that have made substantial contributions to acoustic phonetics that were outlined at the beginning of this chapter are all certain to continue to be important for progress in the field in the future. As a result of the advances in speech physiology, in particular using techniques such as MRI and ultrasound, it is now possible to draw upon a much wider range of vocal tract cross-sectional data allowing more realistic articulatory-to-acoustic models to be developed. Making use of the extensive speech corpora that have become available in the last 15–20 years due largely to the needs of speech technology will be important for expanding our understanding of variability due to different speaking styles, age groups, language varieties, and many other factors. With the existence of larger amounts of training data that are now available from speech corpora, it should be possible in the future to incorporate into acoustic phonetic studies more sophisticated probabilistic and, above all, time-dependent models of the speech signal. Just this kind of information is becoming increasingly important in both phonology and linguistics (Bod et al., 2003). Analyzing large speech corpora will also be essential to ensure a greater convergence in the future between basic speech research and speech technology, so that more of the knowledge that has been described in this chapter can be incorporated more explicitly into the design of human–machine communication systems.

ACKNOWLEDGMENTS

My thanks to David Broad and Frantz Clermont for a helpful discussion of Broad and Clermont (1987), to Janine Lilienthal and Nils Ülzmann for proofreading an earlier version of the manuscript, and to John Laver and Sallyanne Palethorpe for many helpful comments.

REFERENCES

Abramson, A., Nye, P., Henderson, J., & Marshall, C. (1981) Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America*, 70, 329–39.

Ali, A., Van der Spiegel, J., & Mueller, P. (2001) Acoustic-phonetic features for the automatic classification of fricatives. *Journal of the Acoustical Society of America*, 109, 2217–35.

Assmann, P. (1991) The perception of back vowels: Centre of gravity hypothesis, *Quarterly Journal of Experimental Psychology*, 43, 423–8.

Atal, B. S. (1985) Linear predictive coding of speech. In F. Fallside & W. A. Woods (eds.), *Computer Speech Processing* (pp. 81–124). Englewood Cliffs, NJ: Prentice-Hall.

Atal, B. S. & Hanauer, S. (1971) Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50, 637–55.

Aylett, M. & Turk, A. (2006) Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, 119, 3048–58.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992) Prosodic structure and tempo in a sonority model of articulatory dynamic. In G. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, and Prosody* (pp. 68–86). Cambridge: Cambridge University Press.

Beddor, P. S., Krakow, R. A., & Goldstein, L. M. (1986) Perceptual constraints and phonological change: A study of nasal vowel height, *Phonology Yearbook* 3, 197–217.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003) Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, 113, 1001–24.

Bird, S. & Harrington, J. (2001) Speech annotation and corpus tools. *Speech Communication*, 33, 1–4.

Bladon, R. A. W. (1982) Arguments against formants in the auditory representation of speech. In R. Carlson and B. Granström (eds.), *The Representation of Speech in the Peripheral Auditory System* (pp. 95–102). Amsterdam: Elsevier Biomedical.

Bladon, R. A. W. & Al-Bamerni, A. (1976) Coarticulation resistance in English /l/, *Journal of Phonetics*, 4, 137–50.

Bladon, R. A. W. & Lindblom, B. (1981) Modeling the judgment of vowel quality differences, *Journal of the Acoustical Society of America* 69, 1414–22.

Blumstein, S. E. & Stevens, K. N. (1979) Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America* 66, 1001–17.

Blumstein, S. E. & Stevens, K. N. (1980) Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* 67, 648–62.

Bod, R., Hay, J., & Jannedy, S. (2003) *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Brancazio, L. & Fowler, C. (1998) The relevance of locus equations for production and perception of stop consonants. *Perception and Psychophysics*, 60, 24–50.

Broad, D. & Clermont, F. (1987) A methodology for modeling vowel formant contours in CVC context.

Journal of the Acoustical Society of America, 81, 155–65.

Broad, D. & Fertig, R. H. (1970) Formant-frequency trajectories in selected CVC utterances. *Journal of the Acoustical Society of America* 47, 1572–82.

Broad, D. J. & Wakita, H. (1977) Piecewise-planar representation of vowel formant frequencies. *Journal of the Acoustical Society of America*, 62, 1467–73.

Bybee, J. (2000) Lexicalization of sound change and alternating environments. In M. Broe & J. Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon* (pp. 250–68). Cambridge: Cambridge University Press.

Carlson, R., Fant, G., & Granström, B. (1975) Two-formant models, pitch and vowel perception. In G. Fant & M. A. A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 55–82). New York: Academic Press.

Carré, R. & Mrayati, M. (1992) Distinctive regions in acoustic tubes: Speech production modeling. *Journal d'Acoustique*, 5, 141–59.

Cassidy, S. & Harrington, J. (1995) The place of articulation distinction in voiced oral stops: Evidence from burst spectra and formant transitions. *Phonetica*, 52, 263–84.

Chang, S., Plauché, M. C., & Ohala, J. J. (2001) Markedness and consonant confusion asymmetries. In E. Hume & K. Johnson (eds.), *The Role of Speech Perception in Phonology* (pp. 79–101). San Diego CA: Academic Press.

Chen, M. Y. (1995) Acoustic parameters of nasalized vowels in hearing impaired and normal-hearing speakers. *Journal of the Acoustical Society of America*, 98, 2443–53.

Chen, M. Y. (1997) Acoustic correlates of English and French nasalized vowels. *Journal of the Acoustical Society of America*, 102, 2360–70.

Chennoukh, S., Carré, R., & Lindblom, B. (1997) Locus equations in the light of articulatory modeling. *Journal of the Acoustical Society of America*, 102, 2380–9.

Chistovich, L. A. (1985) Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America*, 77, 789–805.

Chistovich, L. A. & Lublinskaya, V. V. (1979) The "center of gravity" effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, 1, 185–95.

Cole, R. A. & Cooper, W. E. (1975) Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America*, 58, 1280–7.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences of the United States of America*, 37, 318–25.

Cox., F. (1998) The Bernard data revisited. *Australian Journal of Linguistics*, 18, 29–55.

Darwin, C. & Seton, J. (1983) Perceptual cues to the onset of voiced excitations in aspirated initial stops. *Journal of the Acoustical Society of America*, 73, 1126–35.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955) Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27, 769–73.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, F. J. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesised from spectrographic patterns. *Word*, 8, 195–210.

Denes, P. (1955) Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 27, 761–4.

Dunn, H. (1950) The calculation of vowel resonances in an electrical vocal tract. *Journal of the Acoustical Society of America*, 22, 740–53.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991) The articulatory kinematics of final lengthening, *Journal of the Acoustical Society of America*, 89, 369–82.

Engstrand, O. & Lindblom, B. (1997) The locus line: Does aspiration affect its steepness? *Reports from the Department of Linguistics: Umea University (PHONUM)*, 4, 101–4.

Espy-Wilson, C. Y. (1992) Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. *Journal of the Acoustical Society of America*, 92, 736–57.

Espy-Wilson, C. Y. (1994) A feature-based semivowel recognition system. *Journal of the Acoustical Society of America*, 96, 65–72.

Espy-Wilson, C. Y., Boyce, S., Jackson, M., Narayanan, S., & Alwan, A. (2000) Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America*, 108, 343–56.

Essner, C. (1947) Recherche sur la structure des voyelles orales. *Archives Néerlandaises de Phonétique Expérimentale*, 20, 40–77.

Fant, G. (1960) *Acoustic Theory of Speech Production*. The Hague Mouton.

Fant, G. (1973) *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Fischer-Jørgensen, E. (1954) Acoustic analysis of stop consonants, *Miscellenea Phonetica*, 2, 42–59.

Fischer-Jørgensen, E. (1972) Tape-cutting experiments with Danish stop consonants in initial position, *Annual Report, Institute of Phonetics, University of Copenhagen*, 6, 104–68.

Flanagan, J. L. (1972) *Speech Synthesis, Analysis, and Perception*. New York: Springer-Verlag.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988) Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, 84, 115–24.

Fourakis, M. (1991) Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, 90, 1816–27.

Fowler, C. A. (1994) Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, 55, 597–611.

Fowler, C. A. & Housum, J. (1987) Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction, *Journal of Memory and Language*, 26, 489–504.

Fry, D. B. (1965) The dependence of stress judgements on vowel formant structure. In E. Zwirner & W. Bethge (eds.), *Proceedings of the Fifth International Conference of Phonetic Sciences* (pp. 306-3–1). Basel: Karger.

Fujimura, O. (1962) Analysis of nasal consonants. *Journal of the Acoustical Society of America*, 34, 1865–75.

Gay, T. (1968) Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America*, 44, 1570–73.

Gay, T. (1970) A perceptual study of American English diphthongs. *Language and Speech*, 13, 65–88.

Gottfried, M., Miller, J. D., & Meyer, D. J. (1993) Three approaches to the classification of American English diphthongs. *Journal of Phonetics*, 21, 205–29.

Hajek, J. (1997) *Universals of Sound Change in Nasalization*. Oxford: Blackwell.

Harrington, J. (1994) The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants. *Journal of the Acoustical Society of America*, 96, 19–32.

Harrington, J. (2006) An acoustic analysis of "happy-tensing" in the Queen's Christmas Broadcasts. *Journal of Phonetics*, 34, 439–57.

Harrington, J. & Cassidy, S. (1994) Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language and Speech*, 37, 357–73.

Harrington, J. & Cassidy, S. (1999) *Techniques in Acoustic Phonetics*. Dordrecht: Kluwer.

Harrington, J., Fletcher, J., & Beckman, M. E. (2000) Manner and place conflicts in the articulation of accent in Australian English. In M. Broe (ed.), *Papers in Laboratory Phonology V* (pp. 40–55). Cambridge: Cambridge University Press.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1, 1–7.

Hashi, M., Honda, K., & Westbury, J. (2003) Time-varying acoustic and articulatory characteristics of American English [ɹ]: A cross-speaker study. *Journal of Phonetics*, 31, 3–22.

Hattori, S., Yamamoto, K., & Fujimura, O. (1958) Nasalization of vowels in relation to nasals. *Journal of the Acoustical Society of America*, 30, 267–74.

Hawkins, S. & Stevens, K. N. (1985) Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels. *Journal of the Acoustical Society of America*, 77, 1560–75.

Hay, J., Sato, M., Coren, A., Moran, C., & Diehl, R. (2006) Enhanced contrast for vowels in utterance focus: A cross-language study. *Journal of the Acoustical Society of America*, 119, 3022–33.

Heinz, J. M. & Stevens, K. N. (1961) On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 33, 589–93.

Hillenbrand, J., Clark, M., & Nearey, T. (2001) Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109, 748–63.

Hillenbrand, J., Houde, R., & Gayvert, R. (2006) Speech perception based on spectral peaks versus spectral shape. *Journal of the Acoustical Society of America*, 119, 4041–54.

Hillenbrand, J. & Nearey, T. (1999) Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, 105, 3509–23.

Holbrook, A. & Fairbanks, G. (1962) Diphthong formants and their movements. *Journal of Speech and Hearing Research*, 5, 38–58.

Hombert, J-M., Ohala, J., & Ewan, W. (1979) Phonetic explanations for the development of tones. *Language*, 55, 37–58.

House, A. S. & Fairbanks, G. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 25, 105–13.

House, A. S. & Stevens, K. N. (1956) Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders*, 21, 218–32.

Huang, C. B. (1986) The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels. *IEEE International Conference on Acoustics Speech and Signal Processing*, 893–6.

Huang, C. B. (1992) Modeling human vowel identificatin using aspects of format trajectory and context. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 43–61). Amsterdam: IOS Press.

Hunnicutt, S. (1985) Intelligibility versus redundancy: Conditions of dependency. *Language and Speech*, 28, 47–56.

Hunnicutt, S. (1987) Acoustic correlates of redundancy and intelligibility. *Speech Transmission Laboratory, Quarterly Status Progress Report*, 2–3, 7–14.

Ito, M., Tsuchida, J., & Yano, M. (2001) On the effectiveness of whole spectral shape for vowel perception. *Journal of the Acoustical Society of America*, 110, 1141–9.

Johnson, K. (2004) *Acoustic and Auditory Phonetics*. Oxford: Blackwell.

Johnson, K. (2005) Speaker normalization in speech perception. In D. Pisoni & R. Remez (eds.), *The Handbook of Speech*

Perception (pp. 363–89). Malden, MA: Blackwell.

Jongman, A. (1989) Duration of frication noise required for identification of English fricatives. *Journal of the Acoustical Society of America*, 85, 1718–25.

Jongman, A., Blumstein, S. E., & Lahiri, A. (1985) Acoustic properties for dental and alveolar stop consonants: A cross-language study. *Journal of Phonetics*, 13, 235–51.

Jongman, A. R., Wayland, S., & Wong, S. (2000) Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108, 1252–63.

Joos, M. (1948) Acoustic phonetics. *Language*, 24, 1–136.

Jurafsky, D., Bell, A., & Girand, C. (2003) The role of the lemma in form variation. In N. Warner and C. Gussenhoven (eds.), *Laboratory Phonology VII* (pp. 3–34). Mouton Berlin: de Gruyter.

Kataoka, R., Warren, D., Zajac, D., Mayo, R., & Lutz, R. (2001) The relationship between spectral characteristics and perceived hypernasality in children. *Journal of the Acoustical Society of America*, 109, 2181–9.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E., & Kay, B. (1985) A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, 77, 266–80.

Kewley-Port, D. (1982) Measurement of formant transitions in naturally produced stop consonant-vowel syllables, *Journal of the Acoustical Society of America*, 72, 379–89.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (1983) Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73, 1779–93.

Kiefte, M. & Kluender, K. (2005) The relative importance of spectral tilt in monophthongs and diphthongs. *Journal of the Acoustical Society of America*, 117, 1395–1404.

Klatt, D. H. (1982) Speech processing strategies based on auditory models. In R. Carlson & B. Granström (eds.), *The Representation of Speech in the Peripheral Auditory System* (pp. 181–96). Amsterdam: Elsevier Biomedical.

Klein, W., Plomp, R., & Pols, L. C. W. (1970) Vowel spectra, vowel spaces and vowel identification. *Journal of the Acoustical Society of America*, 48, 999–1009.

Koenig, W., Dunn, H. K., & Lacy, L. Y. (1946) The sound spectrograph. *Journal of the Acoustical Society of America*, 18, 19–49.

Kohler, K. J. (1979) Parameters in the production and the perception of plosives in German and French. *Arbeitsberichte, Institute of Phonetics, University of Kiel*, 12, 261–92.

Kohler, K. J. (1985) F0 in the perception of lenis and fortis plosives. *Journal of the Acoustical Society of America*, 78, 21–32.

Krakow, R., Beddor, P., Goldstein, L., & Fowler, C. (1988) Coarticulatory influences on the perceived height of nasal vowels. *Journal of the Acoustical Society of America*, 83, 1146–58.

Krull, D. (1989) Second formant locus patterns and consonant vowel coarticulation in spontaneous speech. *Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm*, 10, 87–108.

Krull, D., Lindblom, B., Shia, B. E., & Fruchter, D. (1995) Cross-linguistic aspects of coarticulation: An acoustic and electropalatographic study of dental and retroflex consonants. *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, Sweden, 3, 436–9.

Kuehn, D. P. & Moll, K. L. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4, 303–20.

Kurowski, K. & Blumstein, S. E. (1984) Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 76, 383–90.

Kurowski, K. & Blumstein, S. E. (1987) Acoustic properties for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 81, 1917–27.

Labov, W. (2001) *Principles of Linguistic Change*, vol. 2: *Social Factors*. Oxford: Blackwell.

Ladefoged, P. (1971) *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press.

Ladefoged, P. (1985) The phonetic basis for computer speech processing. In F. Fallside & W. A. Woods (eds.), *Computer Speech Processing* (pp. 3–27). Englewood Cliffs, NJ: Prentice-Hall.

Ladefoged, P. & Bladon, R. A. W. (1982) Attempts by human speakers to reproduce Fant's nomograms. *Speech Communication*, 9, 231–98.

Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984) A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America*, 76, 391–404.

Lawrence, W. (1953) The synthesis of speech from signals which have a low information rate. In W. Jackson (ed.), *Communication Theory* (pp. 460–9). London: Butterworth.

Lehiste, I. (1964) *Acoustical Characteristics of Selected English Consonants*. Bloomington: Indiana University Press.

Lehiste, I. & Peterson, G. (1961) Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, 33, 268–77.

Liberman, A. M., Delattre, P. C. and Cooper, F. S. (1958) The role of selected stimulus variables in the perception of voiced and voiceless stops in initial position. *Language and Speech*, 1, 153–67.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954) The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68, 1–13.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6, 172–87.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839–62.

Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773–81.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling* (pp. 403–39). Dordrecht: Kluwer Academic.

Lindblom, B. (1996) Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99, 1683–92.

Lindblom, B. & Studdert-Kennedy, M. (1967) On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 42, 830–43.

Lindblom, B. E. F. & Sundberg, J. E. F. (1971) Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50, 1166–79.

Lisker, L. (1975) Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America*, 57, 1547–51.

Lisker, L. (1978) Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Reports on Speech Research*, 54, 127–32.

Lisker, L. (1986) Voicing in English: A catalogue of acoustic features signaling /b/ vs. /p/ in trochees. *Language and Speech*, 29, 3–11.

Lisker, L. & Abramson, A. S. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422.

Lisker, L. & Abramson, A. S. (1967) Some effects of context on voice onset time in English stops. *Language and Speech*, 10, 1–28.

Löfqvist, A. (1999) Interarticulator phasing, locus equations, and degree of coarticulation. *Journal of the Acoustical Society of America*, 106, 2022–30.

Löfqvist, A., Baer, T., McGarr, N., & Story, R. (1989) The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America*, 85, 1314–21.

Lubker, J. (1968) An electromyographic-cinefluorographic investigation of velar function during normal speech production. *Cleft Palate Journal*, 5, 1–18.

Mack, M. & Blumstein, S. E. (1983) Further evidence of acoustic invariance in speech production: The stop–glide contrast. *Journal of the Acoustical Society of America*, 73, 1739–50.

Malécot, A. (1956) Acoustic cues for nasal consonants. *Language*, 32, 274–84.

Mann, V. A. & Repp, B. H. (1980) Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception and Psychophysics*, 28, 213–28.

Massaro, D. W. & Cohen, M. M. (1976) The contribution of fundamental frequency and voice onset time to the /zi/–/si/ distinction. *Journal of the Acoustical Society of America*, 60, 704–17.

Milner, B. & Shao, X. (2006) Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end. *Speech Communication*, 48, 697–715.

Modarresi, G., Sussman, H., Lindblom, B., & Burlingame, E. (2005) Locus equation encoding of stop place: Revisiting the voicing/VOT issue. *Journal of Phonetics*, 33, 101–13.

Molis, M. (2005) Evaluating models of vowel perception. *Journal of the Acoustical Society of America*, 118, 1062–71.

Moll, K. L. (1962) Velopharyngeal closure on vowels. *Journal of Speech and Hearing Research*, 5, 30–7.

Moon, S.-J. & Lindblom, B. (1994) Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96, 40–55.

Munson, B. & Soloman, N. (2004) The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, 47, 1048–58.

Nearey, T. M. & Assmann, P. (1986) Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80, 1297–1308.

Nittrouer, S. (2002) Learning to perceive speech: How fricative perception changes, and how it stays the same. *Journal of the Acoustical Society of America*, 112, 711–19.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nossair, Z. B. & Zahorian, S. A. (1991) Dynamical spectral features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, 89, 2978–91.

O'Connor, J., Gerstman, L., Liberman, A. M., Delattre, P., & Cooper, F. S. (1957) Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, 13, 24–43.

Ohala, J. J. (1993) The phonetics of sound change. In Charles Jones (ed.), *Historical Linguistics: Problems and Perspectives* (pp. 237–78). London: Longman.

Ohala, J. J. & Busà, M. G. (1995) Nasal loss before voiceless fricatives: A perceptually-based sound change. *Rivista di Linguistica*, 7, 125–44.

Ohala, J. J. & Lorentz, J. (1977) The story of [w]: An exercise in the phonetic explanation for sound patterns. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 3, 577–99.

Ohde, R. N., Haley, K., & Barne, C. (2006) Perception of the [m]–[n] distinction in consonant-vowel (CV) and vowel-consonant (VC) syllables produced by child and adult talkers. *Journal of the Acoustical Society of America*, 119, 1697–1711.

Ohde, R. N. & Stevens, K. N. (1983) Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, 74, 706–14.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–68.

Öhman, S. E. G. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41, 310–20.

Palethorpe, S., Wales, R., Clark, J. E., & Senserrick, T. (1996) Vowel classification in children. *Journal of the Acoustical Society of America*, 100, 3843–51.

Palethorpe, S., Watson, C. I., & Barker, R. (2003) Acoustic analysis of monophthong and diphthong production in acquired severe to profound hearing loss. *Journal of the Acoustical Society of America*, 114, 1055–68.

Peterson, G. & Barney, H. L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175–84.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986) Speaking clearly for the hard of hearing, II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29, 434–46.

Polka, L. & Strange, W. (1985) Perceptual equivalence of acoustic cues that differentiates /r/ and /l/. *Journal of the Acoustical Society of America*, 78, 1187–97.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973) Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53, 1093–1101.

Potter, R. K., Kopp, G., & Green, H. (1947) *Visible Speech*. New York: Dover Publications.

Qi, Y. & Fox, R. A. (1992) Analysis of nasal consonants using perceptual linear prediction. *Journal of the Acoustical Society of America*, 91, 1718–26.

Recasens, D. (1983) Place cues for nasal consonants with special reference to Catalan. *Journal of the Acoustical Society of America*, 73, 1346–53.

Redford, M. & Diehl, R. (1999) The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America*, 106, 1555–65.

Repp, B. H. (1979) Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, 27, 173–89.

Repp, B. H. (1986) Perception of the m–n distinction in CV syllables. *Journal of the Acoustical Society of America*, 79, 1987–99.

Repp, B. H. (1987) On the possible role of auditory short-term adaptation in perception of the prevocalic m–n contrast. *Journal of the Acoustical Society of America*, 82, 1525–38.

Repp, B. H. & Lin, H.-B. (1989) Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America*, 85, 379–96.

Repp, B. H. & Svastikula, K. (1988) Perception of the [m]–[n] distinction in VC syllables. *Journal of the Acoustical Society of America*, 83, 237–47.

Schouten, M. E. H. & Pols, L. C. W. (1979a) Vowel segments in consonantal context: A spectral study of coarticulation, Part I. *Journal of Phonetics*, 7, 1–23.

Schouten, M. E. H. & Pols, L. C. W. (1979b) CV- and VC-transitions: A spectral study of coarticulation, Part II. *Journal of Phonetics*, 7, 205–24.

Seitz, P. F., McCormick, M. M., Watson, I. M. C., & Bladon, R. A. (1990) Relational spectral features for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, 87, 351–8.

Shadle, C. H. & Mair, S. J. (1996) Quantifying spectral characteristics of fricatives. In H. Bunnell & W. Idsari (eds.), *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)* (pp. 1517–20). New Castle, DE: Citation Delaware.

Shepard, R. N. (1972) Psychological representation of speech sounds. In E. David & D. P. Denes (eds.), *Human Communication: A Unified View* (pp. 67–113). New York: McGraw Hill.

Slis, I. & Cohen, A. (1969) On the complex regulating the voiced–voiceless distinction. *Language and Speech*, 12, 80–102.

Sluijter, A. M. C. & Heuven, V. J. van (1996) Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471–85.

Sluijter, A. M. C., Heuven, V. J. van, & Pacilly, J. J. A. (1997) Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101, 503–13.

Smiljanić, R. & Bradlow, A. (2005) Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America*, 118, 1677–88.

Smits, R. ten Bosch, L., & Collier, R. (1996a) Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants, I: Perception experiment. *Journal of the Acoustical Society of America*, 100, 3852–64.

Smits, R. ten Bosch, L., & Collier, R. (1996b) Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants, II: Modeling and evaluation. *Journal of the Acoustical Society of America*, 100, 3865–81.

Soli, S. D. (1981) Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *Journal of the Acoustical Society of America*, 70, 976–84.

Stack, J., Strange, W., Jenkins, J., Clarke, W., & Trent, S. (2006) Perceptual invariance of coarticulated vowels over variations in speaking rate. *Journal of the Acoustical Society of America*, 119, 2394–405.

Stevens, K. N. (1971) Airflow and turbulence for noise for fricative and stop consonants: Static considerations, *Journal of the Acoustical Society of America*, 50, 1180–92.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In E. Davis & D. P. Denes (eds.), *Human Communication: A Unified View* (pp. 51–66). New York: McGraw Hill.

Stevens, K. N. (1985) Evidence for the role of acoustic boundaries in the perception of speech sounds. In V. A. Fromkin (ed.), *Phonetic Linguistics* (pp. 243–55). New York: Academic Press.

Stevens, K. N. (1989) On the quantal nature of speech. *Journal of Phonetics*, 17, 3–46.

Stevens, K. N. (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111, 1872–91.

Stevens, K. N. & House, A. S. (1955) Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 27, 484–93.

Stevens, K. N. & House, A. S. (1956) Studies of formant transitions using a vocal tract analog. *Journal of the Acoustical Society of America*, 28, 578–85.

Stevens, K. N. & House, A. S. (1963) Perturbation of vowel articulations by consnantal context: An acoustical study. *Journal of Speech and Hearing Research*, 6, 111–28.

Stevens, K. N. & Klatt, D. H. (1974) Role of formant transitions in the voiced–voiceless distinction of stops.

Journal of the Acoustical Society of America, 55, 653–9.

Strange, W. (1999) Perception of vowels: Dynamic constancy. In J. Pickett (ed.), *The Acoustics of Speech Communication* (pp. 153–65). Boston: Allyn & Bacon.

Strange, W. & Bohn, O.-S. (1998) Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies. *Journal of the Acoustical Society of America*, 104, 488–504.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983) Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 695–705.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976) Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 60, 213–24.

Summerfield, A. Q. & Haggard, M. P. (1977) On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, 62, 435–48.

Sussman, H. M. (1994) The phonological reality of locus equations across manner class distinctions: Preliminary observations, *Phonetica*, 51, 119–31.

Sussman, H. M., Fruchter, D., & Cable, A. (1995) Locus equations derived from compensatory articulation. *Journal of the Acoustical Society of America*, 97, 3112–24.

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993) A crosslinguistic investigation of locus equations as a phonetic descriptor of articulation. *Journal of the Acoustical Society of America*, 94, 1256–68.

Sussman, H. M., McCaffrey, H., & Matthews, S. A. (1991) An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90, 1309–25.

Syrdal, A. K. (1985) Aspects of a model of the auditory representation of American English vowels. *Speech Communication*, 4, 121–35.

Syrdal, A. K. & Gopal, H. S. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79, 1086–100.

Tabain, M. (1998) Nonsibilant fricatives in English: Spectral information above 10 kHz. *Phonetica*, 55, 107–30.

Tabain, M. (2000) Coarticulation in CV syllables: A comparison of Locus Equation and EPG data. *Journal of Phonetics*, 28, 137–59.

Tabain, M. (2001) Variability in fricative production and spectra: Implications for the hyper- and hypo- and quantal theories of speech production. *Language and Speech*, 44, 57–94.

Tabain, M. & Butcher, A. (1999) Stop consonants in Yanyuwa and Yindjibarndi: A locus equation perspective. *Journal of Phonetics*, 27, 333–58.

Taylor, H. C. (1933) The fundamental pitch of English vowels. *Journal of Experimental Psychology*, 16, 565–82.

Terbeek, D. (1977) Cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers in Phonetics, University of California, Los Angeles*, 37.

Traunmüller, H. (1981) Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, 69, 1465–75.

Traunmüller, H. (1984) Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Communication*, 3, 49–61.

Tsao, T-C., Weismer, G., & Iqbal, K. (2006) The effect of intertalker speech rate variation on acoustic vowel space. *Journal of the Acoustical Society of America*, 119, 1074–82.

Turner, G. S., Tjaden, K., & Weismer, G. (1995) The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research*, 38, 1001–3.

Vaissière, J. (2007) Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. In M. J. Solé, P. Beddor & M. Ohala (eds.), *Experimental Approaches to Phonology* (pp. 54–72). Oxford: Oxford University Press.

van Bergem, D. R. (1993) Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12, 1–23.

van Son, R. J. J. H. (1993) Spectro-temporal features of vowel segments: Studies in language and language use. Ph.D. thesis, University of Amsterdam.

van Son, R. J. J. H. & Pols, L. C. W. (1990) Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 88, 1683–93.

van Son, R. J. J. H. & Pols, L. C. W. (1992) Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 92, 121–7.

van Son, R. J. J. H. & Pols, L. C. W. (1999) An acoustic description of consonant reduction. *Speech Communication*, 28, 125–40.

van Son, R. J. J. H. & Pols, L. C. W. (2003) How efficient is speech? *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 25, 171–84.

Wardrip-Fruin, C. (1982) On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. *Journal of the Acoustical Society of America*, 71, 187–95.

Watson, C. I. & Harrington, J. (1999) Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106, 458–68.

Weismer, G., Laures, J. S., Jeng, J.-Y., Kent, R. D., & Kent, J. F. (2000) Effect of speaking rate manipulations of acoustic and perceptual aspects of the dysarthria in Amyotrophic Lateral Sclerosis. *Folia Phoniatrica et Logopaedica*, 52, 201–19.

Whalen, D. H., Abramson, A., Lisker, L., & Mody, M. (1993) F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America*, 47, 36–49.

Winitz, H., Scheib, M. E., & Reeds, J. A. (1972) Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America*, 51, 1309–17.

Wood, S. (1986) The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. *Journal of the Acoustical Society of America*, 80, 391–401.

Wright, J. (1975) Nasal-stop assimilation: Testing the psychological reality of an English MSC. In C. A. Ferguson, L. M. Hyman, & J. J. Ohala (eds.), *Nasalfest* (pp. 389–97). Stanford: Language Universals Project, Dept. of Linguistics, Stanford University.

Wright, J. (1986) The behavior of nasalized vowels in the perceptual vowel space. In J. J. Ohala and J. J. Jaeger (eds.), *Experimental Phonology* (pp. 45–67). Orlando: Academic Press.

Wright, R. (2003) Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in Laboratory Phonology VI: Phonetic Interpretation* (pp. 75–87). Cambridge: Cambridge University Press.

Wuensch, K. (2006) Skewness, kurtosis, and the normal curve. http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm.

Zahorian, S. & Jagharghi, A. (1993) Spectral shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94, 1966–82.

4 Investigating the Physiology of Laryngeal Structures

HAJIME HIROSE

1 Introduction: Basic Laryngeal Functions

In humans, there are four basic laryngeal functions: airway protection, which is particularly important during deglutition; effort closure for fixation of the trunk while moving the upper extremities; airway opening for respiration; and phonation.

The most basic function of the larynx is to protect the airway. This function can be best understood by an appreciation of its origin determined by primitive needs (Negus, 1949). The most primitive larynx is found in the bichir lungfish (*polypterus*) which live in rivers that periodically become dry. The primary lung developed as downgrowths of the pharyngeal pouch in response to their need for oxygen under conditions where the source of supply in water is limited. Development of the lung needed to be protected from the invasion of water and food during periods of submersion and, therefore, the primary larynx evolved as a protective mechanism. In the lungfish, the larynx developed as a simple, circular group of muscle fibers within the upper end of the trachea, constituting an encircling sphincter band. When this simple sphincter closed, the lung could be effectively isolated and its closure during deglutition prevented invasion by food or water.

During the course of evolution, the encircling sphincter became a more complicated structure, and in higher animals like the human, laryngeal sphincteric closure is accomplished by a valvular adduction mechanism at both false vocal fold and true vocal fold levels.

The sphincteric closure essentially serves as a protective mechanism for the airway but it also serves those physiologic functions which are dependent on air being trapped at the larynx when accompanied by increases in or maintenance of intra-thoracic and intra-abdominal pressure. Such functions include: coughing, defecation, micturition, and fixation of the trunk for the stable movement of the upper extremities. Another important modification during evolution was the development of the laryngeal opening or abductor mechanism and the cartilagenous framework of the larynx. Thus, the larynx was able to open the airway when necessary. Basically, the glottis widens during inspiration and narrows during expiration. This movement may be almost imperceptible during quiet respiration, but it becomes more prominent as the depth of respiration increases.

Finally, phonation developed as a principal function of the larynx in which the vocal folds are used as a flutter valve. This type of flutter valve is only seen in vertebrates possessing the respiratory requirements of an effective bellows. This is possible only in vertebrates which have a diaphragm; that is, the mammals. Among all mammals, only humans have acquired the potential for the production of meaningful sounds, i.e., speech, by using the laryngeal valve as a source of vibration.

2 Methods of Investigating Laryngeal Function in Speech

Studies of laryngeal function rely heavily on the methods of investigation. In recent years, various kinds of observation techniques for the assessment and analysis of laryngeal behavior during speech production have been developed. The following is a brief description of the systems currently in use to assess laryngeal dynamics.

2.1 Fiberoptic observation and measurement of vocal fold movement

Many techniques have been used for the observation of the larynx. The most simple and popular method for otolaryngologists is the indirect mirror technique, but using this conventional laryngeal mirror, the larynx can be observed only while the subject's mouth is kept open. Even then photographs cannot easily be taken. The rigid tele-endoscope became available later, and laryngeal photography could be readily undertaken. Figure 4.1 shows a view of the larynx taken during phonation and deep inspiration using a tele-endoscope.

However, there was still a difficulty in the assessment of laryngeal dynamics during speech. In order to find out what happens during speech or singing in natural circumstances, the flexible fiberscope was devised during the late 1960s (Sawashima & Hirose, 1968). The flexible fiberscope basically consists of a hard tip that houses an objective lens and two bundles of glass fibers: the light guide and the image guide. The light guide conducts the light for illumination of the field of view from the light source to the object-end of the scope. The image guide is a bundle of aligned or coherent glass fibers which transmits the image from the objective lens to the eye-piece of the scope.

Specific requirements in the design of the fiberscope were: (1) that it have an outside diameter small enough to pass through the nostril; (2) that it be able to obtain an image with a resolution good enough for the analysis of glottal gestures; and (3) that it should be provided with a light source of sufficient brightness (Sawashima, 1977). In recent years, these requirements have generally been satisfied.

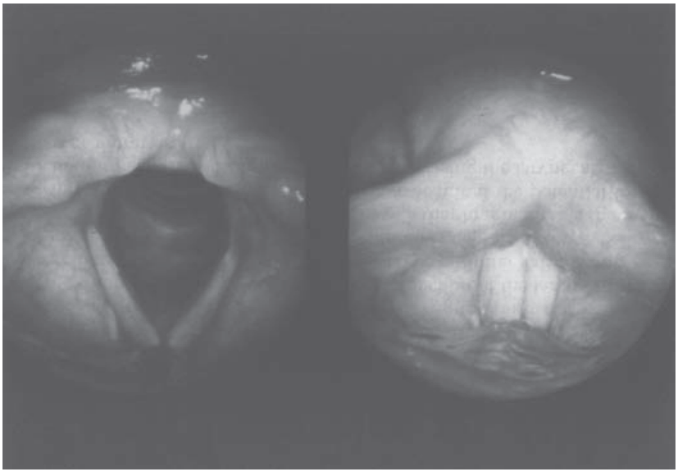


Figure 4.1 The laryngeal views obtained using a rigid tele-endoscope.

Prior to the insertion of the fiberscope, surface anesthesia is applied to the nasal mucosa and to the epipharyngeal wall. After insertion, the tip of the scope is placed down near the tip of the epiglottis to obtain a good laryngeal view. Figure 4.2 illustrates the positioning of the fiberscope in an adult male.

2.2 High-speed digital imaging of vocal fold vibration

Until recently, precise observations of the pattern of vocal fold vibration have generally been made using an ultra-high-speed movie system. Ultra-high-speed photography can provide good resolution images of the vibrating vocal folds. However, this system is usually massive and costly, and it is always very time-consuming to carry out frame-by-frame analysis of the film obtained.

In the late 1980s, a new method of digitally imaging vocal fold vibration was developed using a solid-state image sensor attached to a conventional camera system (Hirose, 1988). Figure 4.3 shows a block diagram of the present system. An oblique-angled laryngeal endoscope is attached to a specially designed camera containing an image sensor and a digital image memory. The laryngeal image obtained through the endoscope is focused on the image sensor. When the shutter is released, the image sensor is scanned at a high-frame-rate, and the output video siginal from the image sensor is fed into the image memory through a high-speed A/D converter. Stored images are then reproduced consecutively on a cathode ray tube monitor as a slow-motion display. At present, the new model achieves frame rates of up to 4,500 per second.

In experimental situations, the electroglottographic (EGG) and speech signals are digitized simultaneously with the image recording by a separate personal computer. These signals are sampled synchronously with the image recording

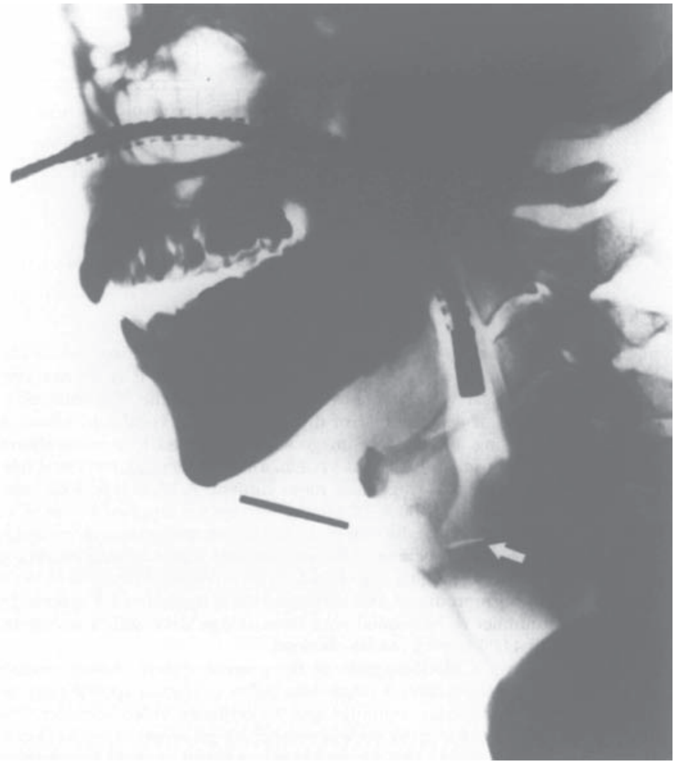


Figure 4.2 Positioning of the fiberscope for laryngeal observation.

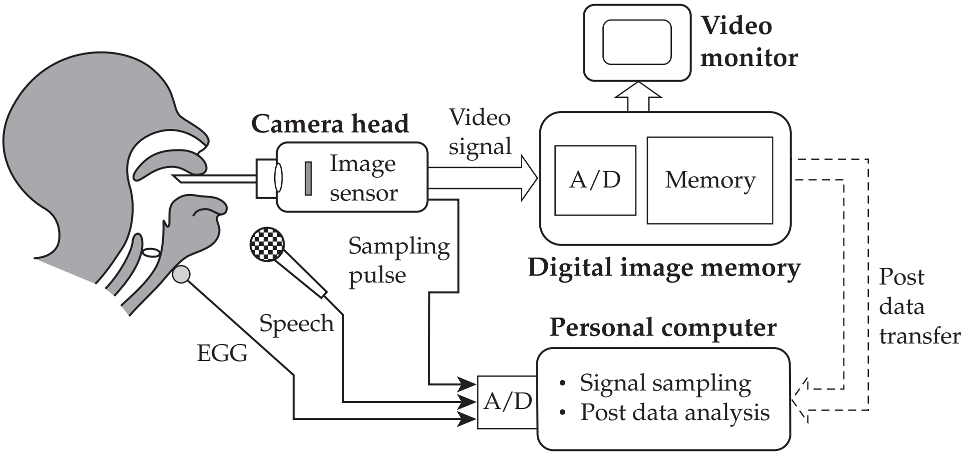


Figure 4.3 Block diagram of the digital imaging system for the analysis of vocal fold vibration.

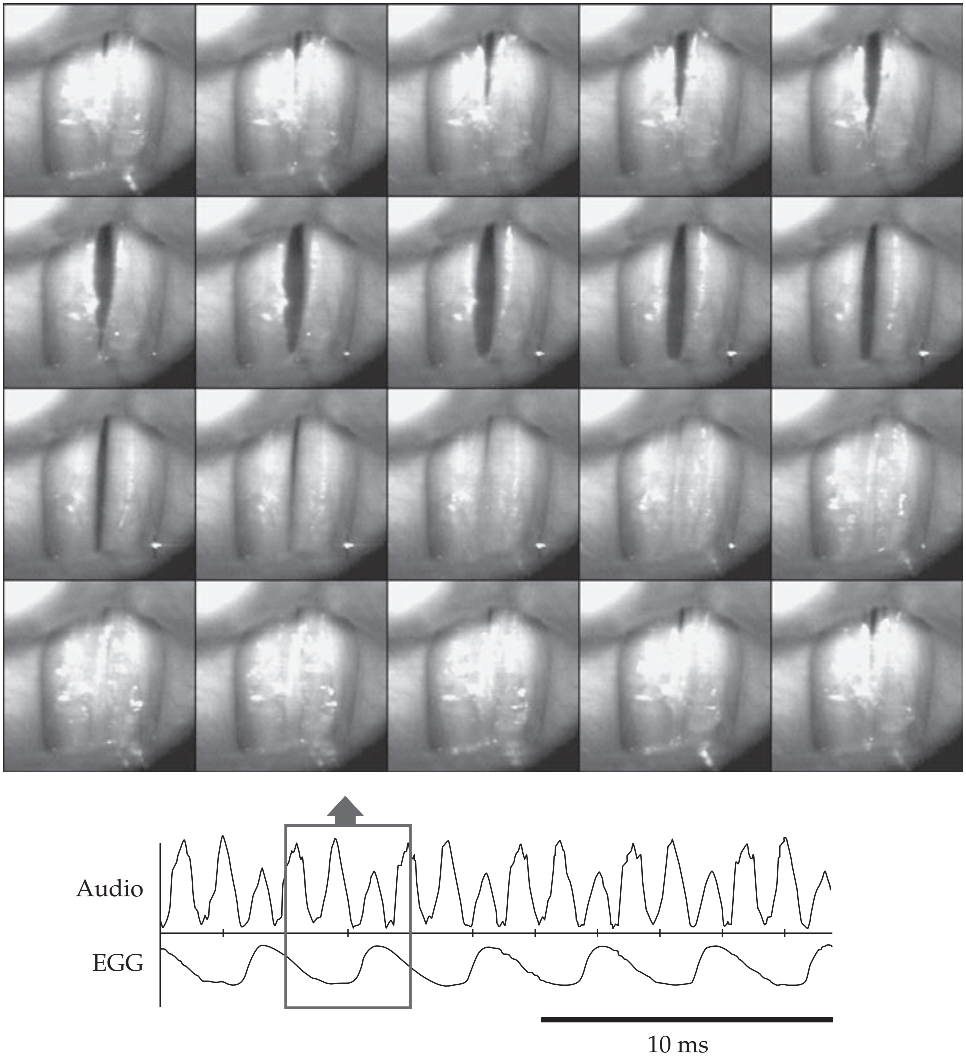


Figure 4.4 An example of laryngeal images recorded using the digital imaging system. A female subject sustained phonation of /e/ at a fundamental frequency (f_0) of approximately 240 Hz. Frame rate: 4,500 frames per second.

using sampling pulses generated in the camera head. At the time of the reproduction of the image data, the sampled EGG and speech data can be displayed simultaneously with the laryngeal images.

Figure 4.4 shows an example of the image display. The display shows the vocal fold vibration during sustained phonation of the vowel /e/ by a normal female subject at approximately 240 Hz. The glottal images in successive time frames are displayed as a two-dimensional array. The two curves shown below the pictures are the acoustic signal at the top and EGG signal at the bottom. The image at the upper left corner is the first frame in the time sequence. The following four frames are displayed in the same horizontal row, and the succeeding frames are displayed in the next row. The image at the lower right corner is the last frame. The recording was made at the rate of approximately 4,500 frames per second, and the rectangle shown on the two simultaneously recorded signals corresponds to about one cycle of vibration.

A pilot system using a fiberscope has also been developed more recently (Hirose et al., 1988; Kiritani et al., 1995). In this system, a charge-coupled device (CCD)-type image sensor is used. The light source for the fiberscope system is a 300 W xenon lamp. The sampling rate of the picture elements is 20 MHz. A film rate of 2,000 per second can be achieved with 200 × 14 picture elements. This type of system makes it possible to observe vocal fold vibration during speech samples containing consonantal gestures.

2.3 Laryngeal electromyography

Electromyography (EMG) is a technique for providing graphic information about the time course of the electrical activity of the muscle fibers that accompanies muscle contraction and subsequent effects, including the development of tension (see Stone, this volume)

Since EMG was established as a scientific discipline, it has been widely used in various fields for studying muscular function and coordination. In particular, EMG has proved to be useful for research into kinesiological aspects of human behavior, where the analysis of the parameters of the individual motor unit action potential may not play an important role. Rather, EMG kinesiology is much more concerned with the biomechanical analysis of various movements or gestures (Harris, 1981).

The EMG system consists of some sort of probe or electrodes for picking up the action potentials, amplifying equipment, recording equipment, and ultimately a graphic display, which may have signal-processing facilities. For laryngeal EMG in the study of speech dynamics, so-called hooked-wire electrodes are used, in which a pair of thin electrically shielded wires are threaded through a needle and inserted in the target muscles (Hirose, 1985). Intrinsic laryngeal muscles, such as the cricothyroid, thyroarytenoid, and lateral cricoarytenoid muscles are reached percutaneously as are two extrinsic laryngeal muscles, the sternohyoid, and sternothyroid muscles. The posterior cricoarytenoid and interarytenoid muscles are reached perorally with indirect laryngoscopy using a specially designed curved probe (Hirose, 1976). As shown in Figure 4.5, the wire-bearing tip of the needle is kept drawn into the shaft of the probe until it is brought closely to the point of insertion, at which time it is pushed out by pulling the trigger of the probe.

Laryngeal EMG technique is also useful for the analysis of nonspeech gestures such as sniff, cough, or throat clearing (Poletto et al., 2004).

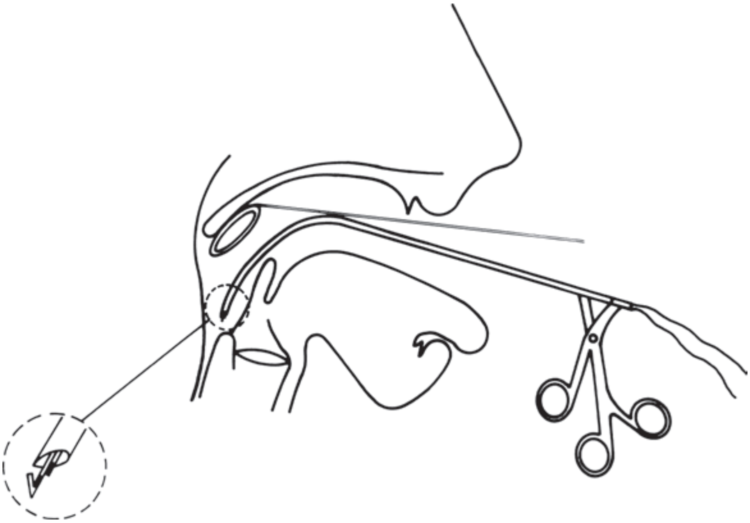


Figure 4.5 Peroral insertion of wire electrodes using a curved probe under indirect laryngoscopy.

It should be emphasized that progress in the strategy of the computer processing of EMG data has led to better analysis of the temporal pattern of the activity of pertinent laryngeal muscles with reference to speech signals (Kewly-Port, 1973).

2.4 Photoglottography (transillumination of the glottis)

Photoglottography (PGG) is a technique for recording glottal area variation by measuring the amount of light passing through the glottis.

In 1960, Sonesson first reported the use of a photo-electric device applied to the normal human subject for assessing the glottal area variation. In his method, a DC light source was placed against the anterior neck, while a light-conducting rod was inserted into the hypopharynx through the mouth under topical anesthesia. A photomultiplier tube was attached to the other end of the rod so that the illuminating light passing through the glottal aperture was transmitted through the light-conducting rod to the photomultiplier tube. The output of the tube was displayed on a cathode ray oscilloscope and the record was called a photo-electric glottogram or photoglottogram. Using this technique, he measured the open period, the opening phase, and the closing phase of the glottal vibratory cycle for sustained phonation. He claimed that the results obtained from his method were in good agreement with the results obtained from high-speed motion picture analysis.

Since Sonesson's technique imposed considerable limitations on articulatory movements, further modification was made by other investigators. For example, Frøkjær-Jensen (1967) introduced a small photo-sensor attached to the tip of the thin flexible plastic tube through the nasal cavity to the hypopharynx, thus making transillumination possible during speech articulation. Sawashima (1968) reversed the positions of light source and photo-sensor relative to the glottis. He used a fiberscopic illumination as a light source while observing the laryngeal gesture, and picked up the photo-electric signals through a photo-sensor attached to the anterior neck.

These modifications have extended the application of the photoglottographic technique to studies on glottal adjustments as well as on the patterns of vocal fold vibration during speech production. It has been assumed that the data obtained by this technique provides a good approximation of the glottal area function, although it is impossible to calibrate the instrument to measure the absolute area of the glottis. Also, it should be taken into consideration that several sources of artifacts may exist during data assessment. A shift in the positioning of the instruments relative to the larynx may be a major source of artifacts. Interruption of the light by the epiglottis during speech utterance should also be carefully monitored during recording to minimize incorrect interpretation of the obtained results.

2.5 *Electroglottography (laryngography)*

Electroglottography (EGG) is a technique for registering glottal vibratory movements by measuring changes in electrical resistance across the neck. In this technique, a pair of plate electrodes are placed on the skin on both sides of the neck above the thyroid cartilage. A weak high-frequency electrical current is applied to the electrodes, and a small fraction of the electrical current passes through the larynx. The transverse electrical resistance of the larynx varies depending on the opening and closing of the glottis, and a modification in the amplitude of the transglottic current occurs in correspondence with the vibratory cycles of the vocal folds. The amplitude modification of the current is detected, from which electrical glottograms are obtained.

In a typical model described by Fourcin (1981), each electrode has a guard ring and an inner conductor. One of the electrodes has a 4 MHz transmitting voltage applied between the center conductor and guard ring. The other serves as a current pick-up. According to Fourcin, typically about 30 mW is dissipated at the subject's neck with only microwatts being involved at the level of the vocal folds. Contact between the vocal folds increases current flow as the contact area increases, but movement of the vocal folds without contact, giving an increase in glottal area, will not necessarily change the current flow. For this reason, Fourcin claimed that the term "glottograph" is inappropriate, and he proposed that it should be called a "laryngograph."

In making comparisons between electrical and photo-electric glottograms, Frøkjær-Jensen (1968) concluded that the opening of the glottis seemed to be better represented in photo-electric glottograms, whereas the closure of the glottis, particularly its vertical contact area, was probably better reflected in EGG. One of the advantages of EGG is that the procedure is carried out with a minimum of discomfort for the subject. As stated above, an EGG record reflects the glottal condition during closure better than during the open period, and the presence or absence of glottal vibration, as well as the accurate fundamental frequency, can be readily determined. However, since it is difficult to estimate to what extent the glottal condition contributes to the electrical resistance or impedance variations between the electrodes, a quantitative interpretation of EGG seems to be less direct than PGG.

2.6 New imaging techniques including Magnetic Resonance Imaging (MRI)

X-ray imaging techniques were adapted for use in movie and video recordings from the 1960s to the 1970s, mainly for the analysis of the laryngeal positioning during phonation or deglutition. However, the invasiveness of the X-ray imaging has been a concern since that time. Recently, less invasive techniques have been developed and, among others, the magnetic resonance imaging (MRI) technique appears to be safe and have substantial advantages in offering superior temporal and spatial image resolution over X-ray imaging.

In the recording session of MRI, the subject is lying prone in a tube and a strong magnetic field surrounds the subject. It aligns atoms in the target structures and radio frequency pulses tip the atoms off their axes of spin. As they return to their previous axes, they give off a signal that can be recorded on a photographic plate. The image shows the distribution of hydrogen atoms in the pictured tissue. MRI has been applied mostly for the imaging of the central nervous system, but the technique has made it possible to obtain the precise shape and area function of the vocal tract during phonation. Although the MRI technique cannot be applied for the analysis of vocal fold vibration per se, it shed light on the analysis of the role of individual variations in the fine structure of the vocal tract for speaker characteristics (Kitamura et al., 2005). MRI analysis was also applied for the analysis of the laryngeal adjustment mechanism of whispering (Tsunoda et al., 1997).

3 Laryngeal Structures and the Control of Phonation

3.1 Laryngeal framework and laryngeal muscles

The framework of the larynx consists of four different cartilages: the epiglottis, thyroid, cricoid, and arytenoid cartilages. The thyroid and cricoid cartilages are connected by the cricothyroid joint, while the arytenoid and cricoid cartilages are connected by the cricoarytenoid joint. The movement of the cricothyroid joint changes the length of the vocal folds. Movements of the arytenoid cartilage on the surface of the cricoarytenoid joint contribute to the abduction/adduction of the vocal folds. The main movement of the cricoarytenoid joint is a rotation (abduction/adduction) of the arytenoid cartilage around the longitudinal axis of the joint. Other possible movements of the arytenoid are a small degree of sliding motion along the longitudinal axis of the joint and a rocking motion around a fixed point at the attachment of the posterior cricoarytenoid ligament (von Leden & Moore, 1961). (For further details of the anatomy of laryngeal structures, see Bless & Abbs, 1983; Kahane & Folkins, 1984; and Hirano, 1991.)

Movements of the cricothyroid and cricoarytenoid joints are controlled by the intrinsic laryngeal muscles. Elongation and stretching of the vocal folds is achieved by contraction of the cricothyroid muscle (CT). Movements of the arytenoid cartilage and the resultant abduction/adduction of the vocal folds are controlled by the abductor and adductor muscles. The posterior cricoarytenoid muscle (PCA) is the only abductor muscle, while another three – the interarytenoid (INT or IA), lateral cricoarytenoid (LCA), and the thyroarytenoid (TA) muscle – are the adductor muscles. Contraction of the cricothyroid muscle may also result in a small degree of glottal abduction. The vocalis muscle (VOC), which is the medial part of the thyroarytenoid muscle, contributes to the control of the effective mass and stiffness of the vocal folds rather than to abduction/adduction movements.

The entire larynx is supported by the extrinsic laryngeal muscles and the ligaments, of which suprahyoid and infrahyoid muscles form the important members. These muscles contribute to the elevation and lowering of the larynx, which may relate to the pitch control of voice, as well as to articulatory adjustments such as jaw opening (Erickson et al., 1977).

3.2 Layered structure of the vocal fold

The layered structure of the vocal fold edge described by Hirano (1974) is shown in Figure 4.6. As can be seen in the figure, the vocal fold consists of the mucosa epithelium, the lamina propria mucosa, and the vocalis muscle. In the lamina propria, the superficial layer is the loose connective tissue, and the intermediate and deep layers correspond to the so-called vocal ligament. Based on the concept of this layered structure, Hirano proposed a structural model of the vocal fold. In his model, the vocal fold basically consists of the three layers – cover, transition, and body. The cover consists of the epithelium and the superficial layer of the lamina propria; the transition includes the intermediate and the deep layers; and the body includes the vocalis muscle. For simplification, the transition can be considered as part of the body so that the entire structure can be regarded as cover and body.

This cover–body model proposed by Hirano is quite useful for explaining variation in the mode of vocal fold vibration with different laryngeal adjustments and with various pathological conditions. Contraction of CT elongates the vocal fold, and its effective mass decreases. Due to the elongation of the vocal fold, the stiffness of both cover and body increases. This is the situation of the vocal

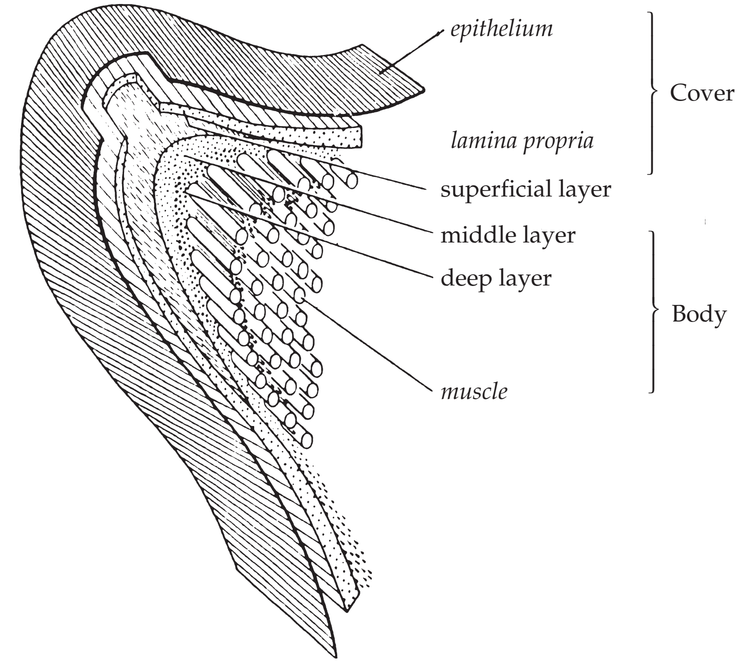


Figure 4.6 Schematical presentation of the layered structure of the human vocal fold.

fold for phonation in the light or head register. Contraction of VOC, in contrast, shortens the vocal fold, its effective mass being increased. At the same time, stiffness of the body increases, while that of the cover decreases. Contraction of VOC in combination with different degrees of contraction of CT usually takes place for phonation in the modal or chest register. Thus the difference in the mode of vocal fold vibration between the head and the chest registers can be accounted for by the different conditions of the cover and body of the vocal fold (Hirano, 1974).

3.3 Vocal fold vibration during phonation

According to the almost universally accepted myoelastic-aerodynamic theory of vocal fold vibration during phonation, one cycle of the vibration of the vocal fold is produced as follows (see also Stevens & Hanson, this volume).

1 The bilateral vocal folds are appropriately approximated toward the midline by the activation of the adductor laryngeal muscles accompanied by suppression of the abductor muscle.
2 Air is then forced through the vocal tract from the lungs and the vocal folds are sucked together by the combined effect of Bernoulli's aerodynamic law and the elasticity of the tissues (see Shadle, this volume).

3 When the vocal folds have been sucked together, the flow of air from the lungs continues but the flow through the glottis ceases and the subglottal air pressure rises.
4 When the subglottal air pressure becomes greater than the medial compression of the vocal folds, the folds are blown apart and a puff of air escapes into the supraglottal space. Consequently, the subglottal pressure falls and the vocal folds return to their adducted position at the beginning of the vibratory cycle as a result of their tissue elasticity.
5 A second cycle starts as a repetition of the first cycle.

Several preconditions are required for normal phonation. The transglottal pressure (the difference between the subglottal and supraglottal pressure) and the airflow must be high enough, the glottal width small enough and the glottal resistance sufficiently low.

4 Laryngeal Adjustments for Different Phonetic Conditions

The basic features of laryngeal adjustments for different phonetic conditions can be classified as follows:

1 abduction vs. adduction of the vocal folds;
2 constriction of the supraglottal structures;
3 adjustment of the length, stiffness, and thickness of the vocal folds;
4 elevation and lowering of the entire larynx.

4.1 Abduction vs. adduction of the vocal folds

This type of adjustment is used for the distinction between respiration and phonation, as well as for the voiced vs. voiceless distinction during speech production. For deep inspiration, the vocal folds are fully abducted by an increase in the activity of PCA and a suppression of the adductor muscles. For quiet respiration, the extent of the glottal opening is approximately half that for deep inspiration and the vocal fold position observed in laryngoscopy in quiet respiration is described as the intermediate position. In this condition the activities of both the abductor and the adductor muscles are minimal.

The general picture of the glottal condition in the abduction vs. adduction dimension during speech is that the glottis is closed or nearly closed for voiced sounds including vowels, whereas it is open for voiceless sounds, the degree of the glottal opening and its timing relative to the articulatory gestures varying with different phonetic environments.

The principal mechanism underlying abduction vs. adduction of the vocal folds during speech production is reciprocal activation of the abductor and adductor

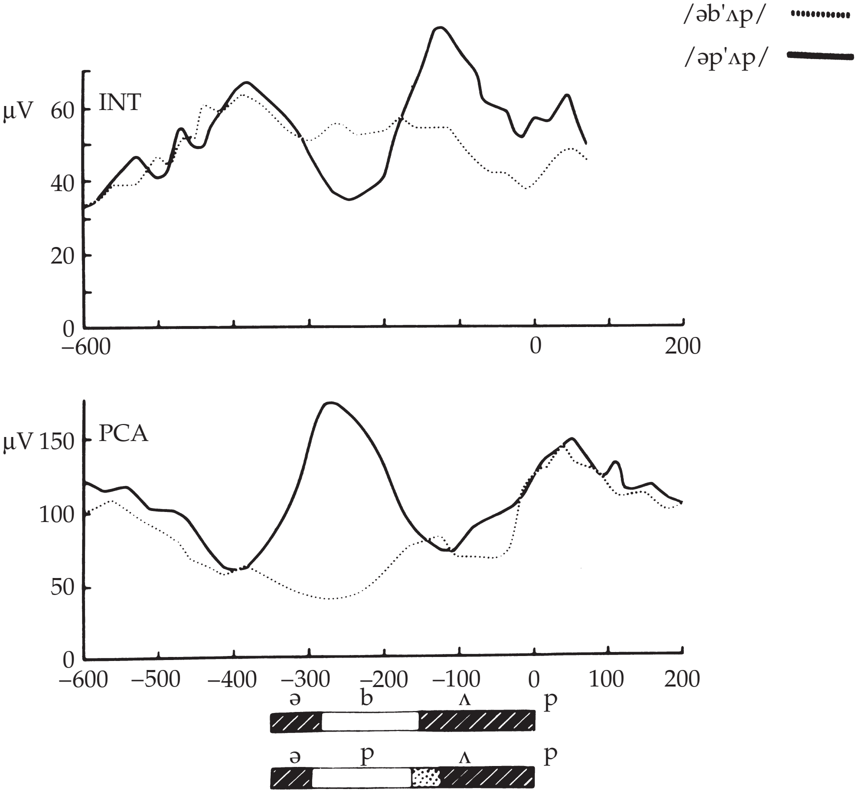


Figure 4.7 Superimposed averaged EMG curves of INT and PCA for the utterances /əp'ʌp/ (solid line) and /əb'ʌp/ (dotted line). The line-up point for averaging (zero on the abscissa) indicates the voice offset of the stressed vowel. (From Hirose & Gay, 1972, with kind permission of S. Karger, Basel)

muscle groups. The reciprocal activity pattern between the two groups of laryngeal muscles has been revealed by recent EMG studies combined with fiberoptic observation. In particular, reciprocity between PCA and INT is found to be important for realization of the voiced–voiceless distinction. The reciprocity between PCA and the adductor muscles has been observed for different languages, including American English (Hirose & Gay, 1972), Japanese (Hirose & Ushijima, 1978), Danish (Hirose et al., 1979) and French (Benguerel et al., 1978).

Figure 4.7 shows an example of averaged EMG curves of the INT and PCA, for a pair of test words /əp'ʌp/ and /əb'ʌp/ produced by an American English speaker. It can be seen that PCA activity is suppressed for the voiced portion of the test words, whereas it increases for the production of the intervocalic voiceless stop /p/ as well as for word-final /p/. On the other hand, INT shows a reciprocal pattern when compared with that of PCA, in that its activity increases for the voiced portion and decreases for the voiceless portion of the test words.

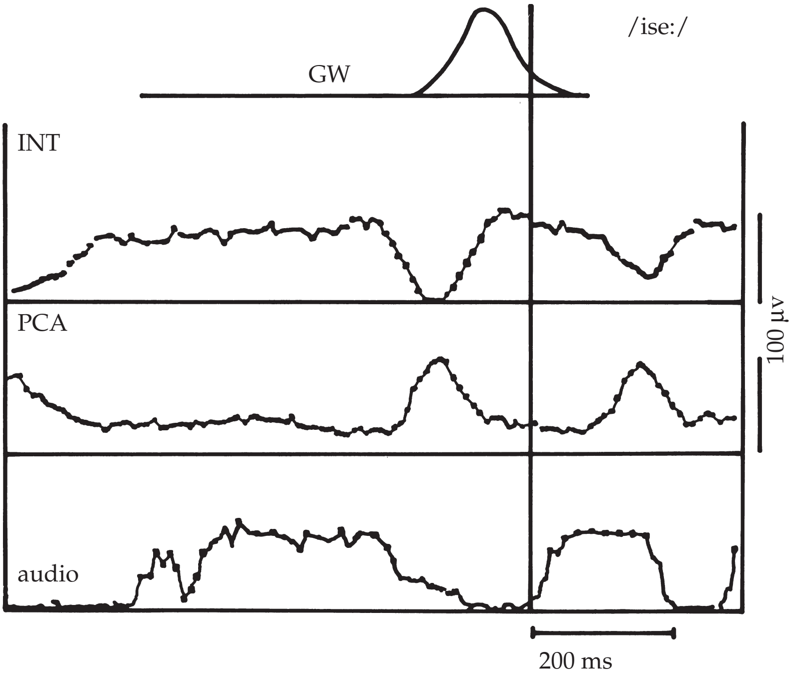


Figure 4.8 Time curves of the glottal width (GW), the smoothed and integrated EMG curves of the INT and PCA, and the speech envelope (audio) for the test word /ise:/ produced by a Japanese subject. The curves are aligned on the same time axis. The vertical line indicates the voice onset for the vowel /e/.

Figure 4.8 shows a typical example of the relationship between the glottal size and the pattern of the averaged laryngeal EMG activity of PCA and INT for the production of the Japanese test word /ise:/. The glottal width (GW), measured by means of fiberoptic analysis, increases for the voiceless consonant /s/, for which PCA activity increases and INT activity is reciprocally suppressed.

Some languages, such as Hindi and Chinese, show a phonemic distinction between aspirated and unaspirated stops. Previous EMG and fiberoptic studies revealed that the degree and timing of glottal abduction–adduction gestures are well controlled by coordinated laryngeal muscle activities (Sawashima & Hirose, 1983). In particular, the degree and timing of PCA activation seem quite important for the distinction between different phonemic types associated with glottal opening, i.e., arytenoid separation at the vocal processes observed by a fiberscope.

Figure 4.9 shows the relationship between the pattern of PCA activity and the time course of the glottal width measured at the vocal process, for the three labial stop types showing arytenoid separation: voiceless aspirated, voiceless unaspirated, and voiced aspirated. The curves are lined up at the articulatory release taken as time 0 on the abscissa, and durations of oral closure and aspiration are also

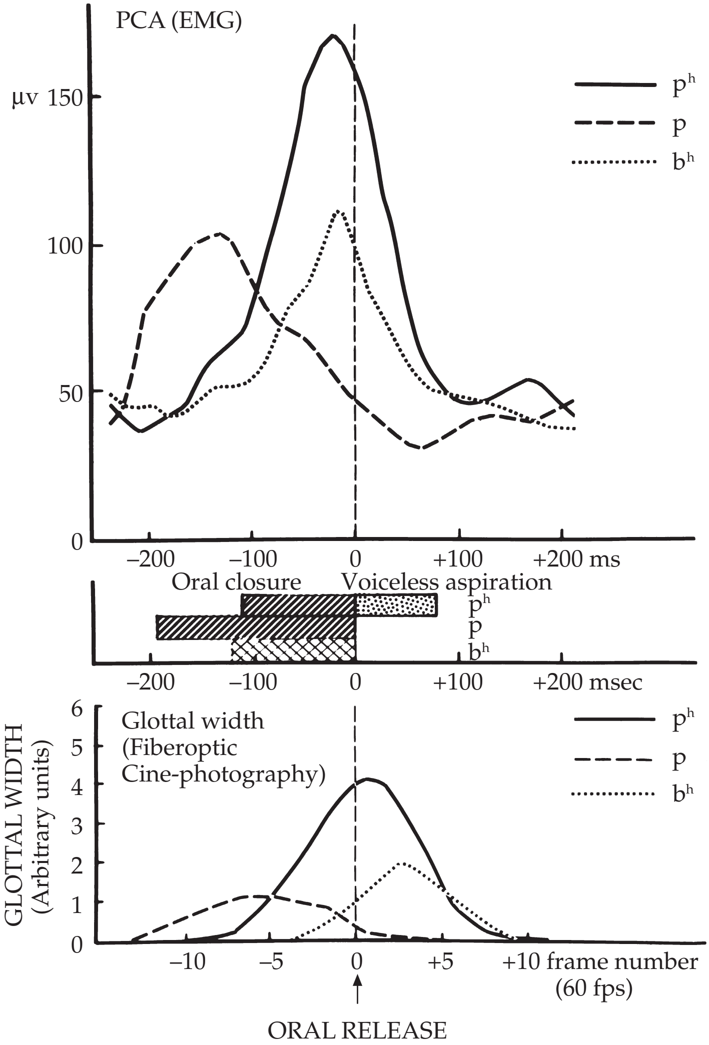


Figure 4.9 Comparison of the time courses between averaged PCA activity and glottal opening gesture. All curves are lined up at the oral release.

illustrated (Hirose, 1977). The figure shows good agreement not only in degree but also in timing between PCA activity and the opening gesture of the glottis. Thus, we must fully realize that, in addition to the control of the degree of glottal abduction vs. adduction, the control of laryngeal timing is also essential in phonetic realization of different types of consonants. As explicitly discussed by Abramson (1977), various languages of the world make extensive use of the timing of the valvular action of the larynx relative to supraglottic articulation in order to distinguish classes of consonants, although certain nonlaryngeal features such as pharyngeal expansion may also be linked with laryngeal timing.

It should be noted, however, that adjustment of glottal width is only one parameter that determines whether or not the vocal folds will vibrate during the consonantal interval. In addition, there must be an adequate glottal airflow through the glottis for generating vocal fold vibration, the amount of which will depend on both subglottal pressure and on the configuration of the supraglottal articulators. Further, the physical properties of the vocal folds, particularly the stiffness, is an important factor that relates to initiation–cessation as well as the mode of vocal fold vibration.

In order to clarify the relationship between transglottal pressure difference and the glottal configuration during the production of voiceless consonants, a physiological experiment was performed in which the sub- and supraglottal pressure was measured by means of pressure transducer systems and the glottal size was estimated using the photoglottography technique (Löfqvist & Yoshioka, 1980). The data were obtained at the offset of the vibration at oral closure of voiceless consonants /s/ and /t/, at the onset of the vibration after the oral release, and during the maximum glottal opening for each consonant. The transglottal pressure (ΔP) was calculated by subtracting the subglottal pressure value from the supraglottal value.

Figure 4.10 shows the relationship between the ratio of the transglottal pressure to the subglottal pressure (ΔP/Pa) and the relative size of the glottal width (GW) for word-initial /s/ and word-initial /t/. In this figure, the 90 percent range of the distribution is represented by circles for each of the following sets of data: the voice offsets for /s/ and /t/, and voice onsets after /s/ and /t/.

It can be seen here that both /s/ and /t/ demonstrate a difference in the physiological conditions for the cessation and initiation of voicing related to obstruent production. Namely, in both cases, voicing following the consonantal closure period occurred with a relatively small glottis and a higher ΔP/Ps ratio compared to those values with which voicing ceased around the implosion of the consonant.

It can also be seen that there is a subtle difference in the patterns of the distribution of data between the fricative /s/ and stop /t/ in terms of the laryngeal conditions for voice offset. In the case of /s/, the vocal fold vibration ceases with a relatively wider glottis than for /t/, whereas the ΔP/Ps ratio is comparable. On the other hand, there is no apparent difference between /s/ and /t/ distribution for the initiation of vocal fold vibration.

Thus, it appears that there is a hysteresis in the glottal mechanism defined by the initiation and cessation of oscillation. That is, vocal fold vibration tends to be maintained at the implosion of obstruents with relatively favorable physiological conditions for oscillation, while vibration does not start after the voiceless period until more favorable conditions are obtained by a narrowing of the glottis. These more favorable conditions are associated with an elevation of the transglottal pressure difference, although the reason why the vocal folds continue to vibrate with a wider glottis for /s/ than for /t/ is sti1l unclear (Hirose & Niimi, 1987).

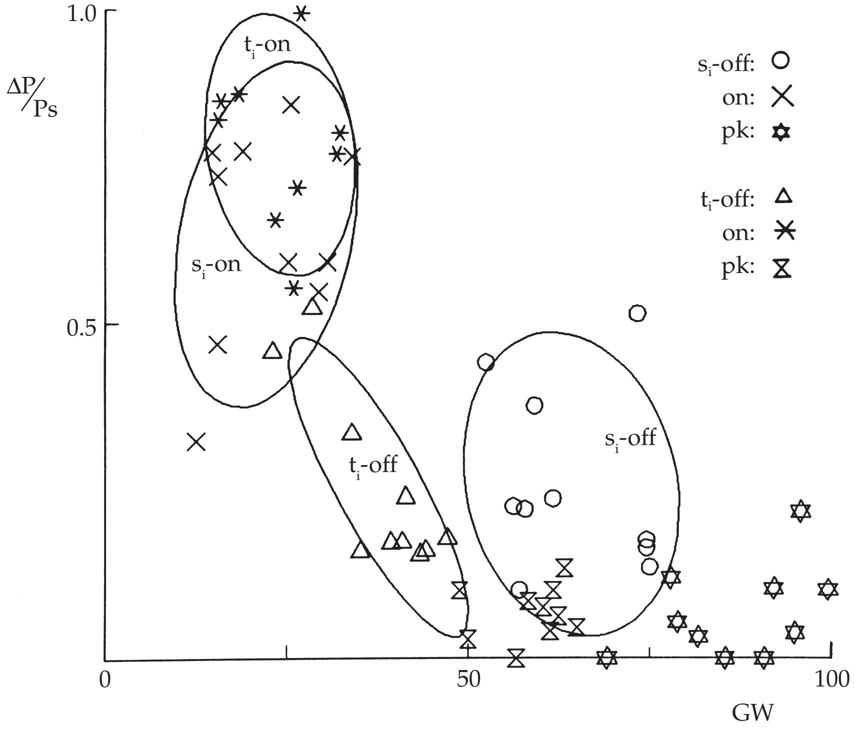


Figure 4.10 Pattern of data distribution for word-inital /s/ and /t/ representing the relationship between transglottal pressure (ΔP) vs. subglottal pressure (Ps) ratio and relative glottal width (GW) (the largest glottal opening during the consonantal period of [s] was taken as 100 percent, and relative glottal width was calculated as a percentage of that value for each token). In the figure, the 90 percent range of distribution is circled for each of the following datasets: voice offset for /s/ and /t/ (s_i-off and t_i-off) and voice onset after /s/ and /t/ (s_i-on and t_i-on), respectively. The symbol for "pk" indicates the coordinate for values at the time of maximum glottal opening for each token.

4.2 Constriction of the supraglottal structures

A typical example of supraglottal laryngeal constriction with the open glottis is observed in whispered phonation. In whisper, there is arytenoid separation at the vocal process with an adduction of the false vocal folds taking place with a decrease in the size of the anterior–posterior dimension of the laryngeal cavity. For this type of laryngeal adjustment, PCA continues to be active and the thyropharyngeal activation is also observed most likely for realization of supraglottal constriction (Tsunoda et al., 1994). This particular gesture for whispering is considered to contribute to the prevention of the vocal fold vibration by the transglottal airflow, as well as to facilitate the generation of turbulent noise in the laryngeal cavity.

Supraglottal laryngeal constriction with closed glottis is typically observed for glottal stop production. A similar gesture is often seen for the syllable-final stops in American English (Fujimura & Sawashima, 1971). The gesture prevents the air from the lungs from passing through the glottis. In laryngeal EMG, it has been observed that LCA appears to show a high degree of activity for this particular gesture together with activation of TA.

A lesser degree of supraglottal constriction with the closed glottis can be regarded as characterizing the laryngeal gesture known as "laryngealization." This type of adjustment may be observed for the production of Korean forced or tense stops and the so-called stød in Danish, where strong activation of VOC has been reported (Sawashima & Hirose, 1983).

4.3 *Adjustment of the length, stiffness, and thickness of the vocal folds with respect to pitch control*

The best example of this type of laryngeal adjustment is control of the pitch of the voice, f_0, during phonation. f_0 control at the larynx is considered to be achieved mainly by adjusting the effective mass and the stiffness of the vocal folds. The main contributor to pitch regulation is CT, while TA also appears to participate to some extent. The activity of CT increases to raise pitch and decreases to lower pitch. As mentioned earlier, contraction of CT elongates the vocal folds, resulting in a decrease in the effective vibrating mass and an increase in the stiffness of both the cover and body of the vocal folds. Contraction of TA results in a thickening of the vocal folds, their effective mass being increased. The stiffness of the body increases while that of the cover decreases. It has been observed that in the chest or modal register, a rise in pitch is characteristically achieved by contraction of both CT and TA. The most remarkable difference in muscle control between the chest and head registers is observed in the activity of TA. In the head register, as compared to the chest register, there is a marked decrease in TA activity, accompanied by an increase in CT activity. The difference in the muscle control between the two registers results in a difference in the physical conditions of the cover and body of the vocal folds, which is reflected in the mode of vocal fold vibration (Hirano et al., 1970).

In the realization of pitch accent in Japanese, different types of tones in tone languages such as Chinese, and word stress in English and other languages, CT is found to be uniquely related to f_0 changes. In particular, the increase in longitudinal tension and stretch of the vocal folds is obtained by CT activation. Figure 4.11 compared the curves of averaged CT activity and f_0 contours for five test words having different stress positions. It is obvious for all words that CT activation occurs slightly ahead of the pitch peak associated with the stressed syllable.

Although the mechanism of pitch elevation seems quite clear, the mechanism of pitch lowering is not so straightforward. The contribution of the extrinsic laryngeal muscles such as sternohyoid is assumed to be significant, but their activity often appears to be a response to, rather than the cause of, a change in conditions. The activity does not occur prior to the physical effects of pitch change.

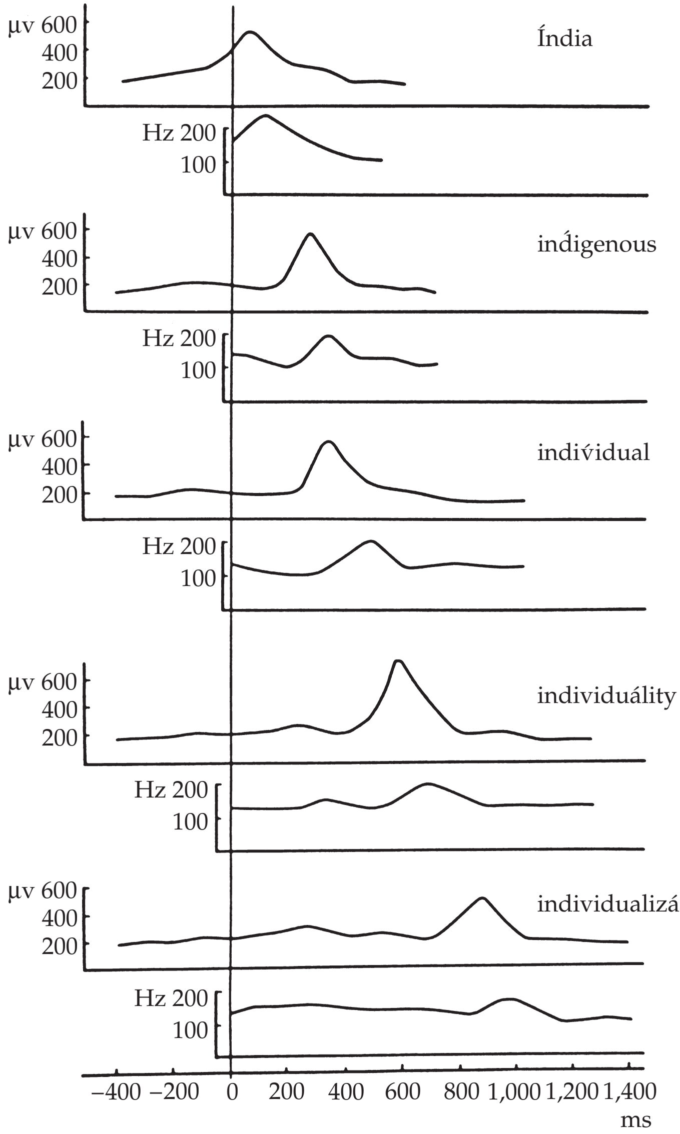


Figure 4.11 Comparison of the time courses between the averaged EMG curves of CT and f_0 contours for test words having different stress positions.

4.4 Elevation and lowering of the entire larynx

This type of laryngeal adjustment is typically observed in the action of swallowing, as well as during speech for vocal pitch control and voiced vs. voiceless distinction.

Recently, Honda and his colleagues (1999) proposed a mechanism of f_0 control by vertical larynx movement based on the measurement of magnetic resonance images (MRI). They claim that the larynx moves vertically in f_0 changes along the cervical spine, which displays anterior convexity (lordosis) at the level of the larynx, and the vertical larynx movement results in the rotation of the cricoid cartilage and vocal fold tension changes. In their MRI analysis, they observed that the hyoid bone moved horizontally while the larynx height remained relatively constant in the high f_0 range. In the low f_0 range, on the other hand, the entire larynx moved vertically, and the cricoid cartilage rotated along the cervical lordosis. They concluded that these results would indicate the vertical movement of the larynx comprises an effective f_0 lowering mechanism. Further study appears to be needed to explore the precise f_0 lowering mechanism related to the vertical larynx movement in speech production.

Also, the contribution of vertical larynx movement for phonetic distinctions still needs to be investigated, except for specific laryngeal adjustment such as ejective and implosive sound production in which the entire larynx is elevated or lowered respectively, and for generating or maintaining vocal fold vibration while the vocal tract is closed.

5 Current Main Issues and the Direction of Future Research

The science of speech production is an inherently interdisciplinary endeavor. Thus, in recent years, multidisciplinary approaches including physiological, engineering, and linguistic aspects have attempted to disclose the fine nature of laryngeal behavior in voice and speech production. For the purpose of facilitating the exchange of information among different research domains, a series of conferences on vocal fold physiology have been held since 1981, and the latest conference was held in 2008.

In the domain of physiological research, simultaneous recordings of multiple parameters are widely performed. For example, ultra-high-speed observation of the vocal fold vibratory pattern was made in combination with precise acoustic measures together with the assessment of other physiological parameters such as EGG (Childers et al., 1983). From an engineering standpoint, numerical simulation and modeling of the voice source based on physiological data were often reported (Bickley, 1991; Cranen, 1991).

Another important issue is to investigate the nature of pathological voice production. Evaluation of abnormal voice quality associated with laryngeal diseases has attracted the interest of laryngologists, and the measurement of many different acoustic parameters has been proposed to quantitatively represent the degree of voice abnormality (Imaizumi, 1985).

Further, simultaneous recordings of vibratory patterns of the vocal folds and voice signals have led to a direct comparison between the temporal variation in vocal fold vibration and perturbation of voice, thus giving a physiological basis of abnormal voice production (Kiritani et al., 1993).

As for future research, it seems that basic studies on laryngeal structure and function are still needed. In particular, we still lack details of neural control in the human larynx, including: (1) efferent nerve cell distribution in the brainstem, and exact neural pathways to the laryngeal muscles from the central nervous system; (2) the control of the larynx by the autonomic nervous system; and (3) the cerebellar control of laryngeal timing, etc. In future research, these points need to be investigated.

In addition, further study is needed of the physical properties of the laryngeal framework, for example the network of blood vessels within the larynx, the surface microstructure and the physical properties of the laryngeal mucosa, and the vocal fold vibratory patterns in different laryngeal conditions under different emotional states. All of these should be suitable topics for future basic research.

REFERENCES

Abramson, A. S. (1977) Laryngeal timing in consonant distinction. *Phonetica*, 34, 295–303.

Benguerel, A.-P., Hirose, H., Sawashima, S., & Ushijima, T. (1978) Laryngeal control in French stop production: A fiberscopic, acoustic and electromyographic study. *Folia Phoniatrica*, 30, 175–98.

Bickley, C. (1991) Vocal-fold vibration in a computer model. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology* (pp. 37–46). San Diego: Singular Publication Group Inc.

Bless, D. & Abbs, J. H. (eds.) (1983) *Vocal Fold Physiology: Contemporary Research and Clinical Issues.* San Diego: College Hill Press.

Childers, D. G., Naik, J. M., Larar, J. N., Krishamurthy, A. K., & Moore, P. (1983) Electroglottography, speech, and ultra-high-speed cinematography. In I. R. Titze & R. C. Scherer (eds.), *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control* (pp. 202–20). Denver: The Denver Center for Performing Arts.

Cranen, B. (1991) Simultaneous modeling of EGG, PGG and glottal flow. In J. Gauffin & B. Hammerberg (eds.), *Vocal Fold Physiology* (pp. 57–64). San Diego: Singular Publication Group Inc.

Erickson, D., Liberman, M., & Niimi, S. (1977) The geniohyoid and the role of the strap muscle. *Haskins Laboratories Status Report on Speech Research*, SR-49, 97–102.

Fourcin, A. J. (1981) Laryngographic assessment of phonatory function. In C. L. Ludlow & M. O. Hart (eds.), *ASHA report 11: Proceedings of the Conference on the Assessment of Vocal Pathology* (pp. 116–27). Rockville, MD: The American Speech-Language-Hearing Association.

Frøkjær-Jensen, B. (1967) A photo-electric glottograph. *Annual Report of the Institute of Phonetics of University of Copenhagen*, 1, 5–19.

Frøkjær-Jensen, B. (1968) Comparison between a Fabre glottograph and a photo-electric glottograph. *Annual Report of the Institute of Phonetics of University of Copenhagen*, 3, 9–16.

Fujimura, O. & Sawashima, M. (1971) Consonant sequences and laryngeal control. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, University of Tokyo, 5, 1–13.

Harris, K. S. (1981) Electromyography as a technique for laryngeal investigation. In C. L. Ludlow & M. O. Hart (eds.), *ASHA report 11: Proceedings of the Conference on the Assessment of Vocal Pathology* (pp. 70–87). Rockville, MD: The American Speech-Language-Hearing Association.

Hirano, M. (1974) Morphological structure of the vocal cord as a vibrator and its variations. *Folia phoniatrica*, 26, 89–94.

Hirano, M. (1991) Phonosurgical anatomy of the larynx. In C. Ford & D. Bless (eds.), *Phonosurgery: Assessment and Surgical Management of Voice Disorders*. New York: Raven Press.

Hirano, M., Vennard, W., & Ohala, J. (1970) Regulation of register, pitch and intensity of voice: An electromyographic investigation of intrinsic laryngeal muscles. *Folia phoniatrica*, 22, 1–20.

Hirose, H. (1976) Posterior cricoarytenoid as a speech muscle. *Annals of Otology, Rhinology and Laryngology*, 85, 334–43.

Hirose, H. (1977) Laryngeal adjustment in consonant production. *Phonetica*, 34, 289–94.

Hirose, H. (1985) Laryngeal electromyography. In G. M. English (ed.), *Otolaryngology*, vol. 3 (pp. 1–14). St. Louis: Harper & Row.

Hirose, H. (1988) High-speed digital imaging of vocal fold vibration. *Acta Otolaryngol* (Stockholm), Supplement 458, 151–3.

Hirose, H. & Gay, T. (1972) The activity of the intrinsic laryngeal muscles in voicing control: An electro-myographic study. *Phonetica*, 25, 140–64.

Hirose, H., Kiritani, S., & Imagawa, S. (1988) High-speech digital image analysis of laryngeal behavior in running speech. In O. Fujimura (ed.), *Vocal Physiology: Voice Production, Mechanism and Functions* (pp. 335–45). New York: Raven Press.

Hirose, H. & Niimi, S. (1987) The relationship between glottal opening and the transglottal pressure differences during consonant production. In T. Baer, C. Sasaki, & K. Harris (eds.), *Laryngeal Function in Phonation and Respiration* (pp. 381–90). Boston, MA: College-Hill Press.

Hirose, H. & Ushijima, T. (1978) Laryngeal control for voicing distinction in Japanese consonant production. *Phonetica*, 35, 1–10.

Hirose, H., Yoshioka, H., & Niimi, S. (1979) A cross language study of laryngeal adjustment in consonant production. In H. Hollien & P. Hollien (eds.), *Current Issues in the Phonetic Sciences* (pp. 443–9). Amsterdam: John Benjamins.

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999) Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech*, 42, 401–11.

Imaizumi, S. (1985) Acoustic measures of pathological voice quality. *Journal of Phonetics*, 457–62.

Kahane, J. C. & Folkins, J. W. (1984) *Atlas of Speech and Hearing Anatomy*. Columbus, OH: Bell & Howell Co.

Kewly-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. *Haskins Laboratories Status Report on Speech Research*, SR-33, 173–84.

Kiritani, S., Hirose, H., & Imagawa, H. (1993) High-speed digital image analysis of vocal cord vibration in diplophonia. *Speech Communication*, 13, 23–32.

Kiritani, S., Imagawa, H., & Hirose, H. (1995) Vocal cord vibration in the production of consonants: Observation by means of high-speed digital imaging using a fiberscope. *Journal of the Acoustical Society of Japan*, (E) 17, 1–8.

Kitamura, T., Honda, K., & Takemoto, H. (2005) Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoustical Science and Technology*, 26, 16–26.

Leden, H. von & Moore, P. (1961) The mechanism of the cricoarytenoid joint. *Archives of Otolaryngology*, 73, 541–50.

Löfqvist, A. & Yoshioka, H. (1980) Laryngeal activity in Swedish obstruent clusters. *Journal of the Acoustical Society of America*, 63, 792–801.

Negus, V. E. (1949) *The Comparative Anatomy and Physiology of the Larynx*. London: Heinemann.

Poletto C. J., Verdun L. P., Strominger R., & Ludlow, C. L. (2004) Correspondence between laryngeal vocal fold movement and muscle activity during speech and nonspeech gestures. *Journal of Applied Physiology*, 97, 858–66.

Sawashima, M. (1968) Movements of the larynx in articulation of Japanese consonants. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 2, 11–20.

Sawashima, M. (1977) Fiberoptic observation of the larynx and other speech organs. In M. Sawashima & F. S. Cooper (eds.), *Dynamic Aspects of Speech Production* (pp. 31–46). Tokyo: University of Tokyo Press.

Sawashima, M. & Hirose, H. (1968) New laryngoscopy technique by use of fiber optics, *Journal of the Acoustical Society of America*, 43, 168.

Sawashima, M. & Hirose, H. (1983) Laryngeal gestures in speech production. In P. F. MacNeilage (ed.), *The Production of Speech* (pp. 11–38). New York: Springer.

Sonesson, B. (1960) On the anatomy and vibratory pattern of the human vocal folds. *Acta Otolaryngologica*, Supplement 156, 1–58.

Tsunoda, K., Niimi, S., & Hirose, H. (1994) The roles of the posterior cricoarytenoid and thyropharyngeus muscles in whispering speech and human evolution. *Folia Phoniatrica et Logopaedica*, 46, 139–51.

Tsunoda, K., Ohta, Y., Soda, Y., Niimi, S., & Hirose, H. (1997) Laryngeal adjustment in whispering magnetic resonance imaging study. *Annals of Otology, Rhinology and Laryngology*, 106, 41–3.

Part II Biological Perspectives

5 Organic Variation of the Vocal Apparatus

JANET MACKENZIE BECK

1 Introduction

1.1 The relevance of organic variation for phonetic science

The theoretical study of phonetics has been based on many assumptions. One of the chief among these is the notion that all speakers use speech production systems which can be treated as if they were essentially equivalent in terms of their anatomical geometry. This assumption is helpful when the aim is to identify the common strands of phonetic performance which allow a similar phonetic analysis to be made for a range of speakers producing the "same" linguistic content. The focus on commonality does, however, mask much of the intricacy and subtlety of individual phonetic performance. As phonetic science has become more sophisticated in its investigative techniques, allowing us to investigate speech output in finer detail, so individual differences in speech performance have become more apparent, and the motivation to examine the underlying causes of such differences has grown.

It may be useful to draw an explicit distinction between two major sources of variation in speech performance, which we will call phonetic and organic factors, following Laver (1980, p. 9). Phonetic variation results from differences in the way in which an individual uses his or her vocal apparatus, whilst organically based variation depends on individual differences in shape and proportion of the vocal organs. The word "organic" will be used here to describe any factors which are to do with anatomical structure or morphology, and with the constraints which that structure imposes on the potential for physiological action. A study of organic features in this sense may be seen as analogous to the study of architecture, being concerned principally with the mechanical properties of the materials, or tissues, which form the vocal apparatus, and the way in which they are arranged. Organic variation, therefore, may encompass any anatomical features for which individual differences in size, shape, or mechanical properties may be observed. The term "organic" could also be said to include the physical features underlying neurological control, but these will not be considered here. This is not to suggest that individual differences in physiological activity and neurological control of speech are unimportant, but it would not be realistic to include all these aspects within the scope of a single chapter.

The most casual inspection of the general population will show how misleading any assumption of vocal tract equivalence must be. Even among the genetically related members of a family, there will almost always be differences in the size and shape of the dental arches, the palatal contour, the relationship of the upper and lower jaws, and the size of the larynx. All these factors have implications for an individual's speech production, and to ignore them is to ignore a rich store of information which may help to explain at least some of the observed individual differences in speech production. In 1991, Laver (1991, p. 211) observed that (within phonetics) "Inter-speaker differences of anatomy within the normal distribution have been largely ignored." Since then, it is true, there has been an increase in the volume of published research linking phonetic output to specific types of organic variation, fueled largely by data generated by improved articulatory measurement techniques such as electropalatography, electromagnetic articulography, ultrasound, and Magnetic Resonance Imaging (MRI). Examples include electropalatographic studies of dental anomalies (Wakumoto et al., 1996; Cayley et al., 2000) and MRI investigations of tongue movement following partial glossectomy (Mády et al., 2001; Murano et al., 2008). Nonetheless, systematic research into the organic bases of phonetic variation is still somewhat scarce.

The phonetic effects of relatively minor organic differences between individuals may be quite trivial, and observable only through careful articulatory or acoustic measurement. For example, individual variations in dentition may cause subtle differences in anterior tongue placement and fricative airflow in front oral fricatives, but are not likely to have much impact on perceived speech quality. Much more substantial phonetic effects result from changes in overall size and shape of the vocal organs during normal growth from childhood to maturity, or from gross anomalies of the vocal tract such as are found in cleft palate or oral cancer. These latter distortions may impose serious limitations on potential phonetic performance. In addition, we all, at one time or another, experience changes in speech output associated with more transient changes within the vocal apparatus. Day-to-day fluctuations in speech output may result directly from such things as nasal obstruction during a common cold, a broken tooth, or a mouth ulcer.

The aim of this chapter is to begin to explore organic variation as it affects the vocal apparatus, by bringing together some of the phonetically relevant information concerning human growth and variation. For general information on the anatomy of the vocal apparatus the reader is referred to standard references such as Hardcastle (1976), Dickson and Maue Dickson (1982), Kahane (1988), and Seikel et al. (2000). The chapter aims to provide a broad overview of some of the types and sources of organic variation which affect speech output, in a form which should be easily accessible by the phonetician with a basic knowledge of the vocal apparatus. It is not intended to be exhaustive, but seeks to alert students of phonetics to some of the potential organic causes for inter- and intra-speaker variation in phonetic performance. In the interests of conciseness, only a selection of illustrative references is presented. A fuller review of some of the older relevant literature may be found in Beck (1988).

1.2 Sources of individual variation

The sources of individual differences in vocal apparatus structure fall into three main categories:

- normal life-cycle changes which affect each individual as they grow, develop and age;
- genetic and environmental factors which differentiate between individuals;
- the organic consequences of trauma or disease.

1.2.1 Intra-individual variation: Life-cycle changes The vocal apparatus, in common with every other part of the body, undergoes a complex process of change throughout the life span. Age-related changes in the vocal apparatus can be seen as falling into three main phases. During the first phase, which corresponds to the period between birth and puberty, major changes in the vocal apparatus accompany general patterns of growth and development. There are no very salient differences between the sexes in terms of morphology or size of the vocal apparatus during this first phase.

The second phase, from puberty to maturity, is characterized by a major growth spurt associated with the onset of puberty and by the fact that male and female patterns of growth and development are typically different. It is during this phase that the major differentiation between the male and female vocal apparatus emerges. Despite some discrepancies in growth patterns between different parts of the speech production system, the overall growth of the vocal apparatus during these two phases generally reflects the body growth curves for males and females (see Figure 5.1a and b).

During the final phase, from maturity to senescence, growth process activity is limited to maintenance and repair, and the changes which occur are generally the result of the decreasing efficiency of these maintenance and repair processes, leading to degenerative change.

1.2.2 Inter-individual variation: Genetic and environmental conditioning While some differences between individuals may be due to sampling at different points in the life cycle, this is obviously not the whole story. Given any group of people of the same age and gender, there will still be marked differences in vocal tract morphology. There is considerable variation in the precise coordination and timing of the changes which occur during development and aging, as well as in the genetic template for each person's target adult form. Developmental patterns are influenced by both endogenous and environmental factors, but the ways in which

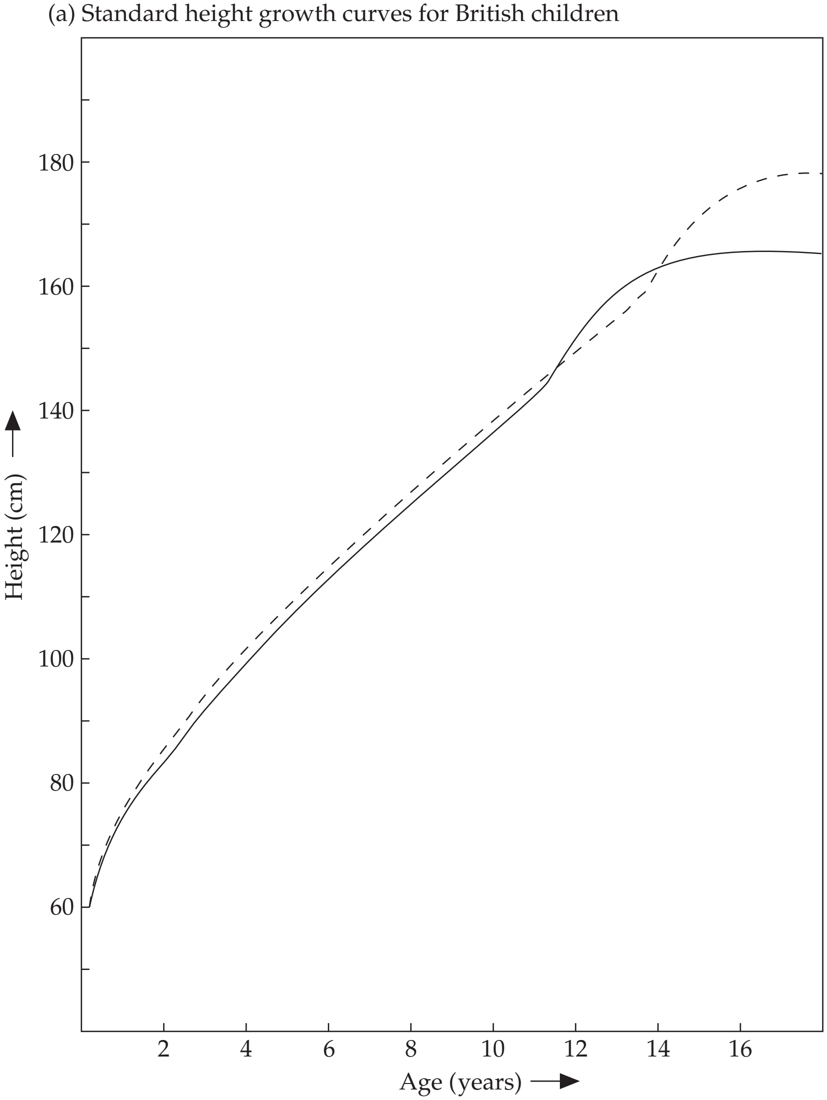


Figure 5.1 Typical overall growth patterns ((a) and (b) both adapted from Tanner, 1978, pp. 169, 170, 177, 178). Female growth patterns are indicated by solid lines, and male growth patterns by the broken lines.

these factors interact in order to coordinate growth and development are still only partially understood.

1.2.3 Variation arising from trauma or disease In addition to the organic variation arising as a result of normal development or degeneration of the vocal

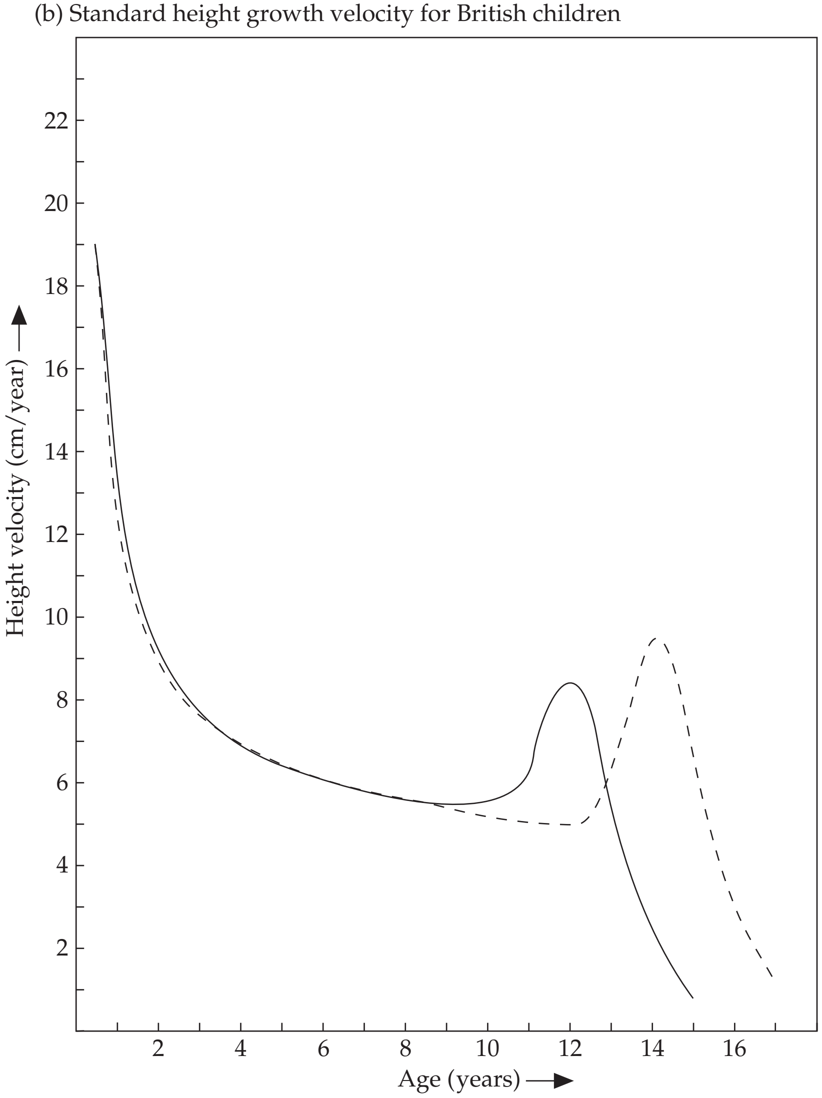


Figure 5.1 (*continued*)

apparatus, there may also be changes within the vocal apparatus which result from traumatic injury or disease. Although these may be defined as "abnormal," they are nevertheless common enough that a high proportion of the population will suffer from some trauma- or disease-related change in their vocal organs at some time in their lives, even if such changes are transient in nature. It therefore seems appropriate to consider such changes as being relevant to general phonetic science, and not solely within the domain of speech and language pathology.

These three sources of organic variation will be illustrated in more detail starting with a discussion of the processes of growth, development, maturation, and aging which affect all of us as we progress through life.

2 Life-Cycle Changes in the Vocal Apparatus

This section will focus on each major area of the vocal apparatus in turn, progressing from lungs and thorax to the larynx, and thence upwards through the resonating cavities. The principal organic changes occurring within each phase of the life cycle will then be summarized.

Discussion of age-related changes within each part of the vocal apparatus will include a description of skeletal aspects, followed by a description of soft tissue changes. Although the extraordinary plasticity of bone growth means that in the longer term it is a mistake to think of the skeleton as rigid or immutable, it is fair to say that at any given point in the life cycle the skeleton does behave as a rigid framework supporting the overlying soft tissues. Soft tissues are subject to constant observable distortion during normal movement of the body, and are prone to significant and short-term alterations in size and consistency in response to infection, hormones, or physiological state, whereas bones are not. When an individual is studied over a short time period, the overall shape and size of that person's vocal apparatus will thus be determined principally by his or her skeletal structure. Each section will therefore begin by considering the growth patterns of the underlying skeletal structures.

2.1 The respiratory system: The lungs and thorax

2.1.1 Skeletal framework: Thoracic skeleton At birth, the whole of the thoracic skeleton and the shoulder girdle is relatively high, as the small size of the pelvis causes the abdominal contents to be compressed upwards towards the diaphragm (Sinclair, 1978, p. 119). Rapid pelvic development during the first two or three years of life allows the abdominal contents, and hence the thorax, to drop. The thoracic skeleton grows to accommodate the lungs, and follows a similar curve (Altman & Dittner, 1962, p. 334). The circumference of the thorax seems to be slightly larger in males than in females in childhood, and this difference increases dramatically at puberty. In adulthood the sternum is shorter in females, and in a slightly higher position relative to the vertebral column. Females also have rather more mobility of the upper ribs, allowing greater expansion of the upper part of the thorax (Davies & Davies, 1962, p. 285). This is assumed to be an evolutionary adaptation for pregnancy, when the lower thorax and diaphragm are constricted by the uterus.

The angle of the ribs has important implications for the efficiency of respiration. In the adult, the ribs are angled downwards, and chest diameter is increased by pulling the ribs to a more horizontal position. During the first two years of life the ribs lie more horizontally (Sinclair, 1978, p. 121; Kahane, 1988), so that raising the ribs has little effect on chest volume. The infant is thus much more dependent on diaphragmatic breathing. Thoracic wall movement increases progressively up to the age of 7, by which time the angle of the ribs is similar to that in adults (Kahane, 1988). In old age the state of the ribs again impedes efficient respiration, as the rib cartilages become calcified and thus lose their ability to twist and allow proper elevation of the ribs during inspiration.

2.1.2 Soft tissues: Lungs, bronchioles, bronchi, and trachea At birth, the lungs are very small, both in mass and volume. During the first few weeks of life they expand greatly, and by the end of the first year the lungs have trebled in weight and increased sixfold in volume (Sinclair, 1978, p. 89). After the first rapid period of growth the lungs follow the general growth curve (Simon et al., 1972). The internal structure of the lungs shows considerable change following birth. Most of the alveoli of the lung are formed after birth, and the number of alveoli increases until some time between 8 years and puberty (Emery, 1979; Kahane, 1988). The density of elastic fibres in the terminal airways increases concomitantly, allowing the lungs to recoil more easily during expiration.

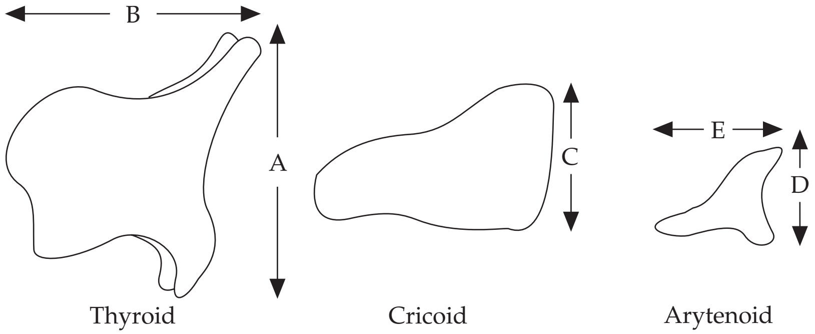
About 50 percent of the solid matter of the lungs is made up of collagen (Bouhuys, 1977), which probably functions to prevent over-extension of the lungs. Changes in the quality of the collagen network occur in old age, as the collagen molecules form cross links and become less flexible. The lung structure becomes less mobile, progressively impairing respiratory function. This, in association with increasing rigidity of the thoracic skeleton, results in a reduction of vital capacity from a range of approximately 3.5 to 5.9 liters in young adult males to a range of 2.4 to 4.7 liters after the age of 60 years (Sinclair, 1978, p. 223).

2.2 *The phonatory system: The larynx*

2.2.1 General features Laryngeal growth during childhood has been relatively little studied. The position of the larynx in the newborn is very high relative to other structures of the vocal tract, and the epiglottis makes contact with the soft palate. This contact is lost through progressive lowering of the epiglottis and larynx during the first year of life. At the age of 6 months the epiglottis and soft palate are well separated, although they make contact during swallowing, and by 12 to 18 months the contact even during swallowing is inconsistent.

2.2.2 Skeletal framework: Laryngeal cartilages Dickson and Maue-Dickson (1982, p. 176) report that growth of the laryngeal cartilages is more or less linearly related to growth in height in both sexes, and that a rapid increase in size of the male cartilages at puberty results in significant adult sex differences. Maue (1970) and Maue and Dickson (1971), both cited in Dickson and Maue-Dickson (1982, pp. 142, 148), give some measurements for male and female laryngeal cartilages which are summarized in Figure 5.2. It is clear that significant growth during

(a) Average dimensions of adult laryngeal cartilages (based on data from Dickson & Maue-Dickson, 1982, pp. 142–8)



CARTILAGE	DIMENSION	FEMALE	MALE
Thyroid	height (A) anterior-posterior (B) weight	38 mm 29 mm 4 gm	44 mm 37 mm 8 gm
Cricoid	height (C) weight	19 mm 2.89 gm	25 mm 5.8 gm
Arytenoids	height (D) anterior-posterior (E) weight	13 mm 10 mm 0.20 gm	18 mm 14 mm 0.39 gm

(b) Gender differences in contour of the thyroid cartilage
(i) Superior view (adapted from Dickson & Maue-Dickson, 1982, p. 142)
(ii) Lateral view (adapted from Kahane 1988, p. 13)

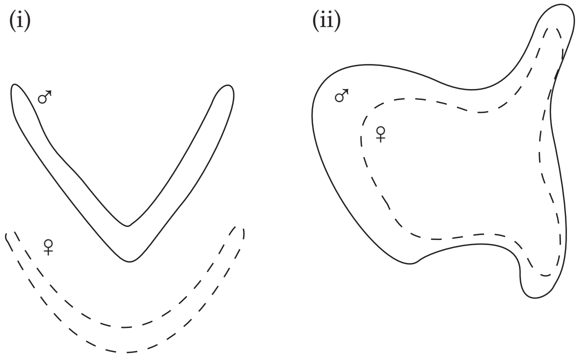


Figure 5.2 Summary of adult gender differences in laryngeal cartilages.

childhood is followed by marked sexual differentiation of the cartilaginous laryngeal skeleton at puberty (Kahane, 1988), following the general growth curve.

During aging, laryngeal cartilages are subject to calcification, with consequent changes in elasticity of the cartilages and to the mechanical characteristics of the insertion zones of the vocal ligaments (Paulsen et al., 2000). The age of onset of calcification varies considerably. It may begin in men in their thirties, but the thyroid cartilage may still be unaffected in some 70-year-olds. In women, ossification generally begins later and is less extensive (Pantoja, 1968; Kahane, 1987; Greene & Mathieson, 1989).

2.2.3 Soft tissues: The vocal folds The whole larynx is extremely small at birth, but reported overall vocal fold length measurements are rather discrepant, varying between 2.5 and 9 mm (Negus, 1949; Terracol et al., 1956; Ballenger, 1969, cited in Aronson, 1980, p. 44; Hirano et al., 1983). Growth seems to be most rapid in the first five years, and again during the pubertal growth spurt, especially in males.

There seems to be less disagreement about average adult vocal fold length, which is usually reported to be between 23 and 25 mm in males, and about 17 mm in females (Romanes, 1978; Greene & Mattheson, 1989). Hirano et al. (1983) report slightly smaller adult measurements, suggesting a total vocal fold length of 17–21 mm in males, and 11–15 mm in females. This is based on data for Japanese subjects, but it may be that there are geographical differences in laryngeal dimensions which reflect genetic variation between populations.

The relative proportions of the ligamental and cartilaginous parts of the vocal folds are usually reported to be broadly similar in both sexes, with the ligamental part constituting about two thirds of the total vocal fold length in adults. Hirano et al. (1983) however, have shown that the ratio of cartilaginous to ligamental portions of the vocal fold changes throughout childhood, and that there is a slight gender difference in the adult ratio. In newborns, the cartilaginous portion of the vocal fold constitutes only slightly less than half the total length of the vocal fold, but disproportionate growth of the ligamental portion results in a relative as well as an absolute increase in size of the ligamental vocal fold. This is slightly more marked in boys, so that in adult males the ligamental portion of the vocal fold constitutes rather more than a third of the vocal fold length.

The structure of the vocal fold at birth is very immature. The fibers of the vocalis muscle are poorly developed, and von Leden (1961) suggests that neuromuscular maturation of the larynx is not complete before 3 years. The tissue layers which make up the vocal ligament are also poorly differentiated, and adult tissue-layer relationships are not seen until after puberty.

In newborn infants there seems to be no clearly differentiated vocal ligament, and the entire lamina propria seems to be rather uniform and pliable. The only areas of increased fiber density are at the ends of the ligamental portion of the vocal folds, and probably represent precursors of the maculae flavae. By 4 years of age an immature vocal ligament is present, but the differentiation between the elastic intermediate layer and the collagenous deep layer of the lamina propria does not begin until between 6 and 12 years. By 15 years of age a clear differentiation

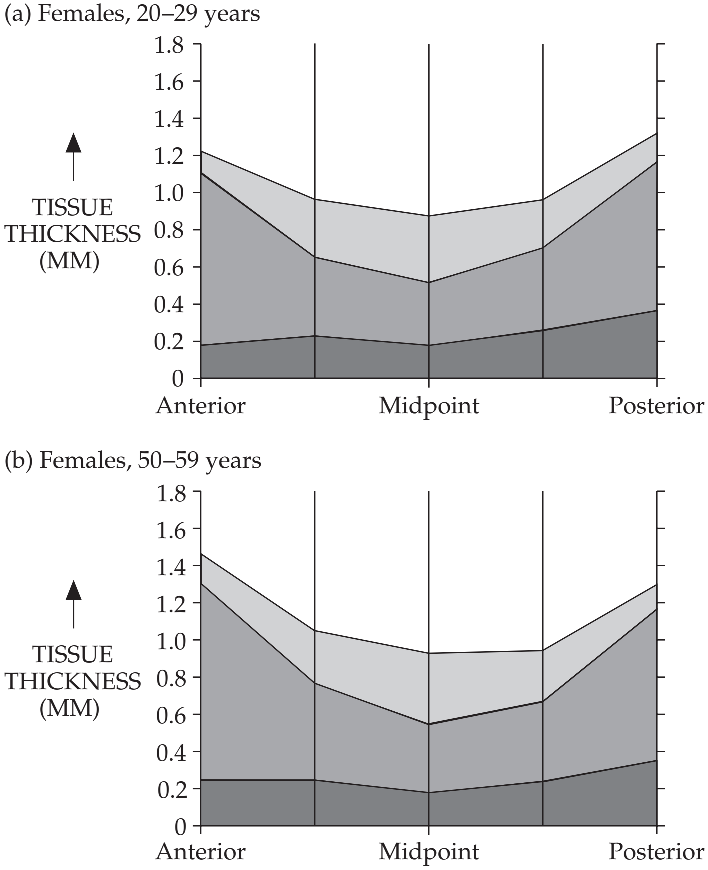


Figure 5.3 Age-related changes in the tissue layer structure of the vocal folds. (Based on data from Hirano et al., 1982, p. 274)

is typically observed. Full maturation may not occur before 20 years of age; before this the vocal ligament is thinner than in the adult, with a looser fiber arrangement. The epithelium shows no significant changes during development (Hirano et al., 1981; Hirano et al., 1982).

After reaching maturity there may be continuing changes in tissue thickness and consistency (see Figure 5.3a, b, and c). Edema of the outer connective tissue cover of the vocal folds combines with a decrease in elastic fibers and an increase and distortion of collagen fiber content in the deeper layers to alter the mechanical properties of the vocal folds (Honjo & Isshiki, 1980; Hirano et al., 1982; Kahane, 1983; Paulsen et al., 2000).

Yellowish or greyish discoloration of the vocal folds seems to occur quite often in older age groups (Honjo & Isshiki, 1980; Mueller et al., 1985), and may indicate a degree of fatty degeneration or keratinization of the epithelium. These localized changes in the mechanical properties of the vocal folds may cause dysperiodic

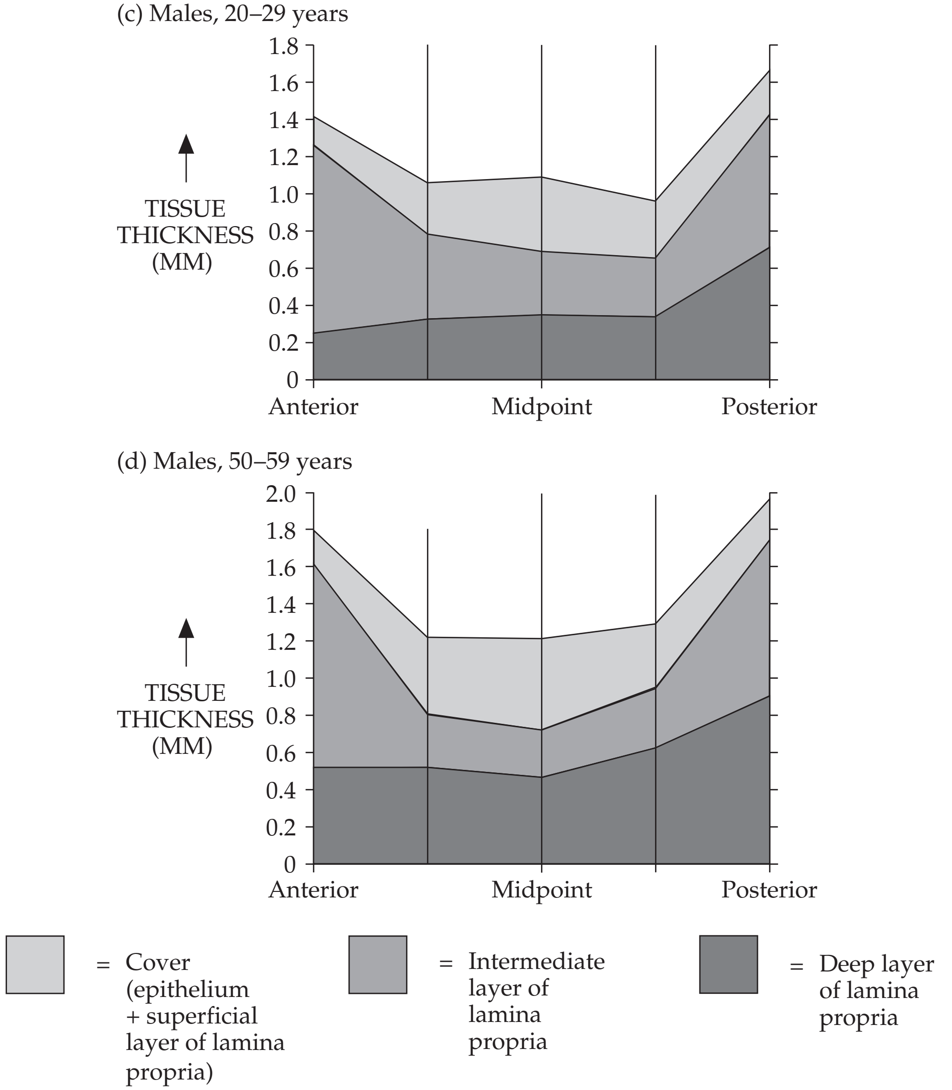


Figure 5.3 (*continued*)

vibration, which would be perceived as harshness. Any significant change in overall mass or stiffness of the vocal folds may also affect fundamental frequency (see Figure 5.16).

Atrophy of the laryngeal musculature, especially of the vocalis muscle which forms the bulk of the body of the vocal fold, is also a commonly reported feature of the aging larynx, which may be more marked in males (Honjo & Isshiki, 1980; Mueller et al., 1985). The decrease in muscle mass and strength may prevent complete adduction of the folds, with consequent air wastage resulting in a whispery phonation and lower intensity. Decreased vocal fold mass may also be associated with an increase in fundamental frequency.

The mechanical structure of the conus elasticus supporting the vocal folds seems also to be subject to degenerative change in old age, especially at the point of union with the vocalis muscle, with males once again being more susceptible (Kahane, 1987).

Some gender differences in the pattern of age-related changes within the larynx have been reported, but findings are somewhat inconsistent and it is possible that varying lifestyle differences between men and women in different cultures may influence the types of degenerative change observed. In general, it appears that men are more prone to vocal fold thinning and tissue changes which reduce vocal fold and cartilage elasticity, while women may be more prone to vocal fold edema (Honjo & Isshiki, 1980; Linville, 2000).

2.3 Resonating cavities: Pharynx, oral cavity, and nasal cavity

2.3.1 Skeletal determinants of the resonating cavities The most important of these is probably the skull, together with the cartilages and bones of the facial skeleton. The skull is usually described as consisting of two parts: the cranium, which encloses and protects the brain, and the facial skeleton. Structurally, these parts form a cohesive whole, but functionally they are rather different, and this difference is reflected in their disproportionate growth patterns. At birth, the cranium is substantially larger than the size of the face, and its relative size increases still further during the first six to twelve months of life as it grows more rapidly than the rest of the skull. Thereafter, facial growth is greater, and continues longer, so that in an adult the cranium is only two to three times the size of the face (Watson & Lowrey, 1967; see Figure 5.4). Growth of the base of the skull, which provides points of articulation with the vertebral column and allows passage of the respiratory and digestive tracts and the spinal cord, is allied with the facial skeleton in terms of its growth behavior.

2.3.2 The cranium Growth of the cranium, as might be expected, reflects quite accurately the growth of the brain, being most rapid during the first one or two years of life, and virtually complete by 10–12 years (Watson & Lowrey, 1967; Sinclair, 1978; Tanner, 1978). The cranium is significantly smaller in females, and the frontal bone may be more prominent (Wei, 1970; Ingerslev & Solow, 1975).

2.3.3 The facial skeleton The facial skeleton imposes much more direct limits on the morphology of the resonating cavities of the vocal tract, and is therefore of more immediate relevance to phonetics. The main constituent parts and landmarks of the facial skeleton are shown in Figure 5.5.

There is a large and often controversial literature concerning development of the facial skeleton. Disagreement about normal patterns of growth arise partly

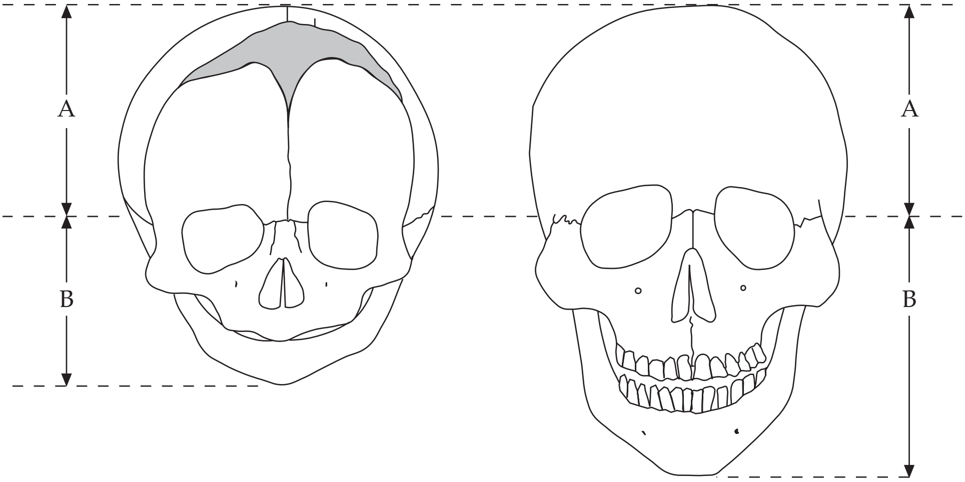
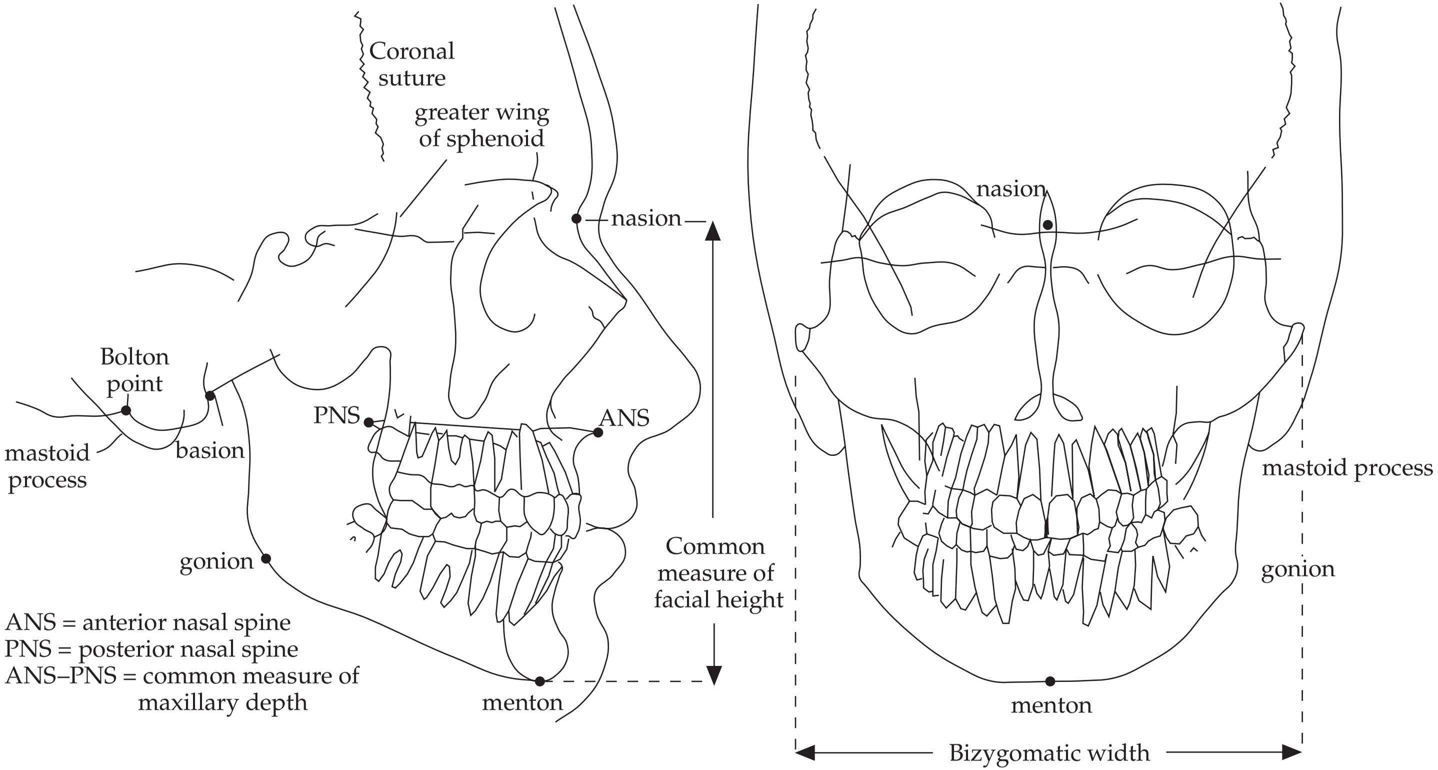


Figure 5.4 Changing proportions of the skull from birth to maturity. Newborn and adult skulls are drawn so that the cranial height is the same, showing the proportionately larger face in the adult. (Adapted from Sinclair, 1978, p. 94)

from the high degree of real variability in facial morphology and growth, and partly from the variety of cephalometric techniques used. A further problem is that descriptions may be biased towards an ideal view of growth, since many studies use only children of "good dental health" (e.g., Walker & Kowalski, 1972, p. 111) or normal occlusal relationships (Shah et al., 1980).

Variability in facial structure obviously has a large genetic component, as evidenced by the observation that different ethnic groups show very different facial characteristics, but facial growth patterns also display a high degree of plasticity, responding quite readily to environmental factors. A certain amount of flexibility in the growth patterns of the various parts of the facial skeleton is presumably an adaptive response to the need for very complex coordination of growth of the many bones and cartilages which make up the facial skeleton. The growth of each part must be carefully timed so as to maintain functional harmony of the overall facial structure, and it may be that the best way of achieving this harmony is for each growth area to be especially sensitive to its skeletal and soft tissue environment. The problem of coordinating growth is not, of course, unique to the face, but the complexity of the skeleton in this area makes it particularly crucial. The observation that facial characteristics are highly prone to disturbance by a wide variety of genetic and environmental abnormalities (Martin, 1961), ranging from Down syndrome to fetal alcohol syndrome, is indicative of the level of sensitivity to growth disturbance displayed by the facial skeleton. Some examples of disturbed growth in this area will be described later in this chapter.

The facial skeleton and the cranial base follow the general body-growth curve much more closely than does the cranium. In early childhood, growth is closely



ANS = anterior nasal spine
PNS = posterior nasal spine
ANS–PNS = common measure of maxillary depth

Figure 5.5 Landmarks of the facial skeleton (adapted from several sources).

related to development of the muscles of mastication, the tongue, and the dentition. There is a pronounced adolescent growth spurt in most measurements (Hunter, 1966; Dermaut & O'Reilly, 1978; Shah et al., 1980), but the precise timing of the growth spurt may depend on the measurements used, the sex of the subjects, and their genetic background. In females, facial growth is usually almost completed in the late teens, but facial growth in males may continue into the mid-twenties (Hunter, 1966). The growth of the mandible seems to show the closest correlation with overall body-growth curves (Hunter, 1966). Generally, growth in facial width is completed earlier than growth in the anteroposterior dimension, and vertical growth of the face may continue into the third decade of life.

The various component sections of the facial skeleton will be considered separately, although vocal tract configuration depends as much on the relationship between these sections as on the shape or absolute size of each.

Palate and maxilla Growth in size of the maxilla (upper jaw) and palate is quite complex. Watson and Lowrey (1967) differentiate three anatomical regions of the nasomaxillary complex, which all show different growth patterns. During the first year of life there is generalized growth of maxilla and palate, but after this period, growth becomes more localized.

1 Length of the anterior portion of the palate and maxilla becomes fixed in early infancy, and palatal width becomes fixed at 4–5 years of age. Thereafter, alveolar width is increased by apposition of bone at the external surface of the alveolar bone.
2 Bizygomatic width (see Figure 5.5) has a very different pattern of growth, increasing at a smoothly and steadily diminishing rate until adulthood. Growth in this dimension is particularly pronounced in males.
3 Maxillary width keeps pace with palatal and bizygomatic widths. Height and length of the maxilla increase concurrently, as growth proceeds in a forward and downward direction.

Figures 5.6a, b and c summarize palatal dimensions for American Caucasians (Shapiro et al., 1963; Redman et al., 1966). Unfortunately these findings were not related to measurements of any other part of the craniofacial skeleton, nor to overall body growth. It can be seen that there is significant sexual differentiation, and this accords with data for Danish subjects (Ingerslev & Solow, 1975). Figure 5.6d shows the changing proportions of the palatal vault which result from these growth patterns. There seems to be considerable variability in the timing and extent of the maxillary growth spurt, at least for females (O'Reilly, 1979).

The maxilla shows some degenerative changes in old age, especially in the area of tooth insertion. As teeth are lost, the requirement for bone thickness in the tooth socket area is reduced, and bone tends to be lost.

The mandible Growth of the mandible (lower jaw) seems to be highly sensitive to a variety of factors. It seems to respond more to growth hormone than most

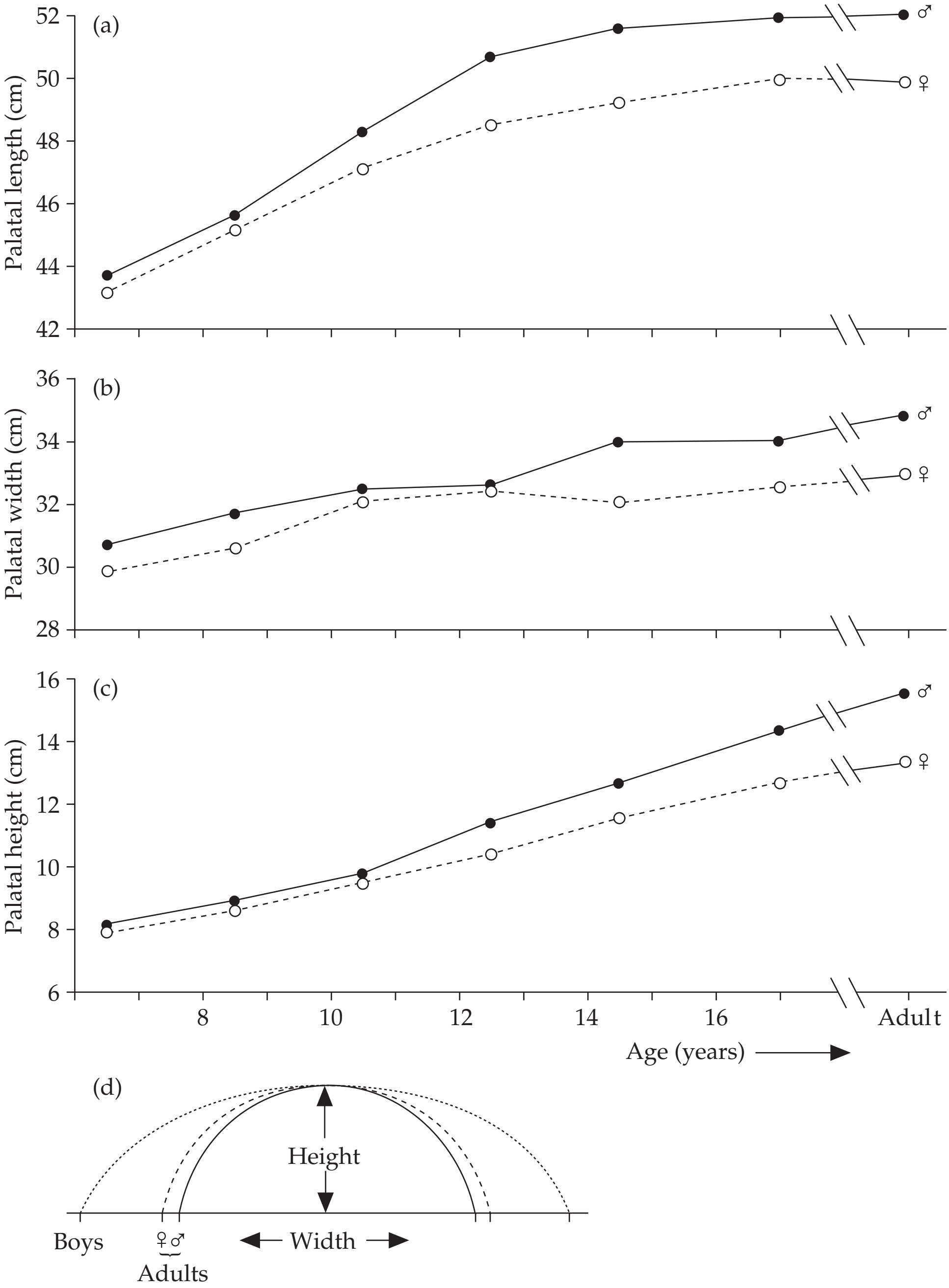
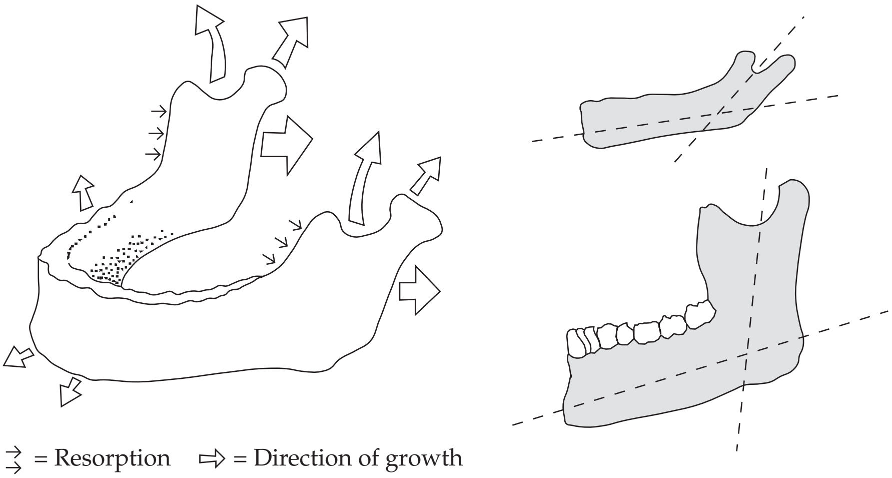


Figure 5.6 Developmental changes in palatal dimensions and proportions. (Based on data from Redman et al., 1966)

(a), (b), and (c): graphic representation of palatal dimensions vs. age; (d) relationship between palatal height and width for 6–7-year-old boys, women, and men, normalized for height.



(a) *Schematic diagram of the main axes of mandibular growth and remodeling (adapted from Enlow & Harris, 1964, p. 50 and Sinclair 1978, p. 58)*

(b) *Resultant change in mandibular angle from infancy to adulthood (adapted from Sinclair, 1978, p. 55)*

Figure 5.7 Patterns of mandibular growth.

other bones (Bevis et al., 1977), and may also be more responsive to testosterone. It is also very sensitive to the muscular forces imposed upon it (Watson & Lowrey, 1967). Mandibular growth seems to be subordinate to maxillary growth, following growth of the maxilla in such a way as to produce adequate occlusion.

The mechanism of mandibular growth is complex and very variable (Enlow & Harris, 1964; Sinclair, 1978, p. 77; Krarup et al., 2005). Increase in length follows the general body-growth curve quite closely, with a greater and longer-lasting growth spurt in males than in females, so that sexual dimorphism in mandibular length becomes quite marked by adulthood (Hunter, 1966; Walker & Kowalski, 1972; Ingerslev & Solow, 1975). Figure 5.7, adapted from Enlow and Harris (1964) and Sinclair (1978, p. 77), shows the main areas of mandibular growth and remodeling. Growth results primarily in a length increase, although width also increases to allow proper articulation with the skull. During the prepubertal phase, there is considerable appositional growth at the head of the mandible. Bone growth behind the ramus, accompanied by bone resorption at the front of the ramus, gradually increases the space available for the dentition. The angle between the ramus and the body of the mandible is gradually reduced from about 140 degrees in infancy to 120 degrees in adulthood. The greatest contribution to overall facial growth at the time of puberty is made by the mandible. During this period, most growth continues in the ramus, but there are also marked increases in the length of the body of the mandible and the vertical distance between the chin and the incisors.

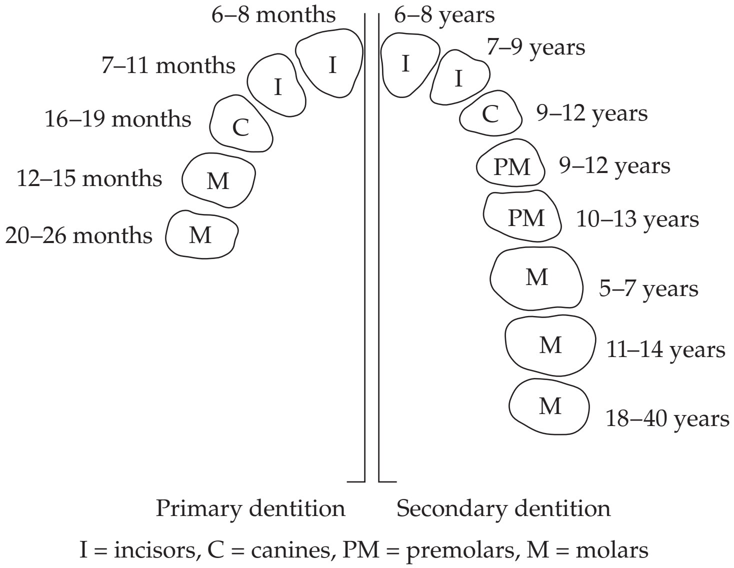


Figure 5.8 Typical ages of tooth eruption.

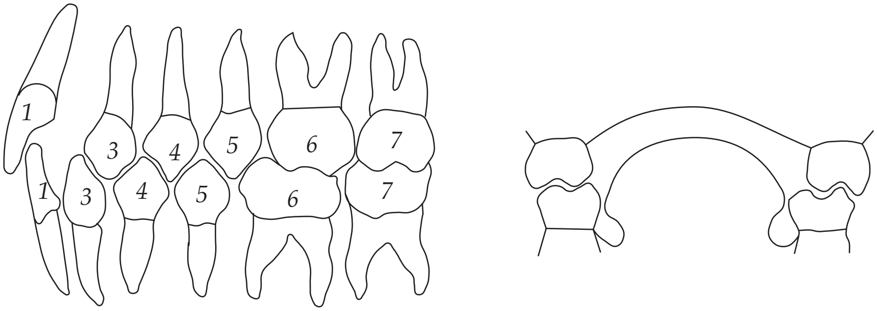
As the mandible grows, the bone remodeling allows teeth to move forwards to create space for the eruption of the molar teeth, and the incisor teeth gradually incline forwards (Sinclair, 1978, p. 78).

As with the maxilla, loss of teeth is associated with bone resorption in the alveolar margin, so that the angle of the mandible becomes more obtuse, as in infancy, and may reach about 140 degrees (Sinclair, 1978, p. 218).

Dentition The first primary teeth usually appear at about 6 months of age. The age of eruption is variable, but usually all have emerged by the age of 2½ years. The eruption of the permanent teeth is also very variable, but usually begins between 5½ and 6 years, and is complete, with the exception of the third molars, at around 12 years. The third molars, or wisdom teeth, do not normally erupt until between 18 and 21 years. Typical ages of tooth eruption are shown in Figure 5.8. The age of eruption of the permanent dentition is slightly earlier in girls, in line with the general trend towards earlier maturity in girls.

Tooth loss through disease is a common feature of old age. The gums begin to recede from the crowns of the teeth in early adulthood, and since the enamel covering the crown of the tooth cannot regenerate, the enamel covering becomes gradually more worn from contact with hard foodstuffs and decay resulting from plaque and infection.

Nasal cavity There is little data available on growth and development of the nasal cavity, but the poor development of the nasal bone at birth, and the marked enlargement of the nasal bone at puberty, together with other changes in



(a) Lateral view
(adapted from Foster, 1990)

(b) Cross section between the molar teeth

Figure 5.9 Schematic representation of normal occlusion. For occlusion to be classified as normal the relationship of upper and lower teeth must be as shown here, and there should be no missing or misplaced teeth.

proportions of the facial skeleton, point to major changes in the internal structure of the nose between birth and maturity. The oral–nasal port will be influenced by the lumen of the pharynx, the size and carriage of the tongue, and the mass of lymphoid tissue which is present at any given stage in development. All these factors may have major consequences for the balance of oral and nasal resonance, since the relative sizes of the posterior entrances to the oral and nasal cavities are thought to be important determinants of nasal resonance (Laver, 1980), but it is unfortunately hard to evaluate their precise effects.

2.3.4 Jaw relationships It was mentioned earlier that growth of the mandible tends to accommodate itself to maxillary growth so that the upper and lower teeth meet (or occlude) in the correct relationship. This accommodation process is not infallible, however, and minor problems of occlusion are very common. Whilst some of these may be transient results of uncoordinated growth between the maxilla and mandible during childhood which are corrected by later stages of mandibular growth, a significant proportion persist into adulthood.

In "normal" occlusion of the teeth, the back surfaces of the maxillary incisor teeth are in contact with the front surfaces of the mandibular incisor teeth. Each lower tooth contacts the corresponding upper tooth, but is relatively slightly further forward so that it also overlaps the adjacent upper tooth. The only exceptions to this are the lower central incisors, which occlude only with the upper central incisors. The vertical overlap (overbite) of mandibular and maxillary incisors is as shown in Figure 5.9, with an overlap of between one third and two thirds (Foster, 1990). The horizontal gap (overjet) between the point of the upper incisors and the nearest point of the lower incisors is about 3 mm. Although this is an accepted description of "normal" occlusion, it might be better described as "ideal," as such a high proportion of the population deviates from this ideal, having some degree of malocclusion (Foster, 1990). This will be discussed further in the section concerned with inter-individual variations.

2.3.5 Nasopharynx The bony nasopharynx appears to expand its volume primarily through vertical growth, and there are some indications that this vertical growth is influenced by any soft tissue obstruction of the airway that may occur (Tourne, 1991).

2.3.6 Soft tissues: Soft palate, lymphoid tissue, tongue The soft tissue structures which are most significant in terms of their effect on resonating cavity volume are the walls of the pharynx, the soft palate and related muscular arches, the lymphoid masses which form the adenoids and tonsils, the tongue, and the lips.

Pharyngeal walls Soft tissue development of the pharyngeal walls seems to have been little studied, but the muscular walls of the pharynx can be assumed to expand quite rapidly as an adaptation to the skeletal and postural changes which occur during infancy. As the head is gradually held in a more upright position with greater extension of the neck, and the larynx adopts a lower position in the neck, pharyngeal volume increases dramatically. At puberty there is another period of pharyngeal enlargement, which is more marked in males, as the larynx descends further. In old age, the general tendency for muscles to atrophy and for mucosal linings to degenerate throughout the body are likely also to affect the pharynx.

Soft palate Growth of the soft palate is most rapid during the first two years of life, continuing more slowly to the age of 18 years. Length increases from about 20 mm at 3 months to 35 mm at 18 years of age, with a relatively smaller increase in thickness (Kahane, 1988, pp. 26–7). Growth progresses in such a way that the velum remains about one third longer than the anterior–posterior dimension of the nasopharynx.

Tonsils and adenoids Growth of the tonsils and adenoids, in common with most lymphoid tissue, shows an unusual growth pattern, reaching a maximum before puberty, and thereafter declining in mass (Sinclair, 1978). Tonsils and adenoids reach a maximum size at about 6 years, and then normally regress, becoming insignificant in adults. The discrepancy between this pattern of growth and that of the skeletal framework of the oral and pharyngeal cavities serves to exaggerate the effects of the skeletal growth spurt on the size of the resonating cavities.

The tongue The tongue, because of its flexible and mobile mass, is notoriously difficult to measure, which may explain the paucity of comment on tongue growth and development. As mentioned earlier, the tongue is entirely contained within the oral cavity at birth, lowering to a relatively stable position within the neck by about the fourth year of life (Laitman & Crelin, 1975, p. 214). Later in childhood, descent of the hyoid bone as the neck elongates allows the tongue to descend more, and further enlarges the oral cavity (Bosma, 1963, p. 101). At birth the tongue effectively fills the oral cavity at rest, but the facial skeleton enlarges relatively more than the tongue (Bosma, 1963, p. 101), so that the oral cavity gradually enlarges. Hopkin's (1967) study of tongue dimensions suggested that the adult tongue is only twice the size of the newborn infant's, but any two-dimensional representation of tongue size must be treated with some caution. The tongue grows differentially at its tip, acquiring what Bosma describes as a "limb like mobility." Eruption of teeth, enlargement of the oral cavity, and maturation of chewing and swallowing patterns are all associated with a more retracted tongue posture.

Lips There is little specific reference in the literature to growth of the labial aperture and labial musculature, but Kahane (1988) suggests that the facial muscles are better developed at birth than most other striated muscles, and links this to their importance in early feeding.

2.4 *Summary of vocal apparatus changes occurring during the three phases of life*

It may be helpful at this point to summarize the overall effect that all these changes have on vocal apparatus size and shape during childhood, adolescence, and senescence, as a lead into a discussion of the consequences of growth and change for speech production.

2.4.1 Birth to puberty It is between birth and puberty that the most obvious changes in size and configuration of the vocal tract occur. At birth, the respiratory system and the larynx are poorly developed, so that phonatory control is rather limited. The human vocal tract is similar to that of other mammals, in that the tongue is held forward within the oral cavity, the larynx lies fairly high in the neck, and the epiglottis can slide up to contact the soft palate. The pharyngeal space is thus very small, and does not constitute a modifiable resonating cavity of any significance during vocalization. The articulators in the oral region, i.e., the lips, jaw, and tongue, are mobile, but immature muscular control limits their voluntary use in modifying vocalizations. The lack of teeth during the first months of life also influences articulatory potential and may have an effect on tongue posture.

The most dramatic changes occur during the first five years. After this time, the configuration of the vocal tract changes more slowly, apart from the temporary changes in dentition as permanent teeth replace the primary dentition, which may have significant, though transient, effects on front oral articulation. By the end of the first decade of life the respiratory system and the larynx are becoming more mature, and the vocal tract approximates to its adult form. Muscular development and increased neuromuscular control allow progressively finer phonetic control of the vocal apparatus during speech.

2.4.2 Puberty to maturity The most striking characteristic of vocal apparatus development during the adolescent years is the rapid increase in size of some areas, which is more marked in males, leading to the emergence of sexual differentiation. The most significant sex-related differences which are evident by early adulthood are to do with overall size of the vocal apparatus, the relative size of the larynx, and the relative proportions of the resonating cavities. Both sexes show some growth in vocal tract size during this period, and full maturation of the larynx and respiratory system will influence the range of phonation available to each individual. A rapid reduction in the mass of lymphoid tissue forming the tonsils will affect the configuration of the oropharyngeal and nasopharyngeal areas. Growth of the vocal apparatus at puberty in girls can be seen mostly as a scaling up of the prepubertal vocal apparatus, but in males there are significant changes in the relative proportions of the vocal apparatus. The male larynx increases rapidly and disproportionately, and the pharyngeal cavity increases its size relative to the oral cavity.

2.4.3 Maturity to senescence General aging of the body is associated with some quite specific changes in the vocal apparatus. Respiratory function is impaired by connective tissue changes in the lungs and thoracic skeleton, and by degeneration of muscle and neuromuscular control. There are marked changes in the larynx, due to calcification of cartilages, muscular atrophy, and degenerative changes in the mucosal covering of the vocal folds. Muscular atrophy and mucosal changes will also affect the form and function of the supralaryngeal vocal tract, and the progressive loss of bone from the maxilla and mandible, together with loss of teeth, may alter the contours of the resonating cavities.

As illustration of the general morphological changes, Figure 5.10 shows a tracing of a lateral xeroradiograph of an adult male vocal tract, together with comparative tracings of lateral radiographs of the vocal tract at various stages during development.

2.5 *Consequences of growth and change for speech production*

Following this summary of organic changes during the life cycle, we can now draw some links between these and changes in phonetic output. There has been relatively little research in this area, partly due to the fact that it is very hard to extricate the relative contributions of organic and sociolinguistic factors when comparing different age and sex groups. The well-documented influences on verbal output of culture and style (e.g., Scherer & Giles, 1979) complicate research design.

This section offers a brief overview of reported age- and sex-linked differences in speech output which may be at least partially related to organic features, but the possibility of cultural determination cannot be excluded. It is well established that gender can be accurately judged from auditory recordings both in adults (Schwartz & Rine, 1968; Coleman, 1971), where there are obvious organic

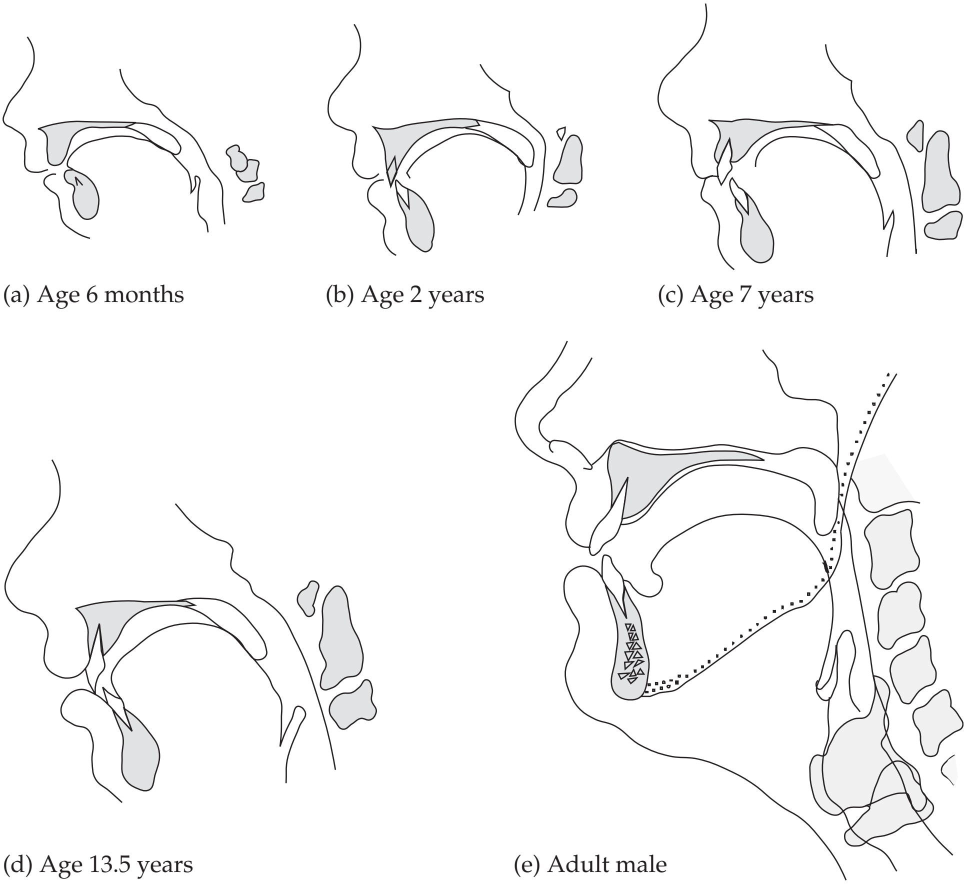


Figure 5.10 Changing proportions of the vocal tract with age. (a), (b), (c), and (d) are cephalometric tracing, and show no velic closure (adapted from Kahane, 1988, p. 24). (e) is a xeroradiographic tracing, showing velic closure during production of [ə].

bases, and in young children (Meditch, 1975; Lee et al., 1995; Nairn, 1995), where potential organic determinants are less obvious. Age, too, is reasonably well judged on the basis of auditory recordings (e.g., Ptacek & Sander, 1966; Hartmann, 1979; Linville & Fisher, 1985). It seems likely that at least some of the features which allow sex and age identification do reflect organic differences, but the task of differentiating those strands of speech quality which are specifically influenced by organic changes in the vocal apparatus from those which are learned remains largely to be done.

There is a very extensive literature relating to age- and gender-related aspects of speech production, so the following summary is necessarily selective and reflects the balance of published research by focusing on suprasegmental aspects of speech. A useful summary of life-cycle changes in voice can be found in Mathieson (2001). For further discussion of the relationship between physical changes in old age and speech, see Linville (2000, 2004).

2.5.1 Phonation characteristics

Subjective impressions Subjective comments on phonatory quality are difficult to interpret, but do indicate some common life-cycle trends. Very poorly controlled, variable vocal behavior at birth becomes rapidly more consistent during the early years of life (Wäsz-Hockert et al., 1968; Stark et al., 1975). Phonation changes at puberty are more obvious in boys, who are having to adjust to much greater changes in laryngeal structure, and who are often described as having a "husky" or "hoarse" voice quality (Greene & Mathieson, 1989; Aronson, 1980), with pitch breaks and fluctuations (Mathieson, 2001). Adolescent girls may also display some "huskiness" during hormonal changes at puberty, and the same authors describe similar changes during menstruation and pregnancy in adult women. Such impressionistic descriptions of voice quality are difficult to evaluate, but "huskiness" and "hoarseness" may usually be interpreted as some combination of whisperiness, i.e., fricative turbulence of the airflow through the glottis, resulting from incomplete vocal fold closure (Laver, 1980) and perturbation of fundamental frequency and/or amplitude of the laryngeal waveform.

The voice in old age has been given such labels as "weak," "tremulous," "hollow," "thin," "hoarse," and "breathy" (Helfrich, 1979; Greene & Mathieson, 1989), but the extent of deterioration in voice quality with age seems to be very dependent on the individual's general state of health and fitness, and on the way in which the voice has been used throughout life (Ramig & Ringel, 1983; Greene & Mathieson 1989, pp. 69–70; Linville, 2000).

Fundamental frequency There is a clear theoretical relationship between laryngeal size and mean fundamental frequency, and a considerable amount of empirical evidence to support the expected general trends (Fairbanks et al., 1949; Mysak, 1959; Hollien & Jackson, 1967; Montague et al., 1974; Benjamin, 1981; see also Helfrich, 1979 for an extensive review). Figures 5.11a and 5.11b summarize some reported average speaking f_0 at different ages, showing the different sex curves. These general trends do, of course, encompass considerable individual variation.

There seems to be general agreement that old age is associated with a slight drop in f_0 in females, which may be due to several factors. Mass increase of the vocal folds due to edema, as reported by Honjo and Isshiki (1980), would certainly be expected to lower f_0. A generalized loss of muscle tone, ossification of laryngeal cartilages, and hormonal changes in old age may all have some effect. The relationship between f_0 and age is less clear in males. The overall trend of studies reviewed in Helfrich (1979, p. 82) was for a slight increase in f_0 after the sixth decade of life, although not all studies reflect this (Wilcox & Horii, 1980). Increased f_0 in older men has been attributed to increased stiffness of the vocal folds and vocal fold atrophy (Honjo & Isshiki, 1980) and to stiffening in the areas of vocal fold-cartilage insertion (Paulsen et al., 2000). The cumulative effect of f_0 lowering in females and f_0 increase in males is a reduction in sexual differentiation of pitch.

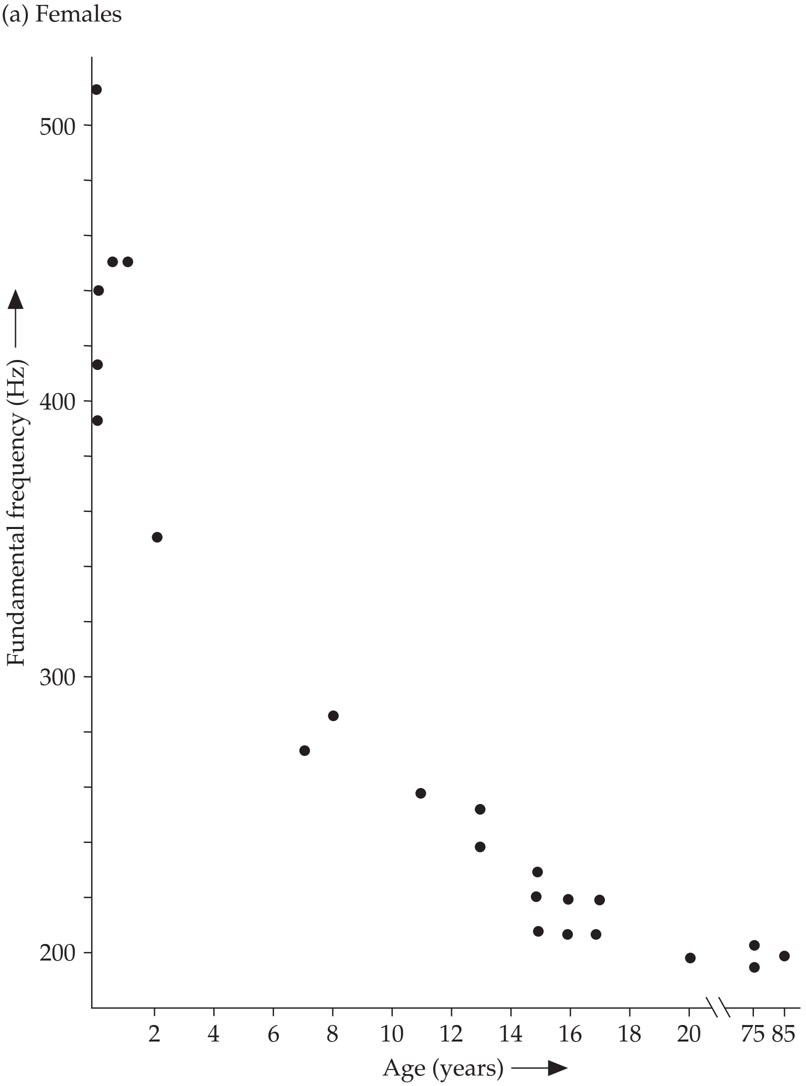


Figure 5.11 A graphic summary of reported speaking fundamental frequency (f_0) as a function of age.

Once speech is established, f_0 range seems to remain relatively constant during childhood, and then to increase between adolescence and adulthood (Helfrich 1979, p. 84). It might be expected that reduced phonatory efficiency and flexibility in old age would be associated with decreased f_0 range, but research findings are inconsistent, especially for male voices (Mysak, 1959; Ptacek et al., 1966; Hollien et al., 1971; Benjamin, 1981). It may be that different measurement procedures can partially explain this disagreement, but the sociolinguistic background and

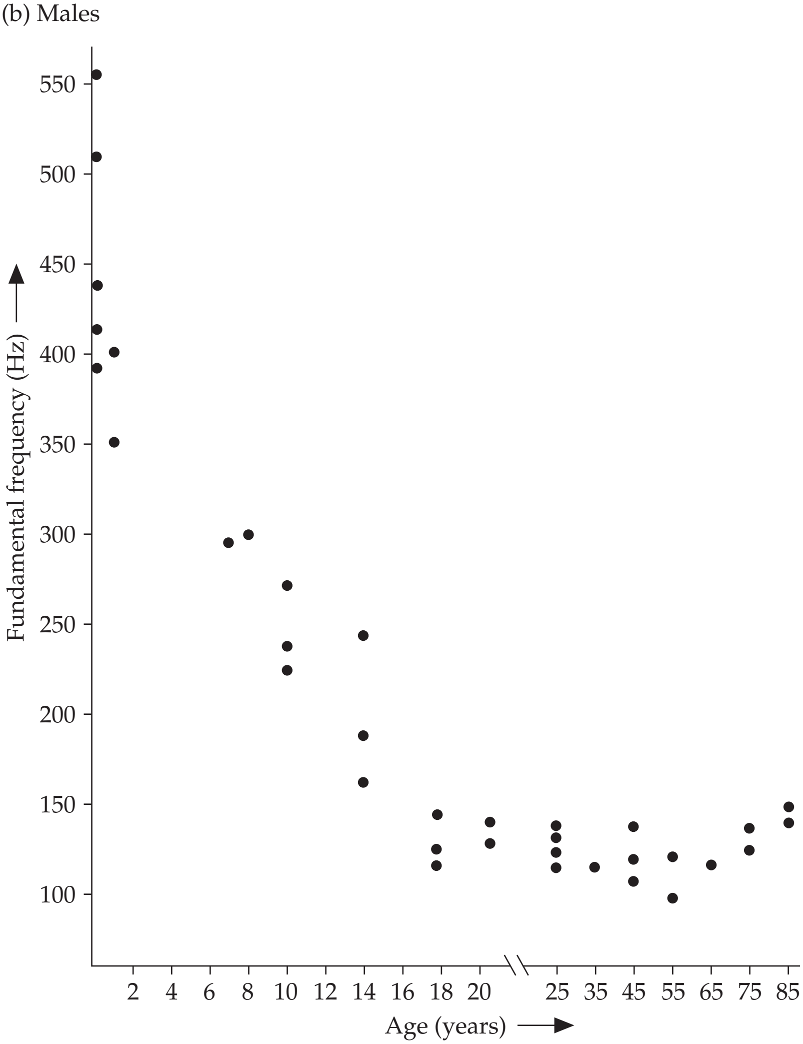


Figure 5.11 (*continued*)

emotional state of speakers may also be important (Helfrich 1979, p. 84). Linville (2000), reviewing studies of the maximum f_0 range that speakers are capable of producing, suggests that women do show a general pattern of f_0 range reduction with aging. There may be some expansion at the lower end of the pitch range, probably due to increased vocal fold mass, but limitations at the higher end of the pitch range typically result in an overall decrease in f_0 range.

Intensity There seem to be few reports on speech intensity changes during childhood, although it might be expected that increased respiratory efficiency would be associated with increasing maximum intensity. Similarly, intensity may be expected to fall as respiratory capacity decreases in old age (Ptacek et al., 1966). A complicating factor affecting habitual intensity in old age may be hearing loss, which could sometimes cause speakers to use inappropriately loud voices (Ryan & Burk, 1974; Helfric, 1979, p. 86). Studies in this area should therefore be careful to draw a distinction between maximum possible intensity and habitual intensity.

Waveform perturbations Pitch has been reported to be very unstable in infancy (Stark et al., 1975) and at puberty (Helfrich, 1979, p. 85), with rapidly varying f_0. Several studies have found indications of increased pitch perturbation (jitter) in aged voices (Benjamin, 1981; Linville & Fisher, 1985) although the increase may be rather small and jitter may be related more to general state of health than to chronological age (Ramig & Ringel, 1983; Ringel & Chodzko-Zajko, 1987; Orlikoff, 1990). Intensity perturbation (shimmer) may also increase in old age (Ramig & Ringel, 1983; Beck, 1988).

Helfrich (1979, p. 85) attributes the pitch perturbations at all ages to lack of cortical control, but variations in the tissue layer structure of the vocal fold and the associated cartilaginous framework are also likely to be important, since these can affect the efficient functioning of the vocal fold as a vibrating body. This may be especially important in the elderly age groups, where the histology of the vocal folds and of the adjacent laryngeal cartilages may be markedly degenerate.

2.5.2 Resonance characteristics The dramatic changes in vocal tract size and configuration which occur in early childhood have direct consequences for the potential range of phonetic production, but it is extremely difficult to extricate the contributions of neuromuscular maturation, language development, and organic change to overall phonetic output of young children.

There are some indications that, at least for women, age-related changes in the resonating cavities of the vocal tract may have detectable acoustic effects on formant patterns and long-term-average spectra (Linville & Fisher, 1985; Linville & Rens, 2001). The authors suggest that these effects may be explained by continuing growth of the craniofacial skeleton in adulthood, and by a lowering of the larynx in old age, albeit this appears to be proportionately greater in men than in women.

3 Interpersonal Variation

The physical characteristics of any individual depend upon the precise patterns of growth during development. It is not feasible to attempt a full discussion of the mechanisms by which the timing, amount, and pattern of growth displayed by an individual are controlled, and the aim here is simply to outline some of the factors which are known to have some influence on growth, as illustration of the complexity of the growth process and the many points at which it may be disturbed.

Factors which have been shown to influence growth fall into two classes: those which are endogenous to the individual, which generally means they are under genetic control, and those which can be loosely classified as environmental. Useful summaries of the genetic and environmental factors which may influence growth can be found in Sinclair (1978), Tanner (1978), Rona (1981), and Tanner et al. (1998). The relative contributions of endogenous and environmental factors is much disputed, and as with any nature/nurture debate, the results of studies in this area will depend on which factors are held constant. If individuals with similar or identical genetic makeup are compared, then it may be shown that environmental factors are responsible for dramatic differences in overall growth. If, on the other hand, environmental factors are held constant, then the enormous contribution of genetic factors may be clearly demonstrated. Normally it is impossible to fully extricate the effects of endogenous and environmental influences, and both obviously play major roles in determining the final shape and size of an individual. Genetic factors will determine the maximum growth potential of each person, whilst environmental factors will determine the extent to which that potential is fulfilled. The influence of environmental and genetic factors is evident not only at the level of the individual, but also when geographically and ethnically distinct populations are compared (Eveleth & Tanner, 1991).

3.1 Sources of interpersonal variation

3.1.1 Endogenous factors Whilst studies of genetically identical twins make it clear that the genetic makeup of a person plays a major role in determining his or her overall size, shape, and rate of growth and maturation, investigation of which genes are responsible is hampered by the fact that the growth process involves so many stages at which genetic control of cells may affect growth. Very many genes play a part in the process, by controlling such factors as the rates of cell division, the rates of intercellular matrix synthesis, the rates of hormone production, or the sensitivity of cells to hormonal effects.

One growth phenomenon which has a clear genetic basis is the differentiation between males and females, including the timing of onset and the duration of the pubertal growth spurt and the earlier skeletal maturation (Sinclair, 1978, p. 142).

Hormonal factors, which play a major part in growth control, are ultimately under genetic control unless there is medical intervention of some sort. A very clear summary of hormonal control of growth is provided in Tanner (1978, ch. 7), and further information can be found in Tanner et al. (1998).

3.1.2 Environmental factors Environmental factors which may be implicated in inhibiting growth potential include poor nutrition, low socio-economic status, emotional disturbance, large family size, being a younger sibling, and disease (Garn & Clark, 1975; Tanner, 1978; Lawson & Mace, 2008). There is also clear evidence that a general trend towards increased size and earlier maturity has been operational in many countries over at least the last century (Tanner, 1978, pp. 150–1; Rona, 1981). This trend seems to have slowed or stopped in Britain and some other countries, but is still continuing elsewhere. Various factors have been proposed as explanations for this phenomenon, including climatic change, a reduction in disease, improved nutrition, and genetic factors.

3.1.3 Integration and co-ordination of growth The growth process is something of an organizational miracle, and the resilience of development to adverse factors is extraordinary. Waddington (1957) used the term "canalisation" to describe the strong tendency for development to return to its original course if anything causes a temporary diversion in the normal stream of development. It is as if the architectural plans of the adult body are laid down in the genes, but the exact timing and sequence of the building stages needed to produce the adult form are fairly flexible. If development is disrupted for a while, later developmental stages can usually be modified to make up for lost time, through a "catch-up" growth phenomenon. If the rate of catch-up growth is inadequate to allow full compensation for growth delay by the normal time of cessation of growth, then maturity may be delayed to allow a longer period of growth. One interesting feature of catch-up growth is that it is more efficient in females than in males, but the reasons for this are not clear (Sinclair, 1978, p. 158).

The mechanisms by which canalization and associated phenomena such as catch-up growth are controlled are very poorly understood, although it has been suggested that the pattern of growth and development is to some extent under neural control (Tanner, 1978, p. 159). The widely varying growth patterns of different parts of the body and different tissue types must be coordinated most exactly if a properly proportioned body is to develop. Some physical characteristics can be clearly linked to specific gene effects, but a certain amount of plasticity is necessary if these physical traits are to harmonize properly. Different parts of the face, as mentioned earlier, must exert some kind of mutual growth control if they are to fit together adequately. In general, the ability of parts of the body which are under different genetic control to grow in such a way as to form an integrated whole is remarkable, although major genetic imbalances may prevent normal development and integration. Down syndrome is an obvious example of such a major, global imbalance in growth and development, and this is discussed further below.

3.2 *Illustrations of organic variations with phonetic relevance*

Two types of organic variation will be used to illustrate the way in which nonstandard anatomy may have implications for phonetic output. The first example concerns individuals who have dental malocclusions; these may be partially genetically conditioned, but environmental or behavioral factors may also play a role. The second concerns people with Down syndrome, whose genetic makeup causes a disturbance of craniofacial growth.

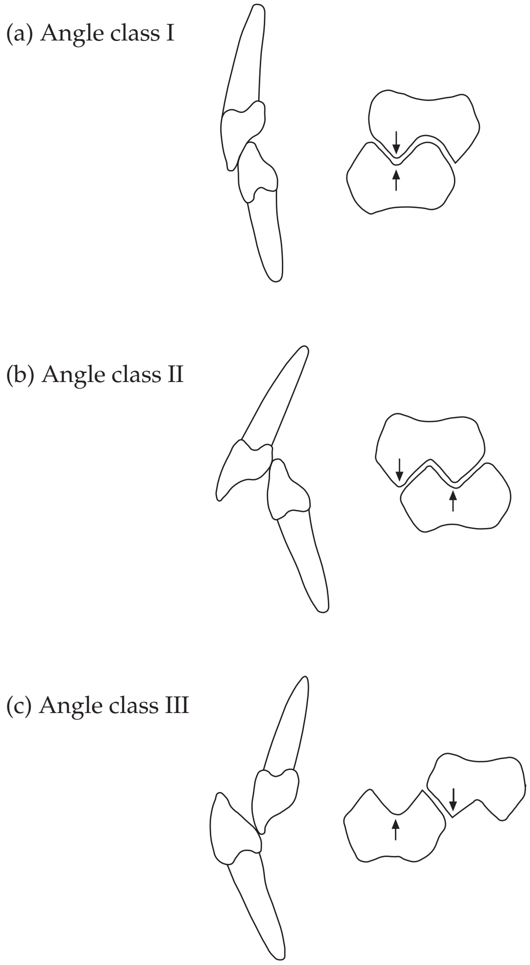
3.2.1 Malocclusion A malocclusion is defined as the abnormal relationship of one or more teeth to adjacent teeth in the same jaw, or to their normal antagonist in the opposing jaw (Hopkin, 1978). The term is commonly used more loosely to describe any dento-facial anomaly, embracing variations in morphology and relationships of the jaws and related craniofacial structures which can affect occlusion of the teeth (i.e., the relative positions of upper and lower teeth when they bite together).

Malocclusions are worthy of comment in the context of this chapter because they are very common and also because there is a strong probability that variations in dentition will affect the fine detail of articulatory patterns, even if they do not cause overt speech abnormalities. Precise incidence figures are hard to give, since studies vary so much in their standards of normality, but it is likely that at least 50 percent of individuals display at least a mild degree of malocclusion (Hopkin, 1978; Foster, 1990). Many of these will involve only the misplacement of a few teeth, and do not result from significant growth imbalances between the mandible and maxilla, but they may still have subtle effects on both the auditory/acoustic characteristics of some segmental articulations and on the precise nature of the muscular adjustments which are necessary to achieve any given lingual articulation. The effects of orthodontic treatment upon speech production should also be considered. In the short term, speakers may have to adapt articulatory patterns to the presence of intrusive orthodontic appliances. In the longer term, the desired occlusal rehabilitation may itself demand some modification of long-established patterns of speech production.

The most commonly used classification of malocclusions was developed by Angle in 1899 (Foster, 1990), and is based on the antero–posterior relationship of the maxillary and mandibular dental arches. The three main classes are summarized below.

Class I: this class shows normal arch relationships, but malpositioning of one or more teeth.
Class II: in this class the mandibular arch is posterior to the maxillary arch. This class is further subdivided according to whether all the maxillary incisors protrude abnormally (= division 1) or only the lateral incisors (= division 2).
Class III: in this class the mandibular arch is anterior to the maxillary arch.

These types of malocclusion are shown schematically in Figure 5.12. In Britain it has been reported that Angle class I malocclusions, where the jaw relationship is essentially normal but there is a variable degree of crowding, spacing, or malpositioning of teeth, accounts for about 44 percent of all malocclusions. The majority (52 percent) fall into Angle class II, while only 3–4 percent fall into Class III (Foster & Day, 1974, cited in Foster, 1990). Comparisons across different studies are complicated by variations in subject age and assessment criteria, but there is little doubt that the relative proportions of occlusal categories vary widely across differing ethnic and geographical populations. The reported percentage of Angle class III malocclusion ranges from nearly 17 percent in a Kenyan population to


(a) Angle class I

(b) Angle class II

(c) Angle class III

Figure 5.12 A schematic representation of Angle classes of malocclusion. (Adapted from Hopkin, 1978)

as low as 1.4 percent in Denmark (Solow & Helm, 1968; Garner & Butt, 1985; both cited in Uysal et al., 2005, p. 809). With regard to their implications for speech, Angle classes II and III are likely to be more important than Angle class I, because of disturbances in overall tongue-to-palate relationships.

The vertical relationship between the upper and lower incisor teeth may also be important for speech. The Angle classification suggests that in an ideal situation the lower edge of the upper incisor should lie level with the middle third of the lower incisor when the teeth are biting together (as in Figure 5.2). If there is less vertical overlap than this, this is described as reduced or incomplete overbite. The situation where there is a vertical gap between the upper and lower incisors is described as open bite.

The development of abnormal jaw relationships is interesting, because although familial trends and ethnic differences show the importance of genetic factors, there is also a large body of evidence suggesting that the development of occlusal patterns is very sensitive to diet, behavioral habits, and to the influence of disturbances in structure and function of other parts of the vocal apparatus (Foster, 1990). Habits such as thumb or finger sucking, for example may distort both the dentition and the palatal contour. Chronic obstruction of the pharyngeal airway may also affect the occlusal pattern and incisor angle, with improvements in occlusal pattern being evident following removal of tonsils or adenoids. This may be partly due to the fact that mouth breathing reduces the usual restraining forces imposed by the labial musculature, and partly due to the adoption of unusual head and tongue postures in an attempt to maintain an open pharynx (Behlfelt, 1990; Linder-Aronson et al., 1993). Malocclusion therefore offers a very clear illustration of the supremacy of the primary function of respiration over speech during development of the apparatus shared by the respiratory and speech mechanisms. The need to maintain an airway may result in a speech mechanism (and possibly a masticatory system) which is not maximally efficient.

The relationship between speech output and specific patterns of malocclusion has attracted a considerable amount of research attention (e.g., Jensen, 1968; Weinberg, 1968; Bloomer, 1971; Barrett & Hanson, 1978; Ruscello et al., 1985; Laine, 1992; Vallino & Tompson, 1993; Konopska, 2006; Hassan et al., 2007). There is a clear consensus that malocclusion may affect speech output, but findings are somewhat variable. Hassan et al. (2007), in an extensive review of research into the effects of surgical correction of abnormal jaw relationships, conclude that there is a need for further research in this area, and that there is currently a lack of clear evidence linking speech output to specific patterns of occlusion, or relating speech improvement to specific types of corrective surgery. Several studies suggest that speech is more likely to be affected in class III malocclusion (Laine, 1992; Vallino & Tompson, 1993; Konopska, 2006), but in general speakers seem to have an extraordinary capacity to compensate for malocclusal problems to produce acceptable speech. The findings of some studies of speech features associated with abnormal tongue-to-palate relationships are summarized in Figure 5.13. Although it is probably generally legitimate to view the reported speech features as being the result of dental anomalies, we should not make assumptions about the direction of the causative relationship. The arrangement of the dentition is itself conditioned partly by the muscular forces acting upon the teeth, and so may be affected by habitual articulatory patterns. Laine et al. (1985), for example, note that the association between lateral articulation of /s/ and unusual spacing of the maxillary teeth could be because a habitual pattern of lateral articulation places high pressure against the central upper incisors and causes the spacing. Although they consider that speech anomalies are more likely to be the result of dental anomalies than the other way round, this cannot be taken for granted.

3.2.2 Down syndrome Down syndrome is characterized by the presence of an additional chromosome, and one of its effects seems to be a disruption of the

Key findings	**Nature of speech assessment used**	**Author(s)**
ANGLE CLASS II		
Retracted placement of front-oral fricatives (becoming less retracted following corrective surgery)	Acoustic and electropalatography data before and after osteotomy	Wakumoto et al., 1996
"Interdental lisp" and "lateral lisp" more common	Perceptual assessment of speech quality	Blyth 1959, cited in Peterson-Falcone 1988, p. 450
/s/ is realized with more incisal opening; tongue tip may be protruded	Perceptual assessment of speech quality	Subtelny et al., 1964
Bilabial closure may be impaired	Perceptual assessment of speech quality	Bloomer, 1971; Witzel et al., 1980
ANGLE CLASS III		
Advanced tongue tip/blade articulation (becoming less advanced following corrective surgery)	Acoustic and electropalatography data before and after osteotomy	Wakumoto et al., 1996
Labiodentals may be realized as dentolabial; alveolar consonants may be realised as linguolabial	Perceptual assessment of speech quality	Witzel et al., 1980
/s/ may be produced with lower jaw position and retracted tongue posture	Perceptual assessment of speech quality	Guay et al., 1978
OPEN BITE/DECREASED OVERBITE		
Less consistent closure for alveolar and velar plosives. More posterior contact Affricates have longer duration	Electropalatography	Cayley et al., 2000
/s/ less acceptable	Perceptual assessment of speech quality	Laine et al., 1985
SPACING OF MAXILLARY INCISORS		
Advanced placement of alveolar sounds; lateral production of /s/	Perceptual assessment of speech quality	Laine et al., 1985

Figure 5.13 Phonetic features associated with malocclusion: some examples of relevant research.

Organic factor	Predicted phonetic consequences
Thick, everted lips	Protruded labial setting
Maxillary underdevelopment	Protruded jaw setting; tongue advanced relative to palate and upper teeth
Short, narrow palate + normal or large tongue	Advanced tip/blade articulations; fronted and raised tongue body setting
Pharynx reduced in anterior–posterior dimension	Pharyngeal constriction
Mucosal disorders affecting the vocal folds	Irregular vocal fold vibration and poor adduction → harshness, whisperiness
Generalized muscular hypotonia	Lax tension settings, increased nasality, open jaw, lowered larynx, minimized range of articulation

Figure 5.14 Characteristic organic features of the vocal apparatus in Down syndrome and predicted phonetic consequences.

narrow canalization of growth and development mentioned earlier. This results in increased variability in many physical characteristics. There are, nonetheless, some features of craniofacial anatomy which may be described as characteristic of the Down syndrome population. These are tabulated in Figure 5.14, together with predictions about the phonetic consequences which might be expected to result from these organic features. It should be noted that the prediction of a tendency towards an apparently "palatalized" quality is based on reports that the chromosomal imbalance in Down syndrome tends to result in palatal constriction of the vocal tract due to underdevelopment of the mid-face, with relatively normal development of the tongue and lower jaw. This contrasts with descriptions sometimes offered, which suggest that front oral constriction may be the result of an over-large tongue.

Figure 5.15 shows the results of a study of the vocal characteristics of a group of adult women with Down syndrome, compared with an age-matched control group, and it can be seen that many of these predictions are borne out by the findings. A full description of this study may be found in Beck (1988), but the results suggest that organic features in these speakers make a very substantial contribution to their overall speech quality.

4 Variation Resulting from Trauma or Disease

The vocal organs, in common with the rest of the body, have to withstand a constant barrage of attack. The vocal apparatus is particularly vulnerable to the

Vocal setting	Mean scalar degree (Max. = 6)	Prediction confirmed (see Figure 5.14)
Lip spreading	0.70	× *Lip rounding expected*
Protruded jaw	1.60**	✓
Advanced tip/blade	1.45	✓
Fronted tongue body	2.60**	✓
Raised tongue body	1.50	✓
Pharyngeal constriction	1.40**	✓
Harshness	2.70**	✓
Whisperiness	3.70**	✓
Lax vocal tract	0.95**	✓
Tense larynx	1.60	× *Lax larynx expected*
Minimized range: lips	2.20**	✓
Minimized range: jaw	1.90**	✓
Minimized range: tongue	3.00**	✓
Nasal	3.70**	✓
Open jaw	0.75**	✓
Lowered larynx	0.75	✓

Figure 5.15 Observed vocal profile characteristics for 20 adult women with Down syndrome.

** indicates vocal characteristics which are significantly different from an age-matched control group.

effects of environmental agents, sharing as it does the routes of ingress for both the respiratory and digestive systems. It is subject to invasion by infectious agents of various sorts, and has to withstand abrasion and chemical and thermal irritation caused by food passing through the mouth and pharynx as well as the effects of airborne irritants inhaled into the respiratory system. As an adaptation to this, the mucosal lining of the vocal tract is highly efficient at repair and regeneration.

Although the body's ability to repair and maintain its structure is extraordinary, tissues do vary in their ability to regenerate themselves. Disease processes and traumatic injuries themselves, and the defensive mechanisms marshaled by the body to combat disease or injury, may all involve some degree of organic change. Such change is complex and varied, and the range of alterations which may occur can be illustrated by reference to a few examples.

Pathology	**Mass change** *Mass increase → f_0 decrease*	**Stiffness change** *Increased stiffness → f_0 increase*	**Protrusion into glottis** *Incomplete adduction → whisperiness*	**Asymmetry** *Asymmetry of mass, stiffness, or contour → irregular vocal fold vibration*	**Disrupted tissue layer geometry →** *irregular vocal fold vibration*
Epithelial					
hyperplasia	+			+	
keratosis	(+)	+	(+)	+	
carcinoma-in-situ	+	+	(+)	+	
squamous carcinoma	+	+	+	+	+
verrucous carcinoma	+	+	+	+	+
adult papilloma	+	+	+	+	+
Lamina propria					
Reinke edema	+		NL		
vocal nodules	+		+	(+)	
vocal polyps: sessile	+	+	+	(+)	(+)
acute laryngitis	+		NL		
chronic hyperplastic laryngitis	+	+	NL		
fibroma		+	+	+	+
vocal polyps: pedunculated	+	+	+	(+)	+

Figure 5.16 A summary of characteristic mechanical changes in a variety of voice pathologies, indicating predicted patterns of phonation. (+) indicates that the mechanical change is sometimes, but not always, present; NL indicates that protrusion into the glottis is non-localized.

4.1 *Illustrative examples of the phonetic consequences of disease or trauma*

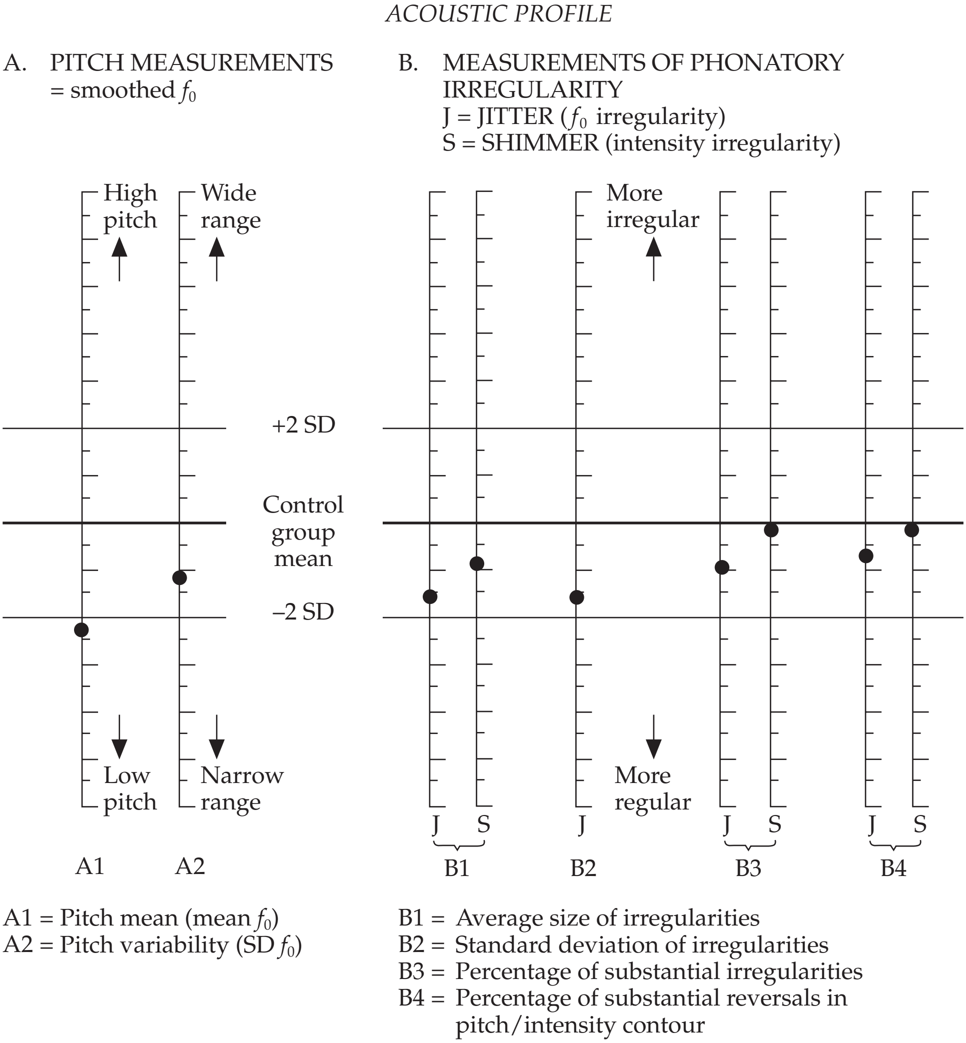
Any organic change which results from disease or injury to the vocal apparatus may have implications for speech production if it alters the morphology of the vocal organs and the resonating cavities, or if it alters the consistency and mechanical properties of the tissues which form the vocal apparatus. Tooth loss is a familiar example; the incisors are particularly vulnerable to traumatic injury, and their loss may cause minor difficulties with front oral articulations. These difficulties are usually transient, as most people adapt quickly to changes in dentition, but subtle differences in fricative quality may continue. Common examples associated with infection include inflammation of the tonsils, blockage of the nasal cavity, and laryngitis. More extreme, although fortunately less common, examples of disease-related changes include tumors of the tongue, pharynx, or larynx. In these cases, the surgical treatment itself may lead to much more severe phonetic disturbance.

4.1.1 Laryngeal disorders Phonetic output of the larynx is especially sensitive to trauma or disease because normal, regular arrangement of vocal fold tissues with varying degrees of stiffness and elasticity is essential for efficient, regular vibration. Any disruption of the tissue layers may interfere either with the mode of laryngeal vibration, or with the ability of the folds to adduct fully so as to limit air leakage during phonation (Hirano, 1981). Structural alterations of the vocal folds can be classified in terms of the mechanical alterations involved and hence the predicted mode of phonation which would be expected (Mackenzie et al., 1991; Hirano & Bless, 1993), and these predictions can be tested. Figure 5.16 is a summary of the structural changes associated with some vocal fold pathologies, and two examples from this list can be used to illustrate a possible relationship between mechanical state and vibratory pattern, as measured from the acoustic laryngeal waveform. Figure 5.17 shows acoustic profiles for two women with contrasting vocal fold disorders.

Case 1 (Figure 5.17a) is a woman with Reinke edema. This is a chronic condition, often associated with a history of smoking, characterized by fluid accumulation in the tissue at the glottal edge of both vocal folds, but without stiffening. The predicted acoustic consequence of such a symmetrical mass increase would be a reduced f_0, without any necessary increase in jitter or shimmer, and it can be seen that the acoustic results fit the predictions, with mean f_0 being the only acoustic parameter which falls outside 2 standard deviations of the normal control values.

Case 2 (Figure 5.17b) is a woman with a benign unilateral sessile polyp on her vocal fold, causing an asymmetrical increase in mass with no significant stiffening. The presence of an asymmetrical mass increase would be expected to result in increased jitter and/or shimmer as well as a reduction in f_0. Again, it can be seen that the acoustic results accord well with the predictions.

(a) Reinke edema

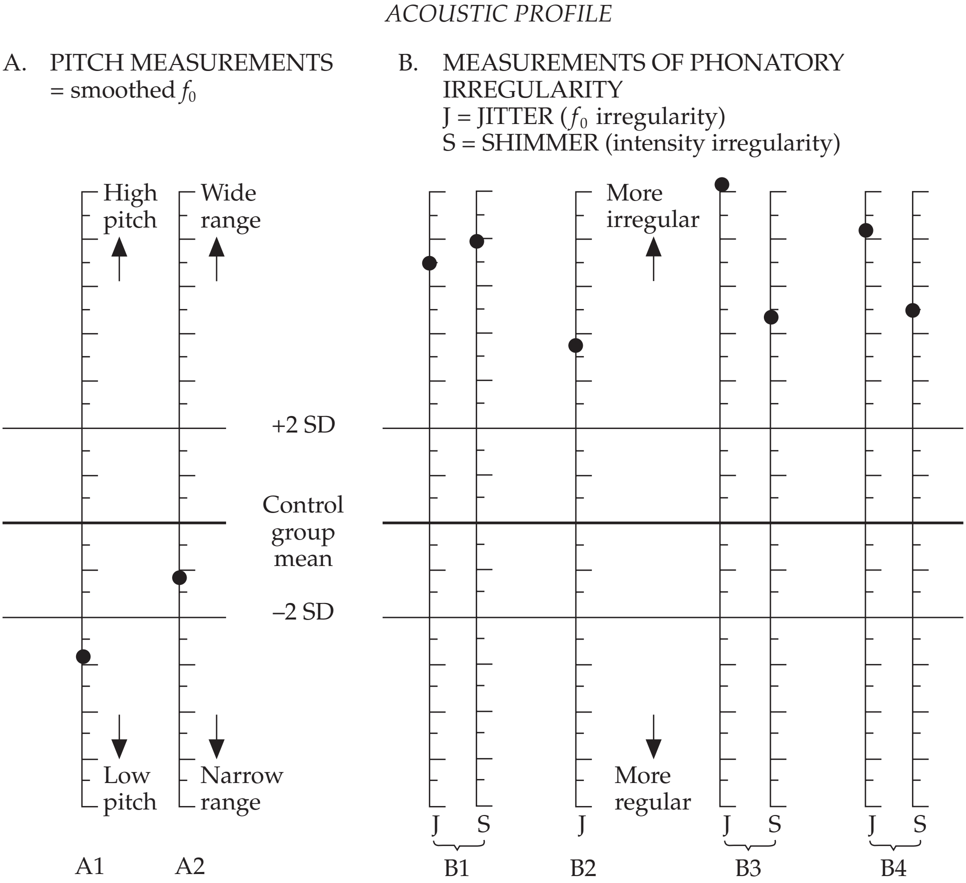


"ACOUSTIC ANALYSIS OF VOICE FEATURES" Research Project.
(MRC Grant No. G8207136) Centre for Speech Technology Research,
Department of Linguistics, University of Edinburgh.

Figure 5.17 Acoustic profiles of two women with vocal fold pathology.

Such relationships between phonetic output and structural state can be utilized in the assessment of voice disorder, and further discussions of the relationship between structural changes within the vocal folds and phonatory output can be found in Hirano (1981), Mackenzie et al. (1991), and Hirano and Bless (1993).

(b) Sessile vocal polyp



A1 = Pitch mean (mean f_0)
A2 = Pitch variability (SD f_0)

B1 = Average size of irregularities
B2 = Standard deviation of irregularities
B3 = Percentage of substantial irregularities
B4 = Percentage of substantial reversals in pitch/intensity contour

"ACOUSTIC ANALYSIS OF VOICE FEATURES" Research Project. (MRC Grant No. G8207136) Centre for Speech Technology Research, Department of Linguistics, University of Edinburgh.

Figure 5.17 (*continued*)

Although the examples given above illustrate pathological changes, which might not be typical of the general population, they differ only in degree from the familiar vocal fluctuations associated with temporary vocal fold inflammation caused by infection or excessive vocal effort.

5 Conclusion

This chapter has indicated some of the sources of variability within the human vocal apparatus and has given some illustrative examples of instances where known organic features may be linked to specific patterns of phonetic production. During our lifespan, each one of us will undergo a series of gradual changes in vocal anatomy and physiology which are the inevitable result of development and degeneration. Many processes are involved in the creation of such changes, and they will interact in subtly different ways so that each one of us is endowed with a unique vocal apparatus. In addition, the consequences of illness or trauma of various kinds may include alterations in the organic state of the vocal apparatus. These alterations may be transient, lasting for a few hours or days, as in vocal fold inflammation following sudden vocal misuse at a football match, for example, or they may be longer-term. In other words, day-to-day variations in vocal anatomy, in response to environmental factors and state of health, may be superimposed upon the types of inter-speaker differences which arise from normal variability in the cycle of development and dissolution.

Since the output of the vocal instrument at any given time depends upon its form and upon its potential for phonetic adjustment, anyone concerned with speech should be aware of the kinds of inter- and intra-personal variation in the vocal apparatus which may occur. The complex interplay between details of individual vocal tract architecture and speech production, both within the normal population and within the area of speech pathology, is largely unexplored.

The ability of widely differing speech production systems to produce utterances which, although different in terms of phonetic detail, are yet similar enough to allow cognitive recognition of linguistic “sameness” is remarkable, and prompts many questions to do with both the nature and the communicative importance of these subtle differences.

In simplistic terms, organically derived speech differences may fall into two categories. In the first, people with organically different vocal tracts might produce some utterances which appear perceptually to be genuinely identical, although the underlying muscular adjustments of the articulators are different. Acoustically this is feasible, since equivalent auditory outputs could theoretically be produced by different vocal tracts as long as the articulators are appropriately adjusted. To illustrate this, let us imagine two speakers who are organically identical except that speaker A has a high, arched palate, and speaker B has a rather shallow palate and hence a small oral cavity volume. Both might be able to produce the initial CV sequence of the word *yam* [jam] with a very similar acoustic output, but the mandibular and lingual movements in each case would be rather different. Speaker A would have to make a relatively large upwards movement of the tongue to create sufficient approximation between the front of the tongue of the palate for the approximant [j], but would be able to produce a fairly open vowel without significant jaw opening being necessary to produce the required oral cavity volume. Speaker B, on the other hand, would need less upward movement of the tongue to constrict the oral cavity for [j], but might need to lower the jaw quite markedly to facilitate sufficient tongue lowering for the vowel to be acoustically equivalent to that of speaker A.

The second type of difference occurs where two organically different speakers produce utterances which, whilst they may be perceived as phonetically similar enough to have linguistic equivalence, show minor differences in auditory quality. This is much more typical. Going back to the previous *yam* example, it is actually rather unlikely that speakers with very different palatal volumes will be able to produce speech which is really perceptually identical. The fact that speaker B has to lower his or her jaw to produce enough oral cavity volume for the vowel [a] is likely to have consequences for the degree of labial opening, and this may affect both the vowel and the nature of transitions to the final nasal consonant. This might well cause minor but detectable differences in the auditory quality.

Such hypothetical examples raise many interesting questions, which, if answered, could inform discussion of many problems in the field of phonetics. For example, how do morphological anatomical relationships within the vocal apparatus influence the dynamics and trajectories of articulatory movements? Can we improve our understanding of the relationships between articulatory factors and acoustic output if we take individual organic characteristics into consideration? To what extent is an individual's potential range of phonetic output constrained by his or her organic status? What are the implications of the trading relationships between organic and phonetic factors in speech acquisition and speech pathology? What is the basis for the concept of phonetic quality in general phonetic theory?

We can begin to answer this last, pivotal, question for general phonetic theory by noting, following Laver (1994, pp. 426–7), that "The auditory quality of every speaker's voice arises from the balance in that speaker between on the one hand organic effects of the dimensions and geometry of the vocal apparatus, and on the other the phonetic adjustments of that apparatus which the speaker habitually makes." When a given phonetic quality is produced by two speakers with different vocal tract dimensions, the balance between organic and phonetic contributions to voice quality will be different, but we can say that they share a *configurational equivalence*. Analogously, one could posit a *phonatory equivalence* between two speakers with whispery voice, where in one it was produced as a result of a learned, phonetic adjustment, and in the other by virtue of semi-paralysis of one vocal fold, preventing full closure of the glottis (Laver, 1994).

Organically-based speech differences are also interesting in the broader field of communication. We know little about the extent to which they impair intelligibility and acceptability of speech. It is likely that for the majority of organic deviations intelligibility is less of a problem than acceptability. Listeners are very willing to make judgments about a wide variety of personal and social attributes on the basis of speech quality, including social, geographical and educational background, physical stature, personality and emotional state, as well as age and gender (Laver, 1991; Scherer et al., 1991; Thomson, 1995); such judgments may have profound implications for an individual's own self-image, as well as for his or her interactions with others. Misattributions where, for example, a voice quality associated with an organic condition is interpreted as a paralinguistic signal, and vice versa, are probably fairly common. It is interesting to speculate to what extent lives may be affected by speech and voice qualities which are derived chiefly from organic states, and over which an individual has very little control. An obvious example would be that of a speaker whose voice, for some organic reason, was harsh. Given that harshness is a phonation type which is often interpreted as a signal of anger or aggression, it is entirely possible that such speakers might be unjustly judged as having aggressive personalities, with significant consequences for daily interactions with listeners making such judgments. Similar but more subtle misattributions could well be common throughout the population of speakers and listeners. We cannot, at the moment, begin to assess the potential misattributions and consequent distortions of self-image which might result from habitual speech patterns associated with such minor organic deviations as an unusually small or large larynx, idiosyncrasies of palatal contours and their relationship with tongue volume, or dental malocclusion.

Some applications of speech science demand an especially good understanding of minor organic differences between speakers. The development of systems for automatic speaker verification and the rapidly burgeoning field of forensic phonetics, for example, might both benefit from a better appreciation of the phonetic limitations imposed by any speaker's organic idiosyncrasies, as well as from an understanding of the phonetic implications of the commoner sorts of short-term organic fluctuation to which we are all prone.

As a closing note, it should be said that phoneticians may benefit from looking beyond the structures which are conventionally described as making up the vocal apparatus; these do not, of course, exist in isolation. Of particular phonetic interest are the structures which provide physical support for the vocal apparatus, such as the spine and shoulder girdle. To demonstrate this, we need only consider how deterioration of postural support in old age may limit respiratory and phonatory activity in speech production (Lieberman, 1998; Mackenzie Beck & Laver, 2004), exacerbating the age-related changes mentioned earlier. The effects of postural alignment on speech production are also of significance for phonetic measurement techniques which impose positional or gravitational constraints on speech production, such as ultrasound or MRI (Stone et al., 2007). An additional motivation for exploration of the relationships between postural support, speech output and gestural movements is that it may throw light on the physical coordination of vocal and nonverbal channels of communication (Mackenzie Beck & Laver, 2004).

The increasing sophistication of measurement techniques within speech science, coupled with a heightened awareness of the importance of organic variation in shaping phonetic output, opens up a rich seam of research, with the potential to enhance our understanding of the way in which physical attributes can color speech and communication.

REFERENCES

Altman, P. L. & Dittner, D. S. (eds.) (1962) *Committee on Biological Handbooks. Growth, Including Reproduction and Morphological Development*. Washington, DC: Federation of American Societies for Experimental Biology.

Aronson, A. E. (1980) *Clinical Voice Disorders: An Interdisciplinary Approach*. New York: Thieme-Stratton Inc.

Barrett, R. & Hanson, M. L. (1978) *Oral Myofunctional Disorders.* St Louis: Mosby Co.

Beck, J. M. (1988) Organic variation and voice quality. Doctoral Dissertation, University of Edinburgh.

Behlfelt, K. (1990) Enlarged tonsils and the effect of tonsillectomy: Characteristics of the dentition and facial skeleton; Posture of the head, hyoid bone and tongue; Mode of breathing. *Swedish Dental Journal Supplement*, 72, 1–35.

Benjamin, B. J. (1981) Frequency variability in the aged voice. *Journal of Gerontology*, 36, 722–6.

Bevis, R. R., Hayles, A. B., Isaacson, R. J., & Sather, A. H. (1977) Facial growth response to human growth hormone in hypopituitary dwarfs. *Angle Orthodontist*, 47, 193–205.

Bloomer, H. H. (1971) Speech defects associated with dental abnormalities and malocclusions. In L. E. Travis (ed.), *Handbook of Speech Pathology and Audiology*. New York: Appleton-Century-Crofts.

Bouhuys, A. (1977) *The Physiology of Breathing*. New York: Grune & Stiatton.

Bosma, J. F. (1963) Maturation of function of the oral and pharyngeal region. *American Journal of Orthodontics*, 49, 94–104.

Cayley, A. S., Tindall, A. P., Sampson, W. J., & Butcher, A. R. (2000) Electropalatographic and cephalometric assessment of tongue function in open bite and non-open bite subjects. *European Journal of Orthodontics*, 22, 463–74.

Coleman, R. O. (1971) Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research*, 14, 565–77.

Davies, D. V. & Davies, F. (1962) *Gray's Anatomy*, 33rd edn. London: Longmans, Green & Co. Ltd.

Dermaut, L. R. & O'Reilly, M. I. T. (1978) Changes in anterior facial height in girls during puberty. *Angle Orthodontist*, 48, 163–71.

Dickson, D. R. & Maue-Dickson, W. (1982) *Anatomical and Physiological Bases of Speech*. Boston: Little, Brown & Company.

Emery, J. L. (ed.) (1979) *The Anatomy of the Developing Lung*. London: Heinemann; Spastics International Medical Publications.

Enlow, D. H. & Harris, D. B. (1964) A study of the postnatal growth of the mandible. *American Journal of Orthodontics*, 50, 25.

Eveleth, P. B. & Tanner, J. M. (1991) *Worldwide Variation in Human Growth*, 2nd edn. Cambridge: Cambridge University Press.

Fairbanks, G., Herbert, E. S., & Hammond, J. M. (1949) An acoustical study of vocal pitch in seven- and eight-year-old girls. *Child Development*, 20, 71–8.

Foster, T. (1990) *A Textbook of Orthodontics*, 3rd edn. Oxford: Blackwell.

Garn, S. M. & Clark, D. C. (1975) Nutrition, growth, development and maturation: Findings from the ten-state nutrition survey of 1968–1970. *Pediatrics*, 56, 306–19.

Greene, M. & Mathieson, L. (1989) *The Voice and Its Disorders*, 5th edn. London: Whurr Publishers.

Guay, A., Maxwell, D., & Beecher, R. (1978) A radiographic study of tongue position at rest and during the phonation of /s/ in Class III malocclusion. *Angle Orthodontics*, 48, 10–22.

Hardcastle, W. (1976) *The Physiology of Speech Production*. New York: Academic Press.

Hartman, D. E. (1979) The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders*, 12, 53–61.

Hassan, T., Naini, F. B., & Gill, D. S. (2007) The effects of orthognathic surgery on speech: A review. *Journal of Oral and Maxillofacial Surgery*, 65, 2536–43.

Helfrich, H. (1979) Age markers in speech. In K. R. Scherer & H. Giles (eds.), *Social Markers in Speech* (pp. 63–107). Cambridge: Cambridge University Press.

Hirano, M. (1981). *Clinical Examination of Voice*. New York: Springer.

Hirano, M. & Bless, D. (1993) *Videstroboscopic Examination of the Larynx.* San Diego: Whurr Publishers.

Hirano, M., Kakita, Y., Ohmaru, K., & Kurita, S. (1982) Structure and mechanical properties of the vocal fold. In N. Lass (ed.), *Speech and Language: Advances in Basic Research and Practice* (pp. 211–97). New York: Academic Press.

Hirano, M., Kurita, S., & Nakashima, T. (1981) The structure of the vocal folds. In K. N. Stevens & M. Hirano (eds.), *Vocal Fold Physiology* (pp. 33–41). Tokyo: University of Tokyo Press.

Hirano, M., Kurita, S., & Nakashima, T. (1983) Growth development and aging of the human vocal cords. In D. M. Bless & J. H. Abbs (eds.), *Vocal Fold Physiology* (pp. 22–43). San Diego: College-Hill Press.

Hollien, H. & Jackson, B. (1967) Normative SSF data on southern male university students. Progress Report to NIH, Grant NB-OX397.

Hollien, H., Dew, D., & Philips, P. (1971) Phonational frequency ranges of adults. *Journal of Speech and Hearing Research*, 14, 755–60.

Honjo, I. & Isshiki, N. (1980) Laryngoscopy and voice characteristics of aged persons. *Archives of Otolaryngology*, 106, 149–50.

Hopkin, G. B. (1967) Neonatal and adult tongue dimensions. *Angle Orthodontist*, 37, 132–3.

Hopkin, G. B. (1978) The Dentition and Speech. Leaflet prepared for Speech Therapy students, Edinburgh.

Hunter, C. J. (1966) The correlation of facial growth with body height and skeletal maturity at adolescence. *Angle Orthodontist*, 36, 44–54.

Ingerslev, C. H. & Solow, B. (1975) Sex differences in craniofacial morphology. *Acta Odontologica Scandinavica*, 33, 85–94.

Jensen, R. (1968) Anterior teeth relationship and speech: Studies using cineradiography synchronized with speech reading. *Acta Radiologica Diagnosis (Stockholm)*, Supplement 276.

Kahane, J. C. (1983) A survey of age-related changes in the connective tissues of the human adult larynx. In D. M. Bless and J. H. Abbs (eds.), *Vocal Fold Physiology: Contemporary Research and Clinical Issues* (pp. 44–9). San Diego: College-Hill Press.

Kahane, J. C. (1987) Connective tissue changes in the larynx and their effects on voice. *Journal of Voice*, 1, 27–30.

Kahane, J. C. (1988) Anatomy and physiology of the organs of the peripheral speech mechanism. In J. Lass, L. V. McReynolds, J. L. Northern, & D. E. Yoder (eds.), *Handbook of Speech-Language Pathology and Audiology* (pp. 2–51). Toronto/Philadelphia: B. C. Decker Inc.

Krarup, S., Darvann, T. A., Larsen, P., Marsh, J. L., & Kreiborg, S. (2005) Three-dimensional analysis of mandibular growth and tooth eruption. *Journal of Anatomy,* 207, 669–82.

Konopska, L. (2006) Abnormal tongue position during production of Polish consonant sounds in persons with mandibular prognathism. *Annales Academiae Medicae Stetinensis*, 52, Supplement 3, 49–52.

Laine, T. (1992) Malocclusion traits and articulatory components of speech. *The European Journal of Orthodontics*, 14, 302–9.

Laine, T., Jaroma, M., & Linnasalo, A. M. (1985) Articulatory disorders in speech as related to the position of the incisors. *The European Journal of Orthodontics*, 7, 260–6.

Laitman, J. T. & Crelin, E. S. (1975) Postnatal development of the basicranium and vocal tract region in man. In J. F. Bosma (ed.), *Symposium on Development of the Basicranium* (pp. 206–19). Bethesda: National Institute of Health.

Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1991) *The Gift of Speech*. Edinburgh: Edinburgh University Press.

Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

Lawson, D. & Mace, R. (2008) Sibling configuration and childhood growth in contemporary British families. *International Journal of Epidemiology* (*IJE* advance access published online June 24, 2008: http://ije.oxfordjournals.org/cgi/reprint/dyn116v1).

Leden, H. von (1961) The mechanism of phonation. *Archives of Otolaryngology*, 74, 660.

Lee, A., Hewlett, N., & Nairn, M. (1995) Voice and gender in children. In S. Mills (ed.), *Language and Gender: Interdisciplinary Perspectives* (pp. 194–204). Longman.

Lieberman, J. (1998) Principles and techniques of manual therapy: Applications in the management of dysphonia. In T. Harris, S. Harris, J. S. Rubin, & D. M. Howard (eds.), *The Voice Clinic Handbook*. (pp. 91–138). London: Whurr Publishers.

Linder-Aronson, S., Woodside, D. G., Hellsing, E., & Enderson, W. (1993) Normalisation of incisor position after adenoidectomy. *American Journal of Dentofacial Orthopaedics*, 103, 412–17.

Linville, S. E. (2000) The aging voice. In R. Kent & M. Ball (eds.), *The Handbook of Voice Quality Measurement* (pp. 359–76). San Diego: Singular Publishing Group Inc.

Linville, S. E. (2004) Voice disorders associated with the aging process. In R. D. Kent (ed.), *MIT Encyclopedia of Communication Disorders* (pp. 359–76). Cambridge, MA: MIT Press.

Linville, S. E. & Fisher, H. B. (1985) Acoustic characteristics of perceived vs actual vocal age in controlled phonation by adult females. *Journal of the Acoustical Society of America*, 78, 40–8.

Linville, S. & Rens, J. (2001) Vocal tract resonance analysis of aging voice using long-term-average-spectra. *Journal of Voice*, 15, 323–30.

Mackenzie Beck, J. & Laver, J. (2004) Organic and performance components in vocal and nonvocal communication: Unifying concepts of description and interpretation. *Clinical Linguistics and Phonetics*, 18, 479–94.

Mackenzie, J., Laver, J., & Hiller, S. M. (1991) Structural pathologies of the vocal folds and phonation. In J. Laver, *The Gift of Speech* (pp. 281–318). Edinburgh: Edinburgh University Press.

Mády, K., Sader, R., Zimmermann, A. et al. (2001) Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction. In B. Maassen, W. Hulstijn, R. Kent, Peters, H., & P. H. H. M. van Lieshout (eds.), *Speech Motor Control in Normal and Disordered Speech: Proceedings of the 4th International Speech Motor Conference* (pp. 142–5). Nijmegen: Uitgeverij Vantilt.

Martin, D. (1961) Some facies in the diseases of childhood. *Medical and Biological Illustration*, 11, 76–84.

Mathieson, L. (2001) *Greene and Mathieson's The Voice and Its Disorders*, 6th edn. London: Whurr Publishers.

Meditch, A. (1975) The development of sex-specific speech patterns in young children. *Anthropological Linguistics*, 17, 421–33.

Montague, J. C., Brown, W. S., & Hollien, H. (1974) Vocal fundamental characteristics of institutionalized

Down's Syndrome children. *American Journal of Mental Deficiency*, 78, 414–18.

Mueller, P. B., Sweeney, R. J., & Baribeau, L. J. (1985) Senescence of the voice: Morphology of excised male larynges. *Folia Phoniatrica*, 37, 134–8.

Murano, E., Stone, M., Liu, X. et al. (2008) Compensatory tongue patterns in glossectomy patients. *Journal of the Acoustical Society of America*, 123, 3321.

Mysak, E. D. (1959) Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research*, 2, 4654.

Nairn, M. J. (1995) The perception of gender differences in the speech of 4$^1/_2$–5$^1/_2$ year old children. *Proceedings of the 13th International Congress of Phonetic Sciences*, 2, 302–5.

Negus, V. E. (1949) *The Comparative Anatomy and Physiology of the Larynx*. London: Heinemann.

O'Reilly, M. T. (1979) A longitudinal growth study: Maxillary length at puberty in females. *Angle Orthodontist*, 49, 2348.

Orlikoff, R. F. (1990) The relationship of age and cardiovascular health to certain acoustic characteristics of male voices. *Journal of Speech and Hearing Research*, 33, 450–7.

Pantoja, E. (1968) The laryngeal cartilages: Physiologic nonmineralization masquerading malignant destruction. *Archives of Otolaryngology*, 87, 416–21.

Paulsen, F., Kimpel, M., Lockemann, U., & Tillmann, B. (2000) Effects of aging on the insertion zones of the human vocal fold. *Journal of Anatomy*, 196, 41–54.

Peterson-Falzone, S. J. (1988) Speech disorders related to craniofacial structural defects, part 1. In J. Lass, L. V. McReynolds, J. L. Northern, & D. E. Yoder (eds.), *Handbook of Speech-Language Pathology and Audiology*. Toronto/Philadeplhia: B. C. Decker Inc.

Ptacek, P. H. & Sander, E. K. (1966) Age recognition from the voice. *Journal of Speech and Hearing Research*, 9, 2737.

Ramig, L. A. & Ringel, R. L. (1983) Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech and Hearing Research*, 26, 22–30.

Redman, R. S., Shapiro, B. L., & Gorlin, R. J. (1966) Measurement of normal and reportedly malformed palatal vaults, II: Normal juvenile measurements. *Journal of Dental Research*, 45, 266–9.

Ringel, R. L. & Chodzko-Zajko, W. (1987) Vocal indices of biological age. *Journal of Voice*, 1, 31–7.

Romanes, G. J. (ed.) (1978) *Cunningham's Manual of Practical Anatomy*, vol. 3: *Head, Neck and Brain*, 14th edn. Oxford: Oxford University Press.

Rona, R. J. (1981) Genetic and environmental factors in the control of growth in childhood. *British Medical Bulletin*, 37, 265–72.

Ruscello, D. M., Tekieli, M. E., and Van Sickels, J. E. (1985) Speech production before and after orthognathic surgery: A review. *Oral Surgery Oral Medicine and Oral Pathology*, 59, 10–14.

Ryan, W. J. & Burk, K. W. (1974) Perceptual and acoustic correlates of aging in the speech of males. *Journal of Communication Disorders*, 7, 181–92.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991) Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15, 123–48.

Scherer, K. R. & Giles, H. (eds.) (1979) *Social Markers in Speech*. Cambridge: Cambridge University Press.

Schwartz, M. F. & Rine, H. E. (1968) Identification of speakers from whispered vowels. *Journal of the Acoustical Society of America*, 44, 1736–7.

Seikel, J. A., King, D. W., & Drumright, D. J. (2000) *Anatomy and Physiology for Speech, Language and Hearing*, 2nd edn. San Diego: Singular Publishing Group Inc.

Shah, P. J., Joshi, M. R., & Darnwala, N. R. (1980) The interrelationships between facial areas and other body dimensions. *Angle Orthodontist*, 50, 45–53.

Shapiro, B. L., Redman, R. S., & Gorlin, R. J. (1963) Measurement of normal and reportedly abnormal palatal vaults, I: Normal adult measurements. *Journal of Dental Research*, 42, 1039.

Simon, G., Reid, L., Tanner, J. M., Goldstein, H., & Benjamin, B. (1972) Growth of radiologically determined heart diameter, lung width, and lung length from 5–19 years, with standards for clinical use. *Archives of Disease in Childhood*, 47, 373–81.

Sinclair, D. (1978) *Human Growth after Birth*, 3rd edn. Oxford: Oxford University Press.

Stark, R. E., Rose, S. N., & McLagen, M. (1975) Features of infant sounds: The first eight weeks of life. *Journal of Child Language*, 2, 205–21.

Stone, M., Stock, G., Bunin, K., et al. (2007) Comparison of speech production in upright and supine position. *Journal of the Acoustical Society of America*, 122, 532–41.

Subtelny, J., Mestre, J., & Subtelny, J. (1964) Comparative study of normal and defective articulation of /s/ as related to malocclusion and deglutition. *Journal of Speech and Hearing Disorders*, 29, 264–85.

Suzuki, N., Sakuma, T, Michi, K. T., & Ueno, T. (1981) The articulatory characteristics of the tongue in anterior openbite: Observation by use of dynamic palatography. *International Journal of Oral Surgery*, 10, 299–303.

Tanner, J. M. (1978) *Foetus into Man: Physical Growth from Conception to Maturity*. London: Open Books.

Tanner, J. M., Ulijaszek, S. J., Johnston, F. E., & Preece, M. A. (1998) *The Cambridge Encyclopedia of Human Growth and Development*. Cambridge: Cambridge University Press.

Terracol, J., Guerrier, Y., & Camps, F. (1956) Le sphincter glottique: etude anatamo-clinique. *Annales d'Otolaryngologic* (Paris), 73, 451.

Thomson, L. (1995) Listeners' judgements of intra-oral cancer patients (with anterior and posterior sites of tumour/surgical lesion), pre-operatively and post-operatively. Dissertation, Queen Margaret College, Edinburgh.

Tourne, L. P. (1991) Growth of the pharynx and its physiologic implications. *American Journal of Orthodontics and Dentofacial Orthopaedics*, 99, 129–39.

Uysal, T., Usumez, S., Memili, B., & Sari, Z. (2005) Dental and alveolar arch widths in normal occlusion and Class III malocclusion. *Angle Orthodontist*, 75, 809–13.

Vallino, L. D. & Tompson, B. (1993) Perceptual characteristics of consonant errors associated with malocclusion. *Journal of Oral and Maxillofacial Surgery*, 51, 850–6.

Waddington, C. H. (1957) *The Strategy of the Genes*. London: Allen & Unwin.

Wakumoto, M., Isaacson, K., Friel, S., et al. (1996) Preliminary study of articulatory reorganisation of fricative consonants following osteotomy. *Folia Phoniatrica et Logopaedica*, 48, 275–89.

Walker, G. W. & Kowalski, C. J. (1972) On the growth of the mandible. *American Journal of Physical Anthropology*, 36, 111–18.

Wäsz-Hockert, O., Lind, J., Vuorenkoski, I. V., Partanen, T., & Valanne, E. (1968) *Infant Cry: A Spectrographic and Auditory Analysis*. London: Heinemann.

Watson, E. H. & Lowrey, G. H. (1967) *Growth and Development of Children*, 5th edn. Chicago: Year Book Medical Publishers.

Wei, S. H. Y. (1970) Craniofacial width dimensions. *Angle Orthodontist*, 40, 141–7.

Weinberg, B. (1968) A cephalometric study of normal and defective s-articulation and variations in incisor dentition. *Journal of Speech and Hearing Research*, 11, 288–300.

Wilcox, K. A. & Horii, Y. (1980) Age and changes in vocal jitter. *Journal of Gerontology*, 35, 194–8.

Witzel, M. A., Ross, R. B., & Munro, I. R. (1980) Articulation before and after facial osteotomy. *Journal of Maxillofacial Surgery*, 8, 195–202.

6 Brain Mechanisms Underlying Speech Motor Control

HERMANN ACKERMANN AND
WOLFRAM ZIEGLER

1 Introduction

As compared to other domains of the human sensorimotor system, e.g., locomotion or upper limb movements, fewer data on the cerebral organization of speech production are available so far. Among others, these discrepancies are due to the more restricted opportunities for kinematic and electromyographic measurements at the level of the vocal tract. As a consequence, analyses of the brain mechanisms subserving articulatory and phonatory functions predominantly had to rely on perceptual as well as acoustic analyses of dysarthric deficits in patients suffering from focal cerebral lesions or neurodegenerative disorders restricted to a distinct functional system such as Parkinson disease or cerebellar atrophy (see, e.g., Kent et al., 2000). However, these data often do not allow for unambiguous inferences on the neural mechanisms underlying motor aspects of speech production. As an alternative, electrophysiological stimulation and recording techniques during, e.g., surgery (craniotomy under local anesthesia) or preoperative diagnostic evaluation of epileptic subjects (subdural or deep electrodes) provide a more direct access to the brain structures subserving speech motor control. Yet, these procedures are confined to rather small patient samples and, in any given individual, to a limited segment of cortical or subcortical structures. Functional imaging techniques such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) now provide a means for the evaluation of task-specific activity across the whole brain volume, and these procedures also have been exploited during recent years for the study of the cerebral networks underlying speech production.

Apart from methodological constraints, the investigation of the neurobiological basis of human verbal behavior is further hampered by the absence of a homologous animal model (Barlow & Farley, 1989). Primates, indeed, use acoustic signals for the sake of intra- and inter-species communication, besides visual display patterns such as facial affective expression or chest beating (Zimmermann, 1992). And beyond emotional and motivational states, monkeys also have been found capable of conveying referential information about environmental events such as the approach of a specific predator (e.g., Cheney & Seyfarth, 1996). In contrast to human speech, as well as the acquired songs of birds, the structure of subhuman primate vocalizations is, however, predominantly shaped by genetic information rather than the imitation of conspecifics (e.g., Jürgens & Ploog, 1981). Notwithstanding these discrepancies, monkey calls provide a model for the investigation of some nonpropositional aspects of human acoustic communication, e.g., affective "vocal outbursts" such as laughing or crying. Since the acoustic speech signal conveys both propositional and emotional-prosodic information, any comprehensive model of the cerebral organization of spoken language must specify how and where the sound structure of verbal utterances, on the one hand, and vocal cues of affective/motivational states, on the other, are integrated into a coherent innervation pattern directed at the vocal tract (see Jürgens, 2002, figure 10, for a recent approach in this regard). Furthermore, the engagement of orofacial and pharyngeal structures in feeding activities precedes speech production both during phylo- and ontogenetic development (Hiiemae, 2000). As a consequence, the motor control mechanisms subserving, e.g., mastication, must be expected to constrain articulatory processes during speaking. Or, in other words, nonverbal orofacial and pharyngeal functions might provide resources exploited by speech production and, thus, appear to represent evolutionary preadaptations. For example, the frame/content theory assumes the syllable structure of sentence utterances to have its origins in the "open–close" cycles of the jaw during preverbal infantile orofacial behaviors such as suckling or chewing ("ingestion-related mandibular oscillations"), reminiscent of nonspeech motor patterns in subhuman primates (MacNeilage, 1998).

This chapter will discuss the brain mechanisms of speech motor control based upon data derived from the three approaches referred to, i.e., electrical surface stimulation of the cortex, lesion studies in patients with neurogenic communication disorders, and functional imaging techniques, preceded by a review of experimental studies in subhuman primates addressing the corticobulbar representation of orofacial muscles as well as the cerebral correlates of their vocal behavior.

2 Macro- and Microstructural Characteristics of the Brain in Subhuman Primates and Man

A series of functionally linked features of the visual system, e.g., front-facing eyes, an expansion of the occipital cortex, etc., concomitant with prehensile and versatile hands, separate the order *Primates* (literally "chieftains") from other mammals, providing the basis for enhanced capabilities of visuomotor coordination (Allman, 2000). The living species of this taxon fall into six natural groups: (a) the lemurs of Madagascar, (b) the lorises and bushbabies, (c) the tarsiers, (d) the New World as well as (e) the Old World monkeys, and (f) the apes including humans. A widely used classification schema divides the primates into two suborders, the prosimians (literally "before the monkeys": lemurs, lorisies/bushbabies, and tarsiers), on the one hand, and the monkeys as well as apes, characterized by distinct novel features (grade shift), on the other. As an alternative, strepsirhines (lemurs, lorises) and haplorhines (tarsiers, monkeys, apes, humans) have been considered separate branches of the primate phylogenetic tree.

Whereas the brain of our species is three times larger than expected, as compared to anthropoid primates of the same body size, it consists, nevertheless, of "the same major cortical and subcortical structures, arranged in the same configurations and composed of neurones with the same cell architectures" (Deacon, 1992, p. 115). Furthermore, prosimians, monkeys, and apes exhibit the same basic cytoarchitectural organization of neocortex (isocortex), i.e., a sheetlike six-layered structure, representing an "innovation" of the mammalian brain (Butler & Hodos, 2005). By contrast, hippocampal and olfactory cortex encompasses only three ("limbic cortex"), the adjacent transitional zones four to five layers. Limbic and transitional areas together are often referred to as allocortex. Each neocortical layer has a distinct set of connections with other parts of the brain. Based upon differences in cytoarchitecture, i.e., shape and arrangement of neurons, the cerebral cortex can be subdivided into a multitude of distinct regions. The most widely used map traces back to the work of Brodmann, published in 1909, hence, the designation Brodmann areas (BA) (Figures 6.1–6.4). Primary motor cortex (M1 = BA 4) is characterized in anthropoids, among others, by prominent giant Betz cells within the lower portion of layer V and by the absence of an inner granular layer IV (Sherwood et al., 2004). Roughly, M1 extends from the fundus of the central sulcus to the convexity of the precentral gyrus and shows a specific pattern of excitability in that low-amplitude electrical stimulation elicits simple movements or discrete contractions of (a portion of) a muscle. Since premotor BA 6 also lacks an inner granular layer, the term agranular frontal cortex covers both BA 4 and BA 6. The latter zone also encroaches upon the medial wall of the frontal lobe (supplementary area 6 = SMA 6). Whereas SMA 6 displays a neocortical, i.e., six-layered structure, the adjacent mesiofrontal zones exhibit a transitional (BA 32, dorsalmost aspect of BA 24) or a limbic type (remaining portions of BA 24) of cytoarchitecture.

Besides cortical regions, the cerebral network of motor control in humans encompasses the basal ganglia and the cerebellum. As their major part, the basal ganglia encompass several nuclei deep in the cerebral hemispheres, separated from the thalamus by the internal capsule, including the caudate nucleus and the putamen as well as the external and the internal segment of the globus pallidus (pallidum). Together, the caudate and the putamen constitute the neostriatum. Usually, two midbrain structures, the substantia nigra and the subthalamic nucleus, as well as the so-called ventral striatum (nuclus accumbens and olfactory tubercule), are also assigned to the basal ganglia complex. Various subcomponents of the caudate nucleus, the putamen as well as the ventral striatum, are embedded into separate parallel circuits, arising from and projecting back, via the thalamus, to distinct frontal areas. Whereas, e.g., the putamen predominantly serves motor functions, the ventral striatum is assumed to mediate emotional and motivational aspects of behavior.

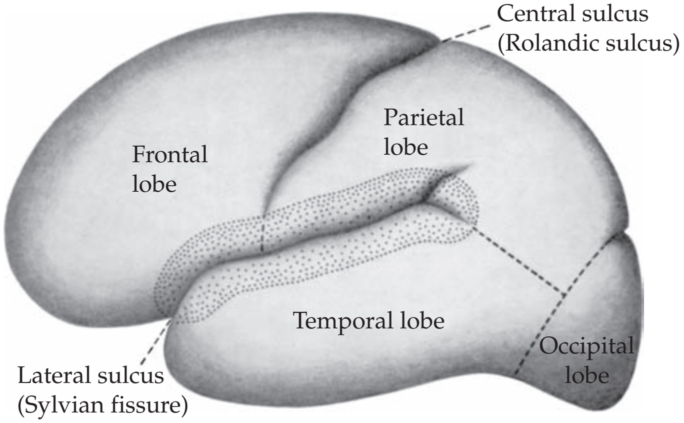


Figure 6.1 Schematic display of the four lobes of the left hemisphere of the human brain (flattened surface, gyri (cortical convolutions) and sulci (cortical furrows) not depicted).

The cerebellum, located within the posterior fossa of the skull, consists at the gross morphological level of two hemispheres, one at either side of a narrow midline ridge (vermis). Deep to the convoluted outer mantle, three pairs of nuclei can be found. All information transfer to and from other components of the central nervous system is restricted to three bilateral fiber tracts, i.e., the superior, middle, and inferior peduncles. Among others, a dense projection of the trigeminal nuclei to the cerebellum has been documented (Elias, 1990). Furthermore, the two cerebellar hemispheres are interconnected each with the contralateral frontal lobe, a network assumed to serve feedforward motor control functions (e.g., Guenther et al., 2006). These efferent projections traverse the superior peduncle (brachium conjunctivum) and target distinct thalamic relay stations.

At least in right-handers, core psycholinguistic functions, i.e., "the production and recognition of the form and literal meaning of words and sentences" (Caplan, 1987, p. 359), depend upon the integrity of an anterior (Broca's area) and posterior zone (Wernicke's area) of perisylvanian cortex in the vicinity of the lateral sulcus of the left hemisphere. Notwithstanding controversies about its exact boundaries, Broca's area is assigned to the caudal portion of the inferior frontal gyrus (IFG), encompassing, as a rule, the opercular and (posterior parts of) the triangular IFG subcomponents, roughly corresponding at the cytoarchitectural level to BA 44 and (parts of) BA 45.[1] Most noteworthy, Paul Broca (1861, p. 333) attributed articulatory functions ("exécuter la série de mouvements méthodiques et coordonnés") rather than the "language faculty" to the inferior frontal convolution. In contrast to adjacent cortex, BA 44 and 45 are characterized by particularly large pyramidal cells (magnopyramidal neurons) within deep layer III (Hayes & Lewis, 1995). Intraoperative electrical-stimulation tract tracing in humans undergoing surgical intervention for medically intractable seizures provided evidence for direct reciprocal functional connections between these posterior IFG segments and the orofacial precentral cortex (Greenlee et al., 2004).

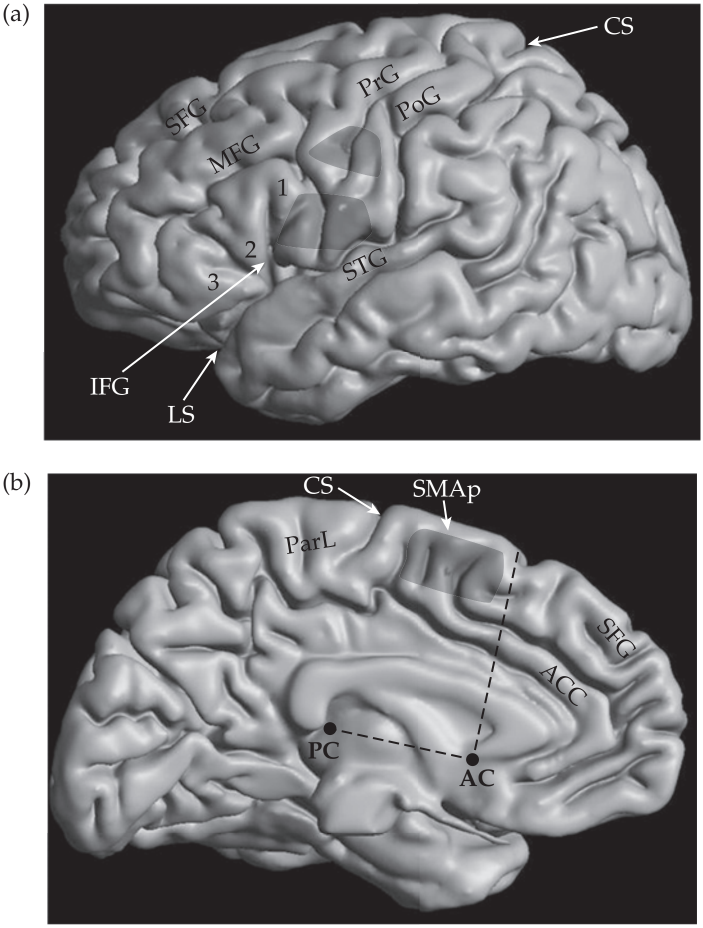


Figure 6.2 (a) Major gyri and sulci of the lateral aspect of left-hemisphere (LH) frontal lobe. Note, the inferior frontal gyrus (IFG) segregates into three components: an opercular (1), a triangular (2), and an orbital part (3). SFG = superior frontal gyrus, MFG = middle frontal gyrus, PrG = precentral gyrus, PoG = postcentral gyrus, STG = superior temporal gyrus, CS = central sulcus (Rolandic sulcus), LS = lateral sulcus (Sylvian fissure). (b) Medial wall of LH frontal lobe: the posterior component of the SFG houses the so-called supplementary motor area (SMA). The dashed line perpendicular to a plane through anterior and posterior commissure (AC–PC) roughly separates preSMA and SMA proper (SMAp). ACC = anterior cingulate cortex, ParL = paracentral lobule, i.e., the medial extensions of PrG and PoG.

Note: the shaded areas (horizontal lines) refer to the cortical areas, presumably, engaged in speech motor control (see section 10.1): bilateral primary motor cortex (Figure 6.2a, upper shaded area), opercular part of left IFG and/or lower left PrG (Figure 6.2a, lower shaded area), LH SMA proper (Figure 6.2b), and the anterior insula at the bottom of LS (not shown). The locations of the first two regions, i.e., primary motor and frontal-opercular "speech cortex," have been derived from the areas of maximum hemodynamic activation in a recent meta-analysis (Fox et al., 2001).

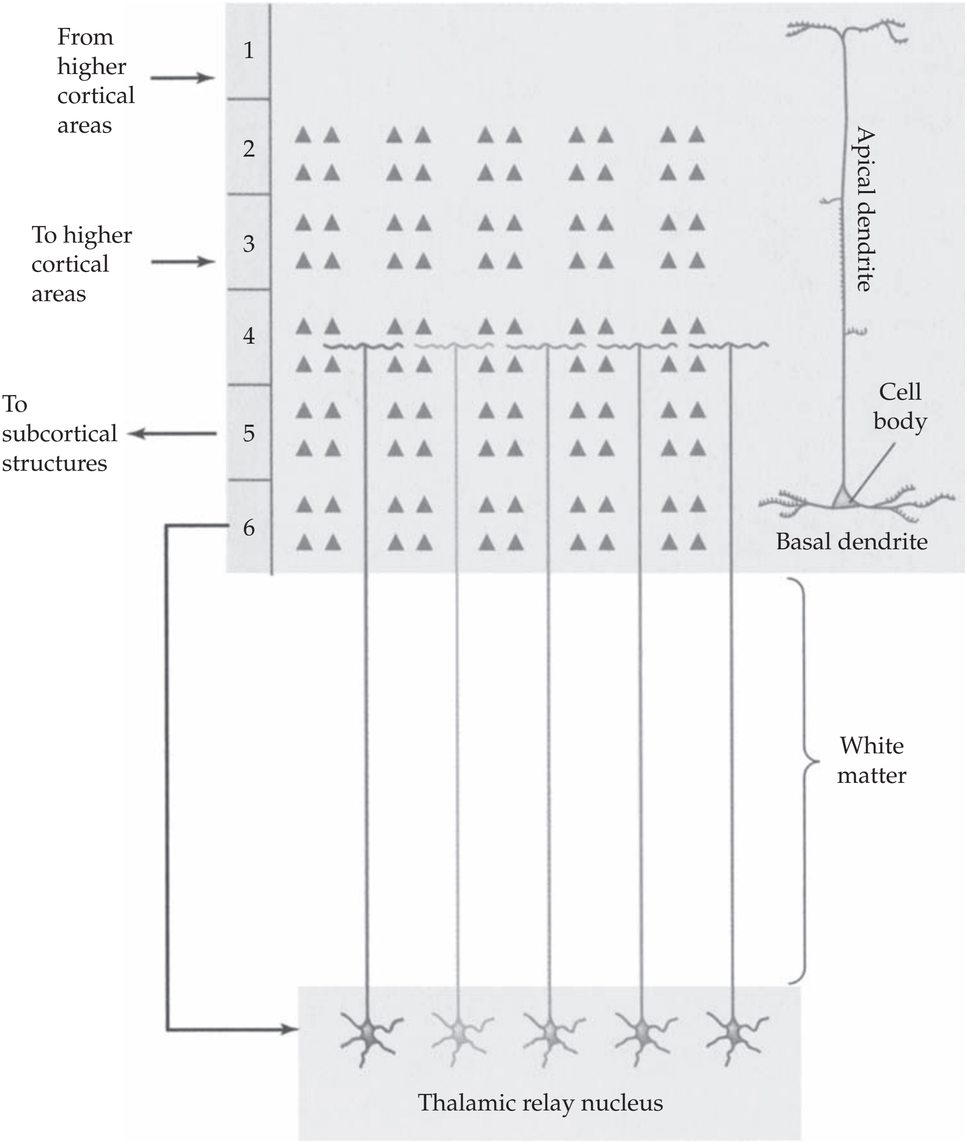


Figure 6.3 Schematic display of the cytoarchitecture of the neocortex of the human brain (modified after Allman, 2000). Each layer is characterized by a distinct pattern of connections with other cerebral components. Based upon differences, e.g., in the shape and arrangement of neurons, the cortex can be subdivided into a multitude of distinct regions.

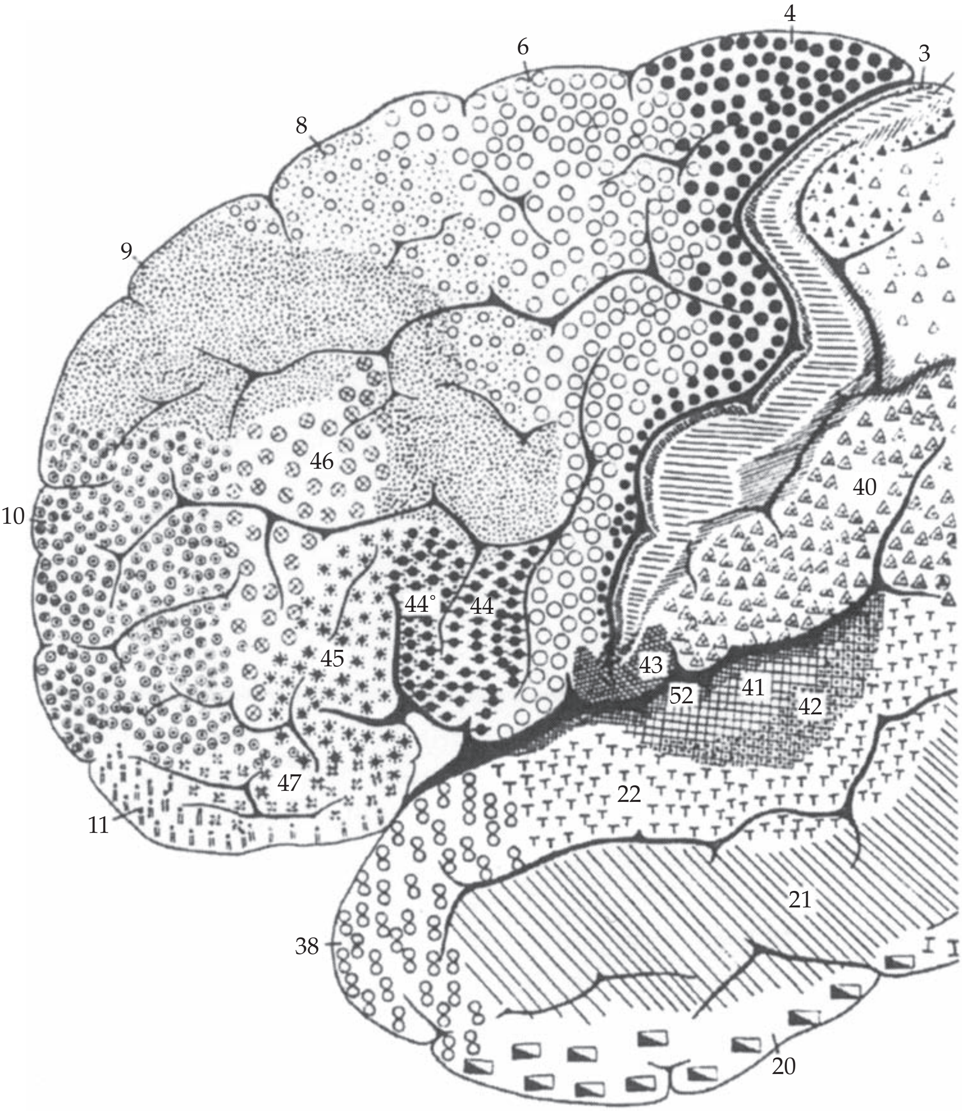


Figure 6.4 The most widely used map of cytoarchitectural cortical areas traces back to the work of Brodmann, published in 1909, hence, the designation Brodmann areas (BA). Neocortical components engaged in speech motor control include BA 4 (filled black circles), BA 6 (open circles), extending to the medial wall (SMA), and BA 44 (filled diamonds), corresponding, more or less, to the opercular part of inferior frontal gyrus.

3 Acoustic Communication in Monkeys and Apes

The neurobiological control mechanisms underlying subhuman primate phonatory functions have been most extensively studied in squirrel monkeys (*Saimiri sciureus*), a New World species with a rich repertoire of vocal behavior (for a review see, e.g., Jürgens & Ploog, 1981, as well as Jürgens, 2002). Electrical stimulation was found to elicit calls within a broad area of cerebral structures, extending from the forebrain down to the lower brainstem. Detailed analyses of concomitant behavioral and autonomic reactions as well as response latency measurements revealed, first, the anterior cingulate cortex (ACC) within the mesial wall of the frontal lobes and, second, the midbrain periaqueductal gray (PAG), including the adjacent tegmentum, to mediate primary, i.e., directly triggered stimulus responses (Figures 6.5 and 6.6). A series of subsequent ablation experiments allowed for a further characterization of the cerebral correlates of acoustic communication in *Saimiri sciureus*. For example, bilateral damage to medial forebrain structures, i.e., the cingulate cortex around the rostral pole of the corpus callosum as well as, in dorso-caudal direction, the adjacent SMA face region, resulted in a diminished rate of spontaneous vocal behavior, in the presence of an undistorted acoustic structure of the produced calls (e.g., Kirzinger & Jürgens, 1982). These mesiofrontal cerebral structures, thus, appear to predominantly mediate vocalizations triggered by an internal impulse, i.e., vocalizations with a strong volitional component, rather than responses to external events. By contrast, complete bilateral destruction of

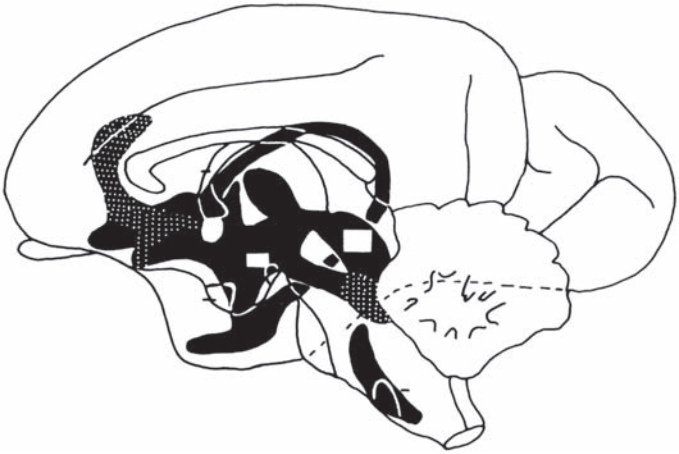


Figure 6.5 Sagittal section through the brain of a squirrel monkey. Electrical stimulation of the black-shaded regions elicits natural, i.e., species-specific calls in these animals. The two stippled areas represent (i) the anterior cingulate cortex (ACC) within the mesial wall of the frontal lobe and (ii) the midbrain periaqueductal gray (PAG), including the adjacent tegmentum. At these locations, electrical excitation gives rise to artificial vocalizations. As a consequence, the two stippled areas seem to be directly engaged in motor aspects of call production. (Adapted from Jürgens & Ploog, 1981)

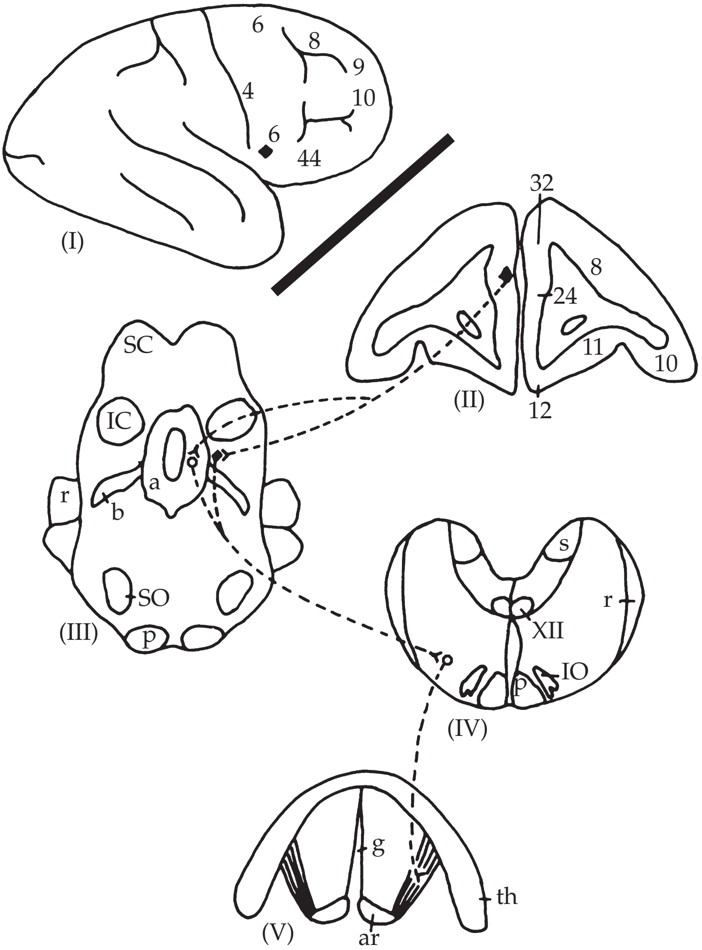


Figure 6.6 Schematic display of the hierarchical organization of the cerebral network subserving vocalizations in squirrel monkeys (I = lateral view of brain surface, II–IV = coronoal slices through ACG, PAG, and brainstem (= stippled areas in 6.5), V = schematic display of the larynx). (Adapted from Jürgens & Ploog, 1981)

the lower sensorimotor cortex, including the adjacent homologue of Broca's area, does not compromise vocal behavior, although this procedure renders the animals unable to chew, lick, and swallow. Finally, more recent investigations, using telemetric single-unit recording techniques in freely moving squirrel monkeys, were able to detect within the ventrolateral pons a vocal pattern generator responsible for the coordination of neural activity across the trigeminal, facial, and ambiguus nucleus during the production of frequency-modulated calls (Hage & Jürgens, 2006).

The cerebral network of vocal behavior, as delineated in *Saimiri sciureus*, extending from the medial forebrain via upper midbrain (PAG) to the lower brainstem, also seems to support acoustic communication in Old World monkeys. For example, electrical stimulation studies documented a rostral "cingulate vocalization region" adjoining the anterior pole of the corpus callosum (see West & Larson, 1995, for references). However, these species appear to be endowed with a more elaborate functional organization of the mesiofrontal areas supporting call generation. Whereas the various vocalizations of squirrel monkeys, more or less, represent genetically determined responses, macaques, by contrast, are capable, within some limits, of conditional vocal learning and of adjusting the structure of their calls, e.g., to reinforcement criteria. For example, bilateral ACC ablation yielded, if at all, a slightly reduced rate of spontaneous calls, but severely compromised conditioned vocal behavior (Sutton et al., 1974). And lesions superior and posterior to ACC, at the level of preSMA, resulted in significantly prolonged response latencies of stimulus-dependent conditioned vocalizations, in the presence of uncompromised nonvocal motor activities (Sutton et al., 1985). Similar to the squirrel monkey, however, damage to lateral aspects of the cerebral hemispheres such as primary motor cortex or ventral premotor areas, including the homologue of Broca's area, has no impact or just a minor impact upon acoustic communication in Old World monkeys (e.g., Aitken, 1981). As concerns apes, only sparse data about the cerebral correlates of acoustic communication are available. These primates seem to be endowed with an even more extensive cortical representation of call production than monkeys, since at least some studies reported electrical stimulation of inferior parts of the dorsolateral frontal surface to elicit vocalizations (see Sutton et al., 1974, for references).

So far, the contribution of the basal ganglia and the cerebellum, essential components of the human central-motor system, to the vocal behavior of subhuman primates has rarely been explored. Stimulation of striatal components, as a rule, failed to elicit acoustic responses in rhesus macaques (*Macaca mulatta*; Robinson, 1967). However, bilateral damage to the cerebellum may compromise conditioned calls of this species, though quite variable effects could be observed across subjects (Larson et al., 1978). By contrast, this procedure had no impact upon electrically elicited vocalizations in squirrel monkeys.

4 Cerebral Representation of Orofacial and Laryngeal Musculature in Subhuman Primates

4.1 Compartmentalization of primary and nonprimary motor cortex

Neurophysiological investigations indicate primary sensorimotor cortex of primates to support the "fractionation" of movements (Brooks, 1986). This notion also holds true for the orofacial domain: Surface stimulation of motor cortex in monkeys and man was found to predominantly elicit activity of individual muscles rather than movement sequences (see McGuinness et al., 1980, for references). The more recent technique of intracortical microstimulation (ICMS) allows for the application of very small currents and, thus, for the analysis of the functional organization of motor cortex in greater detail. ICMS studies revealed a more elaborated micro- and macrostructural organization of the primary motor area as compared to earlier surface stimulation investigations. For example, the cortical face representation was found partially to enclose and to overlap with the smaller and more laterally localized regions of the jaw and tongue musculature. Electrical stimulation of these regions predominantly elicits contralateral activity. In addition, however, considerable ipsilateral muscle innervation could be documented as well.

In macaque monkeys, tongue, face, and jaw-opening muscles appear to be associated with a more dense M1 representation than jaw-closing movements (Murray & Sessle, 1992a, 1992b). And a variety of electrophysiological investigations point at a "functional dichotomy" between tongue, face, and jaw-opening movements, on the one hand, and jaw closing, on the other, at the level of primary motor cortex. These differences in the cerebral organization of various orofacial muscles might be preserved within the domain of speech motor control, constraining articulatory gestures. For example, kinematic measurements indicate lower- and upper-lip movements during speech production to represent different coordinative structures (see, e.g., Hertrich & Ackermann, 1997a). Conceivably, at least some brain mechanisms of speech production, as documented by kinematic recordings, may reflect or even exploit organizational principles of motor cortex tracing back to our primate ancestors. As a further example, the tongue area of macaques appears to rather exclusively receive information from superficial mechanoreceptors of the face and the oral cavity (Murray & Sessle, 1992a). These data corroborate the suggestion of a minor contribution of muscle spindles to the control of articulatory gestures during speech production (e.g., Abbs & Cole, 1982).

Besides primary motor cortex, the brain of rhesus monkeys ecompasses at least four further representations of face musculature, located within (a) SMA 6, (b) rostral as well as (c) caudal parts of the cingulate gyrus, and (d) ventrolateral premotor regions (for references see Morecraft et al., 2001). The latter component of premotor cortex immediately adjoins M1 in rostral direction and exhibits a higher threshold of excitability in unanesthetized animals. The face region of SMA, bound to the anterior pole of this structure, is smaller than the respective dorsolateral representation area of the frontal lobe. Finally, the rostral and caudal cingulate face motor cortices are located within the cingulate sulcus. Anterograde labeling studies demonstrated in macaque monkeys all these five corticobulbar pathways to project to the facial cranial nerve nucleus (Morecraft et al., 2001). However, different distributional patterns of the respective nerve terminals at the level of the brainstem could be detected. As in macaques, the convexity of the frontal lobes in New World monkeys was found to encompass two separate body representations, located within the primary motor area and ventral premotor cortex, respectively (Preuss et al., 1996). Most noteworthy, both tracer and stimulation studies revealed sparse, if any, SMA neurons projecting to M1 in those species (see Tokuno et al., 1997, for references). Thus, New and Old World monkeys appear to differ in the degree of mesiofrontal representation of orofacial muscles.

4.2 *Organization of corticobulbar tracts and cranial nerve nuclei*

The trigeminal motor (Vmo), facial (VII) and hypoglossal (XII) nuclei of the brainstem channel the innervation of most cranial nerve muscles and, thus, must be considered the final common output structure for a variety of quite different movement sequences such as affective expression or food intake. Nucleus (nu.) VII, including its accessory component, predominantly comprises multipolar α-motoneurones and houses few, if any, interneurons and γ-motoneurons (Sherwood, 2005). In accordance with this observation, spindles have been reported to occur in the mandibular muscles, but at very low abundance within the lips and tongue (Loucks & De Nil, 2001; Kent et al., 1990).

Since the enhanced capabilities of facial display in Old World monkeys and in humans as well as the emergence of speech production in our species, presumably, pose increased demands on the versatility of facial musculature, differences in the structural and functional characteristics of the respective cranial nerve nuclei must be expected. In line with these suggestions, quantitative neuroanatomical studies across 102 individuals from altogether 47 primate species found a larger overall volume of the facial nucleus in humans than predicted on the basis of the data derived from nonhuman subjects (Sherwood et al., 2005). By contrast, nu. Vmo did not show these effects. Since the production of a variety of speech sounds requires fast and precise tongue movements, the respective motor cranial nerve nuclei should be more extensively supplied by corticobulbar fibers in humans than in apes (e.g., Fitch, 2000). Indeed, the size of the hypoglossal nucleus was found enlarged as compared to the nonhuman haplorhine regression line, using the medulla oblongata as an independent variable, but did not exceed the respective 95 percent confidence interval. Furthermore, the data obtained from orangutans clearly exceeded the human measures. As a consequence, the human hypoglossal system cannot be considered a straightforward index of speech production capabilities. These suggestions are further supported by the observation that the (absolute and relative) cross-sectional area of the bony hypoglossal canal, a parameter considered to reflect the number of efferent fibers of the respective brainstem nucleus, does not show systematic differences across a variety of primate species (e.g., Jungers et al., 2003).

Complex movement sequences of vocal tract muscles, such as respiration, swallowing, mastication, licking, gaping, coughing, yawning, gagging, vomiting, as well as speech production in humans, require precise spatio–temporal coordination of neural processes distributed across several brainstem nuclei. Central pattern generators (CPG) located within the medulla oblongata and the pons have been proposed to control at least five of these motor activities, i.e., respiration, swallowing, mastication, licking, and gaping (Sawczuk & Mosier, 2001). In addition, a vocal pattern generator recently has been documented in squirrel monkeys (Hage & Jürgens, 2006). Besides peripheral afferent input, these functional units are also the target of suprabulbar efferent neural signals and, presumably, represent relay stations for cortical control mechanisms, transforming, e.g., tonic input from higher-order cerebral areas into a rhythmical output pattern. In contrast, learned movement "programs" rather than brainstem CPG must be assumed to subserve motor aspects of speech production.

Apart from a single exception, i.e., rat vibrissae (Grinevich et al., 2005), there is so far no evidence for direct cortical innervation of orofacial brainstem motoneurons in subprimate species such as the cat, and the respective pathways terminate within the reticular formation of the lateral tegmentum, housing, among others, CPG of orofacial movement sequences (e.g., Kuypers, 1958a). Furthermore, prosimians and New World monkeys are also exclusively endowed with indirect, i.e., polysynaptic, cortical input to cranial nerve nuclei (for references see Sherwood, 2005, and Sherwood et al., 2005). By contrast, neurophysiological and morphological studies found direct cortical projections to brainstem motoneurons in Old World monkeys, great apes, and man (Kuypers, 1958b, 1958c; Morecraft et al., 2001; Jürgens & Alipour, 2002). Whereas the volume of the motor cranial nerve nuclei shows only rather slim differences across the primate phylogenetic lineage (Sherwood et al., 2005), considerable variation between these species seems to characterize the organization of the corticobulbar tracts (for a review see Sherwood, 2005). Presumably, the emergence of monosynaptic projections to the motor centers of the brainstem within the primate order enhances range and gradation of orofacial movements, allowing for more elaborated visual communication signals and providing the basis for voluntary control of orofacial motor activities (Sherwood et al., 2005).

5 Morphological Asymmetries of Primary and Nonprimary Motor Areas in Subhuman Primates and Man

A variety of data documented morphological population-level asymmetries of the planum temporale in terms of, e.g., a larger area at the left side, and these data have been assumed to be related to lateralization effects of speech/language functions. Similar investigations within the anterior perisylvian cortex proved to be more difficult and both postmortem as well as *in vivo* studies using, e.g., MRI scans yielded less consistent findings, because of methodological constraints and small sample sizes, among other things (see, e.g., Uylings et al., 1999). As a principal shortcoming of any investigation of cerebral asymmetries based upon surface contour measures, sulcal landmarks show considerable variation across subjects and do not allow for reliable inferences on the size of cytoarchitectonic areas, the functionally relevant subdivisions of the cortex. More recent volumetric studies based upon histological criteria and thus, circumventing these difficulties, revealed a significantly enlarged size of left-hemisphere BA 44, in the absence of comparable lateralization effects of BA 45 (Amunts et al., 1999; BA 44: left-over-right in all ten subjects, BA 45: left-over-right and right-over-left in five subjects each). These findings are relevant to investigations of the brain mechanisms of speech motor control, because clinical data indicate the opercular part of left-hemisphere

IFG to mediate higher-order aspects of articulatory performance (see below). As concerns the precentral gyrus, humans show a larger hand area, called the "knob," of the hemisphere contralateral to the preferred upper limb (for a review see Hopkins & Cantalupo, 2004). No comparable morphological asymmetries have so far been reported for the orofacial region.

Using MRI techniques, the surface area of the Broca homologue in African great apes, bounded by the fronto-orbital and the inferior precentral sulcus, respectively, was found to exhibit a left-over-right asymmetry in 20 out of a total of 27 individuals (Cantalupo & Hopkins, 2001). Most noteworthy, these primates predominantly use the right hand within the context of referential and intentional manual gestures, especially, when accompanied by vocalizations, and the respective individuals show a larger left-hemisphere IFG as compared to their non-right-handed peers (Halpern et al., 2005). By contrast, side preferences of upper-limb movements during (noncommunicative) feeding behavior in these species are correlated with asymmetries of primary motor cortex rather than the Broca homologue (Hopkins & Cantalupo, 2004). Because of a considerable macro- and microstructural interindividual variability at the level of opercular IFG in apes, these findings still must be considered with some precautions and require further support by cytoarchitecture-based volumetric studies.

Besides Broca's area, the anterior insula (anterior insular cortex) has been assumed to pertain to the cerebral network of speech motor control (see below). Volumetric brain measurements based upon 3-D reconstructions of MRI scans revealed a consistent asymmetry of the size of intrasylvian cortex towards the left hemisphere in great apes and man (Semendeferi, 2000). As concerns our species, furthermore, larger values than expected on the basis of allometric relationships derived from subhuman hominoids could be observed. However, all these observations need further support based upon quantitative analyses of larger samples.

6 Cortical Maps of Vocal Tract Muscle Representation in Humans

Using electrical surface stimulation of the brain during craniotomy, two comprehensive maps of body movements could be documented in humans, extending, first, across the precentral gyrus and, second, across SMA 6 within the medial wall of the frontal lobe (see Woolsey et al., 1979, for a review). The "homunculus" of the primary motor area was assumed to involve BA 4 as well as adjacent portions of BA 6 and to comprise a single topographic representation of the head and the body. More recent noninvasive techniques, based either upon transcranial electrical (application of brief, high-voltage pulses to the scalp) or magnetic stimulation were also able to document an orderly arrangement of leg, hand, and face movements in medio-lateral direction at the level of motor cortex and even allowed for the differential activation of upper face, lip, and tongue muscles, respectively (Rödel et al., 2003). Finally, measurements of the hemodynamic responses to a variety of upper- and lower-limb (single-joint excursions, motor sequences) as well as nonspeech orofacial activities (lip pursing) by means of fMRI corroborated this organizational pattern of the motor strip (Lotze et al., 2000). As expected, lip pursing yielded a highly symmetrical response pattern extending across the pre- and postcentral gyrus.

Several reports, published in the 1930s, noted instances of "grunts and groans" as well as "clear, sustained vowel cries" during intraoperative electrical surface stimulation of the lower motor strip in awake subjects (Penfield & Roberts, 1959). Subsequent systematic "speech-area mapping" studies at the Montreal Neurological Institute within a program of surgical treatment of medically refractory focal epilepsy observed electrical stimulation of precentral and, less frequently, postcentral orofacial areas of either hemisphere to elicit vocalizations in terms of a "sustained or interrupted vowel cry, which at times may have a consonant component" (Penfield & Roberts, 1959, p. 120). More recent investigations, based upon chronically implanted subdural electrode grids in patients with drug-resistant seizures, documented a rather broad "tongue area" across subjects, even in the absence of an organic lesion, extending across a distance of 7.5 cm (4.5 cm anterior and 3.0 cm posterior to the central sulcus) in rostro-caudal direction and encroaching, presumably, upon BAs 44 and 45 (Urasaki et al., 1994). Both arrest of alternating movements as well as involuntary excursions of the tongue (pulling back or lateral deviation) could be observed during electrical stimulation. It remains to be established how far technical constraints, e.g., current spread, account for the large primary motor tongue area or whether these data reflect a "multiplicity of movement representation," instead of "punctuate localization," within motor cortex (Murphy & Gellhorn, 1945).

Besides the lateral surface, electrical stimulation of SMA and also the dorsal bank of the cingulate sulcus have been found to elicit involuntary vocal emissions (Penfield & Roberts, 1959; see also Paus et al., 1993, p. 466). In parentheses, the designation "supplementary motor area" traces back to these observations. SMA effects may emerge at either side of the brain, more pronounced, however, in association with the language-dominant hemisphere. Furthermore, application of this procedure to mesiofrontal cortex may elicit repetition of syllables and words, e.g., during counting. Electrical stimulation of further rostral regions such as ACC fails, by contrast, to evoke involuntary vocal behavior in humans. As compared to precentral areas, SMA responses may show a more complex acoustic pattern, e.g., in terms of loudness and pitch fluctuations. It has been suggested, therefore, that this mesiofrontal area engages into the prosodic modulation of verbal utterances (Penfield & Welch, 1951). As a clear-cut contrast to subhuman primates, thus electrical surface stimulation of both pre-/postcentral cortex and SMA may elicit involuntary vocal behavior in man. Production of intelligible words, however, was never observed. Since, furthermore, the observed involuntary cries, by and large, do not display the perceptual qualities of (well-articulated) speech sounds, a direct impact upon the motor execution apparatus rather than higher-order phonological operations must be assumed. Taken together, these data indicate the presence of cortical control mechanisms in humans acting upon phonatory functions and, thus, laryngeal as well as respiratory motoneurons.

Apart from vocalizations, which were found to be restricted to the motor strip and SMA, electrical surface stimulation may result in "an inability to vocalize spontaneously" ("speech arrest"; Penfield & Roberts, 1959). At the level of the right hemisphere, speech arrest emerged only within vocalization-related areas. By contrast, the respective susceptible zone of the language-dominant side encroaches upon IFG and temporoparietal regions (for a review of more recent data see Ojemann, 1994). Principally, speech arrest could reflect disruption of either speech motor control mechanisms ("motor speech arrest") or preceding higher-order processes of language formulation such as the generation of the sound structure of verbal utterances. Apart from pre- and postcentral areas, systematic exploration of perisylvian cortex found motor speech arrest to be restricted to a small segment of the third frontal convolution rostral to the motor strip, i.e., the cortical zone where orofacial motor responses could be elicited. At that location, electrical stimuli were found to interrupt all verbal utterances, irrespective of task condition, and to result in an inability to mimic single orofacial gestures ("final motor pathway for speech" of the cortex). Besides intraoperative procedures under local anesthesia, electrical surface stimulation can be performed via chronically indwelling subdural grids of electrodes during a patient's preoperative electrocorticographic evaluation. Subsequent investigations based upon this technique were able to corroborate this notion of a cortical "final motor pathway for speech" (Lesser et al., 1984).

Hesitations or slurred speech were observed at about the same locations as speech arrest (posterior parts of the inferior frontal and lower half of the precentral gyrus), however, much less frequently (Penfield & Roberts, 1959). And subsequent stimulation mapping studies reported sporadic instances of dysarthric deficits subsequent to stimulation of the lower motor strip (one in five patients tested; Ojemann, 1979).

7 Electro- and Magnetoencephalographic Measurements of the Time Course of Brain Activity Related to Spreech Production

Electro- and magnetoencephalographic techniques represent the most direct approach available in healthy humans to the investigation of the neural correlates of distinct sensorimotor or cognitive tasks. Furthermore, given a multitude of adequately spaced channels as, e.g., in whole-head magnetoencephalography, functional brain maps can be computed on the basis of these data. Measurements of evoked electrical potentials or magnetic fields at the scalp are a widely used approach to the study of speech sound perception (for a review see Ackermann et al., 2006). So far, only a few electro- and magnetoencephalographic analyses addressing cerebral speech motor control have been conducted. Recordings of the readiness potential preceding word utterances revealed, e.g., a bilateral distribution of slow brain negativity starting at about 1.5 seconds prior to the onset of speech, significant left-lateralization effects being restricted to a very late time interval of the foreperiod (Deecke et al., 1986). Furthermore, neurophysiological investigations in epileptic patients using chronically implanted subdural electrodes were able to record movement-related potentials associated with selfpaced tongue protrusions and vocalizations over left SMA (Ikeda et al., 1992; only left side explored). Similar to the upper limbs, thus mesiofrontal areas encompassing SMA appear to participate in the control of internally triggered human orofacial movements and vocal behavior. The comparison of magnetic field changes bound to the production of visually triggered lexical items and vowel-like sounds, respectively, suggests preparation of verbal utterances to evolve across two steps that tentatively can be assigned to inferior dorsolateral frontal cortex (Broca's area and its right-hemisphere analogue) and to primary sensorimotor areas, in the absence, however, of clear-cut lateralization effects (Sasaki et al., 1996).

8 Clinical Data: Compromised Motor Aspects of Speech Production in Focal Brain Lesions and Degenerative Diseases of the Central Nervous System (CNS)

8.1 Dissociations of verbal and nonverbal vocal tract motor functions

8.1.1 Speech production and swallowing The muscles engaged in speech production also contribute to the oral and pharyngeal stages of swallowing. Nonetheless, there is no straightforward relationship between dysarthric deficits and dysphagia in CNS lesions or diseases. Indeed, patients with pseudobulbar palsy subsequent to bilateral damage to primary motor cortex or its efferent fiber tracts show a frequent co-occurrence of swallowing and speech impairments. In other central-motor disorders, e.g., Parkinson disease, dysarthria has been found to occur more frequently than dysphagia (Hartelius & Svensson, 1994). More specifically, a comparative post-mortem analysis of 83 patients who had suffered from a variety of extrapyramidal diseases presenting with signs of Parkinsonism (Parkinson disease (PD, idiopathic Parkinsonian syndrome, Morbus Parkinson), progressive supranuclear palsy, dementia with Lewy bodies, multiple system atrophy, cortico-basal degeneration) reported similarly high rates of speech motor deficits across these conditions (72–100 percent). By contrast, the prevalence of dysphagia showed a quite heterogeneous distribution, extending from only 21 percent in dementia with Lewy bodies to more than 80 percent in progressive supranuclear palsy (Müller et al., 2001). And even though speech motor deficits consistently preceded the onset of swallowing problems in all patient groups, the temporal delay between these two disorders showed considerable variation, indicating that different pathomechanisms are involved. Furthermore, the syndrome of pure dysarthria, i.e., dysarthria without dysphagia, in stroke patients represents a striking example for a dissociation between speaking and swallowing functions (Kim et al., 2003; Okuda & Tachibana, 2000). These findings cannot simply reflect enhanced demands on motor coordination during speech production, as compared to mastication or deglutition, since cerebrovascular disorders also may give rise to the reverse pattern, i.e., dysphagia without dysarthria (Lee et al., 1999). For example, a recent study found 19 out of a total of 20 patients suffering from a Wallenberg syndrome due to lateral medullary infarctions to exhibit clinical signs of dysphagia, in the presence of less frequent voice abnormalities and unimpaired articulatory functions (Aydogdu et al., 2001). Furthermore, EMG recordings documented submental muscle activity in association with the pharyngeal phase of swallowing, but not other vocal tract activities. In these instances, dysphagia was assumed to reflect a disconnection of the left- and right-sided components of the respective brainstem CPG. Obviously, disruptions of this system do not interfere with motor aspects of speech production. At least at some levels of the neuraxis, the network associated with swallowing functions may, therefore, be spared in dysarthric subjects, and vice versa.

8.1.2 Speech production and facial/vocal expression of emotions Vocal expression of affective states provides a further example of a clinical dissociation between speech production and nonverbal vocal tract functions. Thus, patients with pseudobulbar palsy may show preserved voiced phonation and orofacial movements during laughter and crying, despite severely impaired or even abolished voluntary control of the muscles supplied by the lower cranial nerves. In these subjects, furthermore, compromised innervation of the vocal tract due to a disruption of the upper motor neuron may give rise to a release or disinhibition of nonverbal affective vocalizations ("pathological laughter/crying"; see, e.g., Ackermann, 2006). Conceivably, these emotional "bursts" are mediated by the "limbic vocalization system" as delineated in subhuman primates (see above). Thus, two distinct cerebral pathways appear to act upon the motor nuclei of the lower cranial nerves in humans ("dual pathway" model of acoustic communication; Larson, 1988).

8.2 Compromised motivational mechanisms of verbal communication: The role of mesiofrontal cortex

Patients suffering from left-sided SMA lesions may exhibit reduced spontaneous verbal behaviour, in the absence of any central-motor disorders of the vocal tract muscles and any deterioration of language functions. This pattern of disrupted acoustic communication has, by some authors, been considered a variant of transcortical motor aphasia (Freedman et al., 1984). In addition, dysfluent, i.e., stuttering-like speech utterances in terms of sound prolongations and syllable repetitions have been observed in subjects with mesiofrontal lesions of the dominant hemisphere (Ackermann, Daum, et al., 1996; Ackermann, Hertrich, et al., 1996; Ziegler et al., 1997). Furthermore, bilateral damage to mesiofrontal areas, encroaching, presumably, upon the anterior cingulate gyrus (ACG) and its projections to SMA, may give rise to the syndrome of akinetic mutism, characterized by a lack of self-initiated motor activities, including speech production (for a review see Ackermann and Ziegler, 1995). In consideration of these clinical data, the medial wall of the frontal lobes appears to mediate via ACG motivational aspects of verbal motor behaviour and SMA has been assumed to operate as a "starting mechanism of speech" (Botez & Barbeau, 1971).

Apart from reduced spontaneous speech production, the verbal utterances of patients suffering from mesiofrontal lesions may be characterized by "flat" and monotonous intonation (Rubens, 1975; Jürgens & von Cramon, 1982), resembling the syndrome of motor aprosodia subsequent to dysfunctions of the basal ganglia (Cancelliere & Kertesz, 1990).

8.3 Impaired phonetic planning mechanisms in language production: The role of left-hemisphere inferior dorsolateral frontal and anterior insular cortex

8.3.1 Apraxia of speech The prearticulatory phase of spoken language production is assumed to encompass, among others things, the generation of a motor program or a phonetic plan, providing the input to the execution apparatus (e.g., Levelt et al., 1999). Dysfunctions of this stage of speech motor control may give rise to the syndrome of apraxia of speech (AOS; see Ziegler, 2008, for a recent review). In the majority of cases, AOS arises from ischemic infarctions within the area of blood supply of the left medial cerebral artery and, hence, must be considered a syndrome of the language-dominant hemisphere. Several experimental psycholinguistic studies indicate that this constellation reflects an impaired capability to plan speech movements at the level of syllable-sized or even larger linguistic units (Aichert & Ziegler, 2004; Ziegler, 2005).

The neuroanatomic correlates of AOS have been a matter of dispute since the 1860s. Whereas Broca had assigned the "faculty of articulate language" to the posterior part of the left-hemisphere inferior frontal convolution, a region now bearing his name (Broca, 1861), this notion has repeatedly been challenged since then, mainly because damage to Broca's area is not necessarily associated with persisting articulatory deficits (e.g., Alexander et al., 1990).

8.3.2 Subcortical white matter Déjerine (1891) assumed any impairments of the "faculty of articulate language" to reflect damage to fiber tracts connecting Broca's area with brainstem motor centers. Contrary to this notion, however, the corticobulbar projections engaged in speech production most probably show a bilateral organization, and unilateral dysfunctions of this system can be compensated for within rather short time intervals (Muellbacher et al., 1999). Nevertheless, the suggestion that subcortical mechanisms may contribute to the development of AOS has been upheld until very recently. For instance, a review of 13 cases drawn from the literature, extended by four additional new observations, concluded that besides the opercular part of IFG white matter structures beneath the inferior motor strip, including the anterior limb of the internal capsule of the language-dominant hemisphere, participate in the simultaneous and sequential organization of motor actions into well-articulated speech (Schiff et al., 1983). Furthermore, a subsequent comprehensive analysis of the pathways involved in severe aphasic nonfluency, reflecting at least partially an AOS component, proposed a significant contribution of two distinct left-hemisphere subcortical regions to the development of persisting articulatory deficits: (a) the medial subcallosal fasciculus deep to Broca's area, and (b) the middle portion of the periventricular white matter, beneath the mouth and face representation areas of primary sensorimotor cortex (SMC; Naeser et al., 1989). A subsequent fMRI study, based upon four patients with severe persisting nonfluent speech production, assumed damage to these white matter areas to result in transcallosal disinhibition of right-hemisphere primary motor cortex and SMA, thus compromising spoken language communication production (Naeser et al., 2005).

8.3.3 Left inferior precentral cortex A series of case studies attributed the syndrome of AOS to a lesion of the primary motor representation area of the face, mouth, and larynx muscles within the language-dominant hemisphere (e.g., Tanji et al., 2001). In these instances, a unilateral upper motor neuron syndrome (UMNS; see, e.g., Urban et al., 2001) must be expected, characterized by, among other things, a paresis of the supplied musculature. However, first, dysfunctions of the upper motor neuron cannot explain the articulatory deficits of AOS. Second, while the predominantly bilateral organization of the speech-relevant components of the corticobulbar system usually allows for a compensation of unilateral disorders within a few days or weeks (Muellbacher et al., 1999), many AOS patients show severe and persistent deficits of verbal communication.

8.3.4 Left anterior insular cortex Damage to anterior parts of left-hemisphere insular cortex often has been assumed to impede speech motor functions (Mohr et al., 1978; Shuren et al., 1995). The most substantial evidence derives from a comprehensive investigation conducted by Dronkers (1996), including 25 AOS patients and 19 nonapraxic subjects suffering from a single left-sided ischemic infarction each. A small region (precentral gyrus) of the anterior insula represented the area of maximum overlap of the lesions associated with speech motor deficits. By contrast, this region was found to be spared in the 19 nonapraxic patients. A series of subsequent single-case and group studies also reported AOS subsequent to a left-sided anterior intrasylvian lesion (e.g., Nestor et al., 2003).

8.3.5 Broca's area Recently, Broca's suggestion that "the faculty of articulate language" is bound to posterior IFG received new substantial support, based upon a large-scale study including 80 patients with left-hemisphere nonlacunar strokes, encroaching upon or sparing insular cortex (Hillis et al., 2004). All participants underwent screening of speech motor functions within a time interval of 24 hours after stroke onset. This investigation thus also included subjects with transient disorders of verbal communication, a group which might have gone unnoticed in lesion studies based upon chronic AOS cases. Furthermore, diffusion- and perfusion-weighted MRI measurements were conducted in order to detect, apart from structural abnormalities, areas of hypoperfusion. Most noteworthy, AOS was not linked to lesions of the left-hemisphere anterior insula in this large sample of patients. Rather, verbal apraxia was found to depend upon hypoperfusion of and/or structural damage to Broca's area. In the absence of any morphological or hemodynamic changes at the level of posterior IFG, articulatory performance was unimpaired. Conceivably, the large MCA infarctions, usually associated with chronic AOS, extend by virtue of the organization of the cerebral vascular system to intrasylvian cortex. As a consequence, an association between AOS and damage to the anterior insula, though statistically reliable, does not necessarily represent a causative relationship. Based upon these findings and suggestions, left-hemisphere posterior IFG rather than intrasylvian cortex appears to be crucially engaged in higher-order motor aspects of speech production.

8.3.6 Summary Based upon the available clinical literature, AOS cannot be unambiguously attributed to a circumscribed cerebral lesion. As an alternative model, the various regions found to be associated with this syndrome so far, i.e., left inferior precentral region, the opercular portion of IFG, as well as the anterior insular cortex, concomitant with the underlying white matter tissue, might constitute a specialized motor network integrating simultaneous and sequential vocal tract movements into well-articulated speech (Schiff et al., 1983). It is well established that long-term practice of specific motor skills gives rise to a distinct functional reorganisation of rolandic motor areas (for a review see Ungerleider et al., 2002). Since adult speech capabilities are based upon extensive motor learning processes, these mechanisms may account for a specific contribution of left inferior precentral cortex to motor aspects of speech production. Whereas (unilateral) damage to these structures would spare elementary motor functions such as muscle strength, because of a mostly bilateral organization of the upper motor neuron system projecting to the lower cranial nerve nuclei, a dysfunction of any critical components of this network may compromise the complex interplay of articulatory gestures during verbal communication.

8.4 Impaired motor execution mechanisms engaged in speech production

8.4.1 Syndromes of paretic dysarthria: Disorders of the primary (Rolandic) motor cortex and its efferent corticobulbar pathways (upper motor neuron) Damage to the efferent projections of the face, mouth, and larynx areas of primary (rolandic) motor cortex to the respective brainstem nuclei represents the most salient pathomechanism of dysarthric deficits in stroke patients. For example, an investigation of 68 consecutive subjects with a sudden onset of speech motor deficits subsequent to a single ischemic infarction was able to document in the majority of cases (49 out of 68) a lesion of corticobulbar fiber tracts at the level of lower precentral cortex, centrum semiovale, internal capsule (genu and ventral part of the dorsal segment), cerebral peduncles, basis pontis, or the ventral ponto-medullary junction (Urban et al., 2001). Since the brainstem nuclei subserving the innervation of vocal tract muscles receive input from both cerebral hemispheres, with the exception, by and large, of the lower face, unilateral dysfunctions of the upper motor neuron, i.e., primary motor cortex and corticobulbar tracts, usually fail to elicit persistent dysarthria. By contrast, bilateral damage to these structures either at the cortical or the subcortical level (e.g., pseudobulbar palsy) may give rise to the syndrome of spastic dysarthria, characterized by, among other things, slowed speaking rate, reduced range of orofacial movements, velar insufficiency, and hyper-adduction of the vocal folds. In its extreme, anarthria and/or aphonia may develop under these conditions, e.g., the Foix-Chavany-Marie syndrome. In case of a supranuclear localization of the relevant lesion, orofacial, pharyngeal, and laryngeal reflex mechanisms must be preserved, and emotional mimic or vocal behavior remains intact as long as the efferent bulbar projections of mesiofrontal motor areas are spared. By contrast, damage to the brainstem nuclei and the cranial nerves (lower motor neuron), engaged in the innervation of vocal tract structures, compromises both verbal and nonverbal functions of the respective muscle groups.

Unilateral damage to the corticobulbar system yields, as a rule, only mild and transient dysarthric impairments (see above), developing into, by and large, normal phonatory and articulatory functions within a time interval of several days to a few weeks (Urban et al., 1999). In order to further delineate the neural basis of these rapid recovery processes, Muellbacher and co-workers (1999) studied the role of the intact hemisphere in subjects suffering from a unilateral hypoglossal paresis after hemispheric stroke, using transcranial magnetic stimulation techniques. In five out of a total of six patients, the abnormal conduction characteristics of the crossed cortico-nuclear pathways, arising within the affected hemisphere, persisted even after complete recovery of tongue paresis. These observations suggest that pre-existing uncrossed motor pathways of the intact hemisphere participate in the compensation of these cranial nerve deficits. A subsequent investigation of lingual motor functions revealed the muscle fibers of either side of the tongue to receive the same information from each hemisphere in terms of inhibitory and facilitatory control signals. As a consequence, contra- and ipsilateral functions of the axial musculature engaged in speech production seem to depend upon a common cortical network, conveying identical signals to the brainstem motor nuclei on either side (Muellbacher et al., 2001). The notion that contralesional structures compensate for lost articulatory skills after unilateral infarctions has also been supported by a recent follow-up functional imaging study in a patient with left-hemisphere striato-capsular infarction: The emergence of a mirror-like reversal of hemodynamic activation at the level of motor cortex and cerebellum was found to parallel recovery of speech (Riecker et al., 2002).

Geschwind (1969) assumed left-hemisphere dominance of speech production to extend beyond linguistic aspects of language processing to the cortical areas steering vocal tract motor functions. This organizational pattern of neural control mechanisms might avoid lateral competition during innervation of midline vocal tract muscles and, thus, prevent asynchronous input to the relevant brainstem nuclei. So far, however, clinical data provide only weak empirical support for this model. Indeed, a recent study reported a relative dominance of left-sided lesions of the cortico-nuclear system in a group of 53 patients suffering from dysarthria after unilateral extra-cerebellar stroke (89 percent left- as compared to 11 percent right-hemisphere involvement; Urban et al., 2006). In line with preceding reports, nevertheless, these findings point toward a considerable, i.e., non-negligible, subgroup of patients with transient speech motor deficits subsequent to right-sided damage to the upper motor neuron (Duffy & Folger, 1996; Urban et al., 1997, 2000).

8.4.2 Syndromes of hypo- and hyperkinetic dysarthria: Disorders of the basal ganglia Parkinson disease (PD, also known as idiopathic Parkinsonian syndrome) is primarily characterized by progressive degeneration of dopaminergic pathways, arising within the substantia nigra and projecting to the basal ganglia. The salient motor signs of this disorder such as akinesia, bradykinesia, hypokinesia, and rigidity are assumed to reflect presynaptic depletion of this neurotransmitter at the level of the striatum. Tracing back to the work of Darley et al. (1975), the same pathomechanisms, acting upon vocal tract muscles, are considered to be engaged in PD dysarthria, giving rise to, among other things, monotonous pitch, reduced loudness, breathy and harsh voice quality, and imprecise articulation (see, e.g., Duffy, 2005). Hence, these perceived speech motor deficits are lumped together into the syndrome of hypokinetic or rigid dysarthria. Whereas dopamine substitution and/or dopamine agonists have a significant impact upon, e.g., upper-limb motor functions, dysarthric deficits show, by contrast, a less pronounced or even missing responsiveness to this therapeutic approach. Similar discrepancies have been reported for surgical procedures such as the more recent technique of deep brain stimulation. As a consequence, these data point toward a significant contribution of nondopaminergic mechanisms to PD dysarthria (Pinto et al., 2004, 2005).

Besides PD, Huntington chorea (HC), an autosomal-dominant hereditary disorder, represents, within some limits, a further paradigm of a striatal dysfunction. Whereas the degenerative process, at least in the earlier stages of the disease, predominantly involves the caudate nucleus and the putamen, other components of the basal ganglia, the thalamus, the cerebellum as well as neocortical areas also will be compromised during further follow-up. The characteristic clinical signs of HC encompass choreatic hyperkinesia, personality changes such as increased irritability, and progressive decline of cognitive functions. Perceptual speech evaluation in these patients revealed, among other things, fluctuating pitch and/or loudness, involuntary vocalizations, and "overshooting" articulatory gestures, abnormalities assumed to reflect hyperkinetic activity of vocal tract muscles (Duffy, 2005). The few available instrumental investigations of HC dysarthria, e.g., segment measurements at the acoustic speech signal and kinematic analyses of lower-lip movements, do not yet allow for unambiguous inferences on the pathomechanisms of these speech motor deficits (see, e.g., Hertrich & Ackermann, 1994, as well as Ackermann, Hertrich et al., 1997).

It is widely acknowledged that the various subcomponents of the basal ganglia are embedded into several parallel reentrant cortico–subcortico–cortical circuits, arising in distinct areas of the frontal lobes and projecting back via specific thalamic nuclei to their origin (DeLong & Wichmann, 2007). Based mainly upon anatomical labeling and tracing studies in subhuman species, there is experimental evidence for up to five separate routes. Roughly, the motor loop, the most widely studied pathway, includes primary motor cortex as well as several premotor areas, including SMA, the putamen, a direct and an indirect pathway through the basal ganglia, and, finally, distinct thalamic nuclei. By contrast, cognitive capabilities are assumed to be associated with connections between prefrontal cortex and the caudate nucleus, whereas limbic-affective functions depend on projections of ACG to the ventral striatum. Functional imaging data obtained in humans are consistent with at least a tripartite division of the striatum into motor, associative, and limbic zones (Postuma & Dagher, 2006). The suggestion of parallel circuits running through the basal ganglia does not necessarily imply strictly segregated data processing. As an alternative, the "information funneling" hypothesis proposes that the various loops enter these subcortical nuclear complexes via separate subcomponents of the neo- and ventral striatum, but converge towards the pallidum and/or substantia nigra (Parent & Hazrati, 1995). On these grounds, the implementation of affective prosody during speech production could be bound to a "cross talk" between limbic and motor circuits within the basal ganglia.

Rather than neurodegenerative diseases such as PD or HC, focal pathology might allow for more fine-grained correlations between the various cortico–striato–thalamo–cortical loops and different functional aspects of speech motor control. Unfortunately, the available data, based mainly upon subjects suffering from a striatocapsular ischemic infarction or basal ganglia hemorrhage, do not yet provide unequivocal topographic evidence in these regards. These lesions nearly always encroached upon the adjacent corticobulbar fiber tracts at the level of the corona radiata or the internal capsule, especially its anterior limb. As a consequence, the observed dysarthric deficits cannot be unambiguously assigned to the striatum, especially in the absence of a detailed auditory–perceptual evaluation of spoken language production (see, e.g., Bhatia & Marsden, 1994; as well as Weiller et al., 1993). Notably, however, some patients showed severe and persistent dysarthria even following unilateral damage to subcortical structures.

So far, only few data on the articulatory and phonatory dysfunctions in subjects with lesions confined to the basal ganglia have been reported. Whereas a patient suffering from an isolated lesion restricted to the left caudate nucleus exhibited severe speech motor deficits, as part of a choreatic syndrome (Hesselink et al., 1987), other studies point at a rather limited contribution of this component of the basal ganglia to speech motor control (e.g., Caplan, 1990), in line with the suggestion that the caudate nucleus is embedded into the cognitive, rather than the motor cortico–striato–thalamo–cortical loop. Based upon a group of patients with capsulostriatal infarctions, Mega and Alexander (1994) assumed hypophonic and dysprosodic speech in these instances to reflect disruption of the basal ganglia motor circuit, especially, at the level of the putamen. Some clinical evidence for a participation of the left pallidum in motor aspects of speech production derives from the observation of dysarthric deficits, however, of a mostly transient nature, subsequent to hemorrhagic lesions of this basal ganglia component (Alexander & LoVerme, 1980). The reported profile of speech motor signs, i.e., "hypophonic" and "mumbling" verbal utterances, closely resembles the syndrome of hypokinetic dysarthria as observed in PD. A similar constellation has also been noted in a case of bilateral ischemic damage to the thalamus (Ackermann, Ziegler et al., 1993). Most presumably, the infarction encroached upon target areas of the efferent pallidal projections, giving rise, thereby, to a disruption of the cortico–striato–thalamo–cortical motor loop.

8.4.3 Ataxic dysarthria and the role of the cerebro–cerebellar loop Speech motor deficits have been observed in a variety of cerebellar disorders, e.g., gunshot injuries, hereditary or sporadic degenerative diseases, and ischemic infarctions, predominantly within the territory of the superior cerebellar artery (for a review see Ackermann & Hertrich, 2000). The classical description of cerebellar dysarthria emphasizes irregular articulatory breakdown, harsh voice quality, reduced speaking rate, and "scanning speech" as salient perceptual characteristics (ataxic dysarthria). Relating articulatory and phonatory abnormalities in olivo-ponto-cerebellar atrophy to the distribution of infratentorial glucose metabolism, fluctuations and irregularities of speaking rate, pitch, loudness, and voice quality have been emphasized as the core signs of ataxic dysarthria (Kluin et al., 1988).

Especially in children, surgical resection of cerebellar midline tumors may give rise to transient mutism, arising, usually, within several days after intervention and resolving across a time interval of up to a few months. More rarely, similar constellations have been observed in adults or in other pathological conditions, e.g., occlusion of the basilar artery (Nishikawa et al., 1998). Since transient mutism does not necessarily develop into dysarthric deficits during further follow-up, this condition, most probably, does not exclusively reflect impaired motor control mechanisms (Ozimek et al., 2004; see Ackermann et al., 2007, for a recent review).

So far, discrepant data are available with respect to the topographic basis of cerebellar dysarthria. Whereas an early investigation reported a higher prevalence of articulatory and phonatory deficits in association with damage to the left hemisphere (Lechtenberg & Gilman, 1978), subsequent studies found a predominance of contralateral lesions (Ackermann et al., 1992; Urban et al., 2006). Furthermore, ataxic dysarthria appears to be predominantly bound to ischemia within the area of blood supply of the superior cerebellar artery (for a review, see Ackermann & Hertrich, 2000). Nevertheless, speech motor deficits have also been noted in lesions of more inferior portions of the cerebellum, though less frequently. As a consequence, the "localizing value" of cerebellar dysarthria is still a matter of controversy.

Besides the cerebellum proper, damage both to its afferent and efferent fiber tracts may compromise speech motor control mechanisms. During the early stages of the disease, the articulatory and phonatory dysfunctions associated with Friedreich ataxia reflect, presumably, a disruption of afferent pathways conveying sensory information arising in the vocal tract to the cerebellum (see, e.g., Ackermann & Hertrich, 2000). In addition, ataxic dysarthria has been observed in a patient suffering from an ischemic meso-diencephalic lesion, obviously encroaching upon efferent projections of the cerebellum to thalamic target nuclei (von Cramon, 1981).

9 Cerebral Networks of Speech Motor Control: Functional Hemodynamic Imaging Studies

9.1 Physiological background of positron emission tomography (PET) and functional magnetic resonance imaging (fMRI)

Based upon animal experimentation, an "automatic" increase of regional cerebral blood flow (rCBF) in response to local variations of neural activity was first suggested in the late nineteenth century ("neurovascular coupling"; for references, see Ackermann, Riecker et al., 2004). Thus, registration of hemodynamic changes should provide a feasible means for the identification of cerebral structures engaged in distinct sensorimotor or cognitive tasks.

Early approaches such as the Xenon clearance technique used radioactive inert gases as rCBF tracers. A series of studies considered "automatic speech," i.e., highly overlearned word strings such as the names of the months of the year, as an experimental paradigm of speech production (see Ackermann, Wildgruber et al., 1997, for a review). Because of limited spatial resolution, these procedures provided only coarse topographic information. At the present time, PET and fMRI represent the two most important brain imaging techniques based on neurovascular coupling mechanisms. Alternatives such as near-infrared spectroscopy (NIRS) have not yet been applied to speech motor control issues. PET makes use of the unique radioactive decay characteristics of positrons, i.e., positively charged particles given off by the nucleus of unstable atoms such as ^{15}O. Emitted positrons lose their kinetic energy after traveling just a few millimeters in brain tissue and ultimately are attracted to the negative charge of electrons. Annihilation of these two particles creates two very powerful photons that leave the respective area in exactly opposite directions. Because of their high energy, the photons easily exit the skull at the speed of light. During PET scanning, the subject's head is placed within a corona of radiation detectors, electronically coupled via so-called coincidence circuits. Following the injection of a small amount of ^{15}O-labeled water, the radioactive substance accumulates at the level of the cerebral cortex in direct proportion to local blood flow and, thus, local neural activity. If two detectors record a photon simultaneously, i.e., within a specified time interval, the annihilation event must have occurred on the respective connecting line. These collisions are counted and the data converted into an image of cerebral blood flow during the one-minute interval following injection. Besides rCBF, metabolic processes such as glucose consumption also show a strong correlation with the degree of neural activity. Using a feasible biological probe, PET also allows for the calculation of local cerebral metabolic rates for glucose (lCMRGlc) within distinct brain structures. Because of a rather long delay between administration of the tracer and subsequent measurements, extending across tens of minutes, this method is of limited relevance for the study of speech motor control in normal subjects, but has been applied, e.g., to correlate the distribution of hypometabolic areas with speech motor deficits in patients suffering from olivo-ponto-cerebellar atrophy (Kluin et al., 1988).

By contrast to PET, the more recent fMRI technology represents a completely noninvasive procedure, based on the detection of endogenous tissue contrasts. Furthermore, this technique offers superior spatial resolution approximating that of anatomical MR imaging. Deoxygenated hemoglobin acts as a paramagnetic agent that compromises the T2*-weighted MRI signal. Now, neural activity within circumscribed brain areas gives rise to a transient local increase in rCBF and/or blood volume. The enhanced oxygen supply surpasses the respective metabolic demands. Fast MRI acquisition procedures allow for the monitoring of so-called magnetic susceptibility effects, reflecting the shift in the balance between paramagnetic deoxyhemoglobin and its diamagnetic variant oxyhemoglobin. Thus, neural activity can be detected indirectly as an increase in the local T2*-weighted MRI signals (blood oxygen level dependent (BOLD) effect: more diamagnetic oxyhemoglobin = less paramagnetic deoxyhemoglobin = less distortion of T2*-weighted magnetic resonance signals). Besides hemodynamic PET measurements, fMRI thus also allows for the detection of local changes in cerebral blood flow, serving as an indirect parameter of neural activity. However, articulatory gestures during spoken language production may give rise to motion-induced BOLD signal changes that, eventually, confound the effects of experimental tasks. Some early studies considered silent speech as a surrogate of overt verbal utterances, assuming both behaviors to be bound to overlapping networks of cerebral structures (e.g., Wildgruber et al., 1996). As an alternative, several procedures have been introduced that may disentangle speech-related movement artifacts and task-induced hemodynamic responses. For example, verbal utterances can be restricted to pauses of image acquisition, taking advantage of the physiological delay of the hemodynamic response to a given task (Gracco et al., 2005).

9.2 Cerebral correlates of the phonetic/articulatory stage of speech production

The first systematic account of the cerebral circuitry underlying speech motor control emerged as a byproduct of a PET investigation of lexical aspects of single-word processing (Petersen et al., 1989). Subjects were confronted with a set of tasks of increasing complexity: (a) fixation of a sign appearing on a screen, (b) passive exposure to written or spoken English nouns, (c) spoken repetition of auditorily or visually applied items, and (d) generation of a verb semantically related to the presented noun (hierarchical subtraction design). It was assumed that subtraction of the hemodynamic responses to the second stage of performance, i.e., passive viewing of or listening to words, from the rCBF effects bound to the third level, i.e., loud repetition of the nouns, should isolate the brain areas related to motor aspects of speech production. Besides activation of SMA, although of a rather faint degree, bilateral responses of sensorimotor cortex and anterior–superior portions of the cerebellum could be noted. Furthermore, and quite unexpectedly, an activation spot "buried" in the depth of the lateral sulcus emerged, whereas, by contrast, both Broca's area and basal ganglia did not show any significant hemodynamic effects. In line with several sporadic observations (for references see Ackermann, Riecker et al., 2004), a subsequent PET study based upon the repetition of auditorily applied nouns (versus stimulus anticipation) was able to assign the intrasylvian response to the anterior insula (Wise et al., 1999; significant effects restricted to the left side).

More than 80 functional imaging studies addressing single-word processing (picture naming, verb/noun generation, word/pseudoword reading and listening) have been published so far. Using these data, a recent comprehensive meta-analysis tried to delineate the cerebral correlates of several distinct stages of speech production: lexical selection, phonological code retrieval, syllabification, phonetic/articulatory processes, and self-monitoring (Indefrey & Levelt, 2004). As a basic assumption, this review suggested overt tasks to engage both syllabification and phonetic/articulatory operations, whereas silent verbal performance must be expected, more or less, to spare the latter stage. The contrast of overt (aloud) and covert (silent) conditions thus should allow for the delineation of motor aspects of speech production "downstream" to language formulation.

Comparison of the two modes of single-word production resulted in the identification of altogether 17 cerebral structures associated with phonetic/articulatory operations, 12 of them pertaining to the central-motor system, e.g., ventral parts of sensorimotor cortex, dorsolateral premotor areas, SMA, thalamus, cerebellum, and the midbrain. The remaining five brain regions, among others, right-hemisphere posterior IFG and occipital areas, lack a direct relationship to movement control. Rather than phonetic/articulatory processing, most notably intrasylvian cortex was found to serve phonological code retrieval and Broca's area to mediate syllabification processes. At the lateral surface of the left hemisphere, the cortical area linked to motor aspects of speech production roughly encompasses the lateral third of sensorimotor cortex, bounded by the pre- and postcentral sulci as well as the Sylvian fissure, but sparing posterior IFG. As a consequence, this review failed to differentiate between primary motor and premotor/opercular components of speech motor control, a functional compartmentalization indicated both by clinical data and direct electrical cortex stimulation. Furthermore, these data are at some variance with a probabilistic topographic description of the "mouth region" or, to be more specific, the primary sensorimotor representation of all muscles engaged in speech production, based upon a series of functional imaging studies (overt speech tasks such as reading/repetition of single words, counting etc.; Fox et al., 2001). The observed activation spots, extending within both hemispheres across a distance of about 2 cm along the three Talairach coordinates, were strictly separated from hemodynamic responses arising within Broca's area (BA 44) and the opercular portion of precentral gyrus (BA 6; see Figure 6.2a).

In order to obtain a more refined model of the compartmentalization of posterior IFG and lower precentral convolution, a recent meta-analysis compiled 43 functional imaging studies with a focus on the phonological level of speech production (Vigneau et al., 2006). As a result, the lower frontal convexity was assumed to encompass three distinct levels of phonetic/articulatory processing, i.e., "an upper motor area for mouth motion control, a lower premotor area in the precentral gyrus that is dedicated to pharynx and tongue fine-movement coordination, and a sensory-motor integration region in the Rolandic operculum" (Vigneau et al., 2006, p. 1419). However, visual inspection of the displayed maps reveals a rather continuous band of activation spots, and the observed hemodynamic responses within the lateral convexity of the frontal lobe do not segregate into three distinct clusters. Whereas precentral activation, indeed, encroaches upon the primary motor area of vocal tract musculature (see for comparison the probabilistic description of the "mouth area" in Fox et al., 2001), the separation of a "lower premotor pharynx and tongue region" subserving "fine-movement coordination" from a "sensory-motor integration region in the Rolandic operculum" is not borne out by the presented data. It remains obscure, furthermore, how far "fine-movement coordination" and "sensory-motor integration" differ from each other in terms of motor control processes.

The meta-analyses referred to are based upon studies not specifically designed for the investigation of the cerebral correlates of the articulatory/phonetic stage of spoken language. As an alternative, more focused functional imaging studies can be expected to provide a more fine-grained display of the cerebral organization of speech motor control. A widespread and mostly bilateral pattern of activation spots was found to be associated with the production of the isolated vowel /ah/, including precentral gyrus, mesiofrontal areas, posterior insula, superior temporal lobe, thalamus, basal ganglia, red nucleus, and the cerebellum (Sörös et al., 2006). Since this task had been contrasted with a rest condition, a variety of speech-unspecific responses must be expected, besides the cerebral correlates of articulatory/phonatory functions. For example, the authors refer to "self-awareness of movement" as an explanation of the observed bilateral recruitment of the posterior insula. Furthermore, there is some evidence that rest conditions may represent a specific "activation state", confounding the effects of experimental tasks (e.g., Raichle et al., 2001). Comparison of CV utterances (/pa/, /ta/, or /ka/) and the polysyllabic item /pataka/ with the vowel task revealed activation spots within the temporal lobes (CV and /pataka/) as well as hemodynamic responses of the basal ganglia and the cerebellum (/pataka/; Sörös et al., 2006). Whereas the rather extensive responses of the temporal gyri can be explained in terms of central-auditory and phonological processing, the additional subcortical BOLD signal changes might reflect enhanced demands upon the control of vocal tract muscles in association with syllable sequencing (see below). Unexpectedly, production of the polysyllabic item elicited neither activation of premotor nor of the anterior insular cortex of the left hemisphere, though damage to these regions may give rise to severe disruption specifically of the polysyllabic variant of syllable repetition tasks.

9.3 Functional organization of the cerebral network of speech motor control: Some insights from functional imaging

9.3.1 The impact of syllable sequencing and syllable complexity upon hemodynamic activation patterns during verbal tasks The available clinical data do not yet provide a coherent picture of the cerebral correlates of verbal apraxia (see above). Since this syndrome is assumed to reflect a deficit in the "programming" of vocal tract movements and since the demands on "motor planning" must be expected to increase with utterance length, a recent fMRI study suggested that hemodynamic activation of the insula covaries with this parameter of spoken language (Shuster & Lemieux, 2005). Indeed, overt (versus silent) repetition of mono- and multisyllabic nouns was found to be associated with activation spots arising at the floor of the Sylvian fissure. Whereas, however, the longer items yielded significantly enhanced hemodynamic responses of the inferior parietal lobule, the precentral convolution (BA 6), and posterior parts of IFG (BA 44) of the left hemisphere, a comparable effect of intrasylvian structures did not emerge. Although both the dorsolateral convexity of the frontal lobe and the insular cortex in the depth of the lateral sulcus seem to pertain to the articulatory/phonetic network of speech production, these structures might subserve different control mechanisms.

The phonotactic rules of, e.g., the German or the English language allow for a variety of syllable onset structures (V, CV, CCV). Besides utterance length, in terms of the number of enclosed syllables, consonant clusters pose presumably higher demands on articulatory/phonetic control mechanisms as compared to CV units (Aichert & Ziegler, 2004). A recent elaborate fMRI study used four tri-syllabic items (/ta-ta-ta/, /ka-ru-ti/, /stra-stra-stra/, /kla-stri-splu/), systematically varied in sequence complexity (the same three versus three different items in a row) and syllabic complexity (CV versus CCCV onset) as test materials (Bohland & Guenther, 2006). Since the experimental design included both "go" and "no go" trials, overt task performance could be contrasted with a state of being prepared to produce the same items. The "minimal" cerebral network of overt speech production, in terms of the conjunction of hemodynamic activation across the four test materials considered (versus baseline), was found to encompass, among other things, the post- and precentral gyrus, encroaching upon posterior IFG, the anterior insula, superior temporal cortex, SMA, the basal ganglia, thalamic areas, and the superior cerebellar hemispheres. Significant left-lateralization effects emerged at the level of intrasylvian cortex, by contrast, posterior IFG and, at variance with a preceding study (Riecker et al., 2000b), sensorimotor cortex failed to display comparable side-differences of hemodynamic activation. The contrast of "go" and "no go" trials revealed, by and large, the same response pattern. An increase of stimulus complexity in either dimension yielded, as expected, enhanced activation of at least some components of the "basic speech network." Since CC- and CCC-consonant clusters, as a rule, encompass more articulatory gestures than CV syllables, enhanced demands upon movement execution mechanisms must be expected. Under these conditions, a recruitment of further cortical areas such as parietal structures and additional left-lateralization effects within ventral premotor cortex could be observed. As an explanation, the low-frequency, hyper-complex syllables of the CCC-condition might pose additional demands on motor programming, giving rise to lateralized activation of parietal (sensory) and inferior-frontal (motor) components of the speech planning system.

9.3.2 Contribution of the basal ganglia and the cerebellum to syllable rate control

Hemodynamic rate–response functions As compared to the production of lexical items or pseudo-words, syllable repetitions represent a more direct probe of articulatory performance, lacking any significant impact of perceptual processes, grapheme–phoneme transformation, lexical or syntactic operations. Furthermore, this task performed as fast as possible (oral diadochokinesis) has been claimed to represent a sensitive and, within some limits, a specific overall measure of dysarthric deficits (e.g., Kent et al., 1987). For example, a reduced maximum syllable repetition rate has been observed in patients with spastic or ataxic dysarthria. By contrast, subgroups of patients with Wilson or Parkinson disease may even exhibit "speech hastening," i.e., involuntary acceleration of speaking rate, while speech tempo, otherwise, is found largely unimpaired in these disorders (Duffy, 2005). In order to further elucidate the differential contribution of the various components of the cerebral speech motor network to rate control, a fMRI study of our group measured hemodynamic brain activation during syllable repetitions at different rates (2.0, 2.5, 3.0, 4.0, 5.0, 6.0 Hz), synchronized either to an auditorily applied pacing signal or produced in a self-paced manner (Riecker et al., 2005, 2006). Significant hemodynamic main effects, calculated across all repetition rates (versus passive listening to the acoustic pacing signals), emerged within SMA, precentral areas, dorsolateral frontal cortex, including Broca's region, anterior insula, thalamus, basal ganglia, and cerebellum. Dorsolateral frontal and intrasylvian cortex as well as the caudate nucleus showed lateralized responses in favor of the left side, whereas the other components displayed a rather bilateral activation pattern (Figure 6.7).

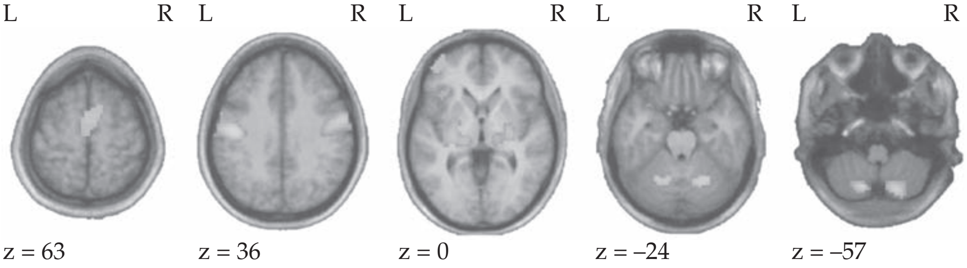


Figure 6.7 Hemodynamic main effects during syllable repetitions (group averages) computed across six frequency conditions (gray spots), displayed on transverse sections of the averaged anatomic reference images (z = distance to the intercommisural plane; L = left hemisphere, R = right hemisphere). Significant responses emerged within SMA (first scan from left), bilateral sensorimotor cortex (second scan), bilateral basal ganglia, left anterior insula, left inferior frontal gyrus (third scan), superior (fourth scan) and inferior parts (fifth scan) of both cerebellar hemispheres.

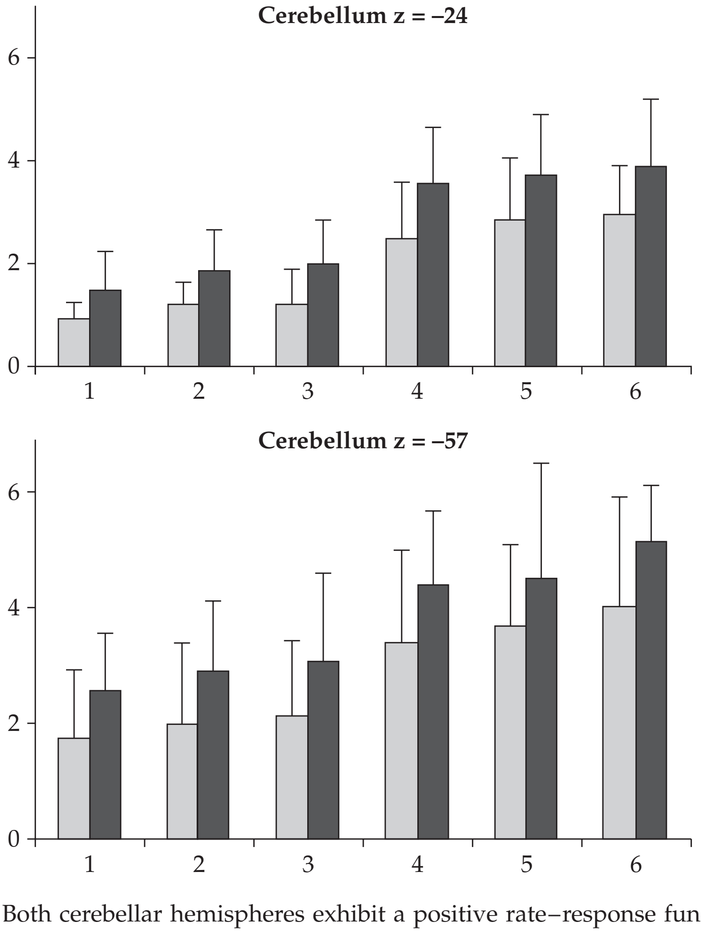


Figure 6.8 Both cerebellar hemispheres exhibit a positive rate–response function, characterized, most notably, by a step-wise increase of the hemodynamic response at about 3 Hz. (From Riecker et al., 2005)

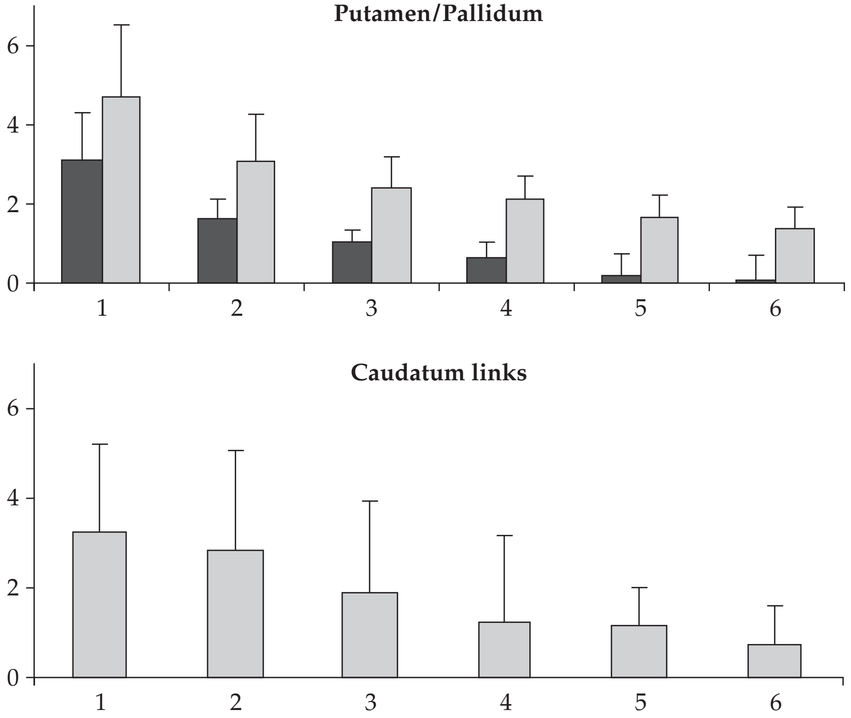


Figure 6.9 Group averages of hemodanymic activation across subjects, calculated separately for the six stimulus rates each. Note, no. 1 refers to 2 Hz, 2 = 2.5 Hz, etc. (dark columns = right hemisphere, light columns = left hemisphere). Bilateral putamen/ pallidum and left caudate nucleus show a negative rate–response relationship. (From Riecker et al., 2005)

Since damage to these cerebral structures compromises verbal behavior, giving rise to dysarthria, apraxia of speech, or transcortical motor aphasia (see above), these findings are in good accord with clinical data. As a second step of signal analysis, hemodynamic rate–response functions were calculated. SMA 6, sensorimotor cortex, anterior insula, and the cerebellar activation spots showed a positive linear rate–response relationship, i.e., the BOLD signal increased in parallel with syllable repetition rate. A variety of functional imaging studies had revealed mass activation effects, i.e., a parallel increase of BOLD response and motor demands, within the cortical hand representation area and SMA 6 during finger tapping tasks and joystick movements (for references, see Riecker et al., 2005). These data have been assumed to reflect a tight relationship between neuronal activity and movement velocity as documented, e.g., by single-cell recordings within monkey motor cortex. Quite conceivably, the same mechanisms are engaged in oral diadochokinesis tasks at the level of frontal and intrasylvian cortex.

In accordance with a previous investigation based upon silent (covert) syllable repetitions (Wildgruber et al., 2001), the cerebellar activation spots at either side showed, though less pronounced, a step-wise increase of the BOLD signal between 3 and 4 Hz (Figure 6.8). Most notably a series of preceding acoustic studies of our group had revealed that syllable rate did not fall below a value of 3 Hz in patients with ataxic dysarthria during oral diadochokinesis and sentence production tasks (Hertrich & Ackermann, 1997b). Taken together, these clinical and functional imaging data appear to indicate that the cerebellum "pushes" speaking rate beyond a level of about 3 Hz. Syllable repetitions and connected speech, indeed, may differ, at least to some extent, in their underlying motor control mechanisms. (see, e.g., Ziegler, 2002, for a further discussion). Hence, any inferences from brain activation data based upon syllable repetition paradigms to sentence production must be considered with some caution.

As a novel finding, parametric signal analysis revealed a negative linear relationship between syllable rate and hemodynamic response within bilateral putamen/pallidum as well as left caudate nucleus (Figure 6.9). In accordance with these data, recent PET studies reported an inverse relationship between a volume–mean normalized measure of rCBF and syllable frequency at the level of right caudate nucleus both in normal speakers and ataxic subjects (Sidtis et al., 2003, 2006). On a global scale, therefore, these subcortical structures seem to be characterized by a decline of hemodynamic activation in response to an increase of motor demands, at least during repetitive movements. Conceivably, the observed negative rate–response profiles reflect a more efficient organization of higher-frequency movements at the level of the basal ganglia. These suggestions could explain why PD, as a rule, fails to disrupt oral diadochokinesis tasks, in contrast to other central-motor disorders, and why this disorder eventually may even give rise to a "hastening phenomenon," i.e., involuntary acceleration of speech tempo (Ackermann, Gröne et al., 1993; Ackermann, Konczak et al., 1997).

Preparation/initiation and movement execution during syllable repetition Besides the calculation of rate–response relationships, the time series of the BOLD signal across syllable repetitions derived from the various components of the cerebral network of speech motor control were compared to each other ("functional connectivity analysis"; for further details see Riecker et al., 2005). Most notably, these areas segregated into two clusters (see Figure 6.10): left SMA, Broca's area, left anterior insula, and right cerebellar hemisphere, on the one hand, left SMC, left thalamus and basal ganglia, and right cerebellum, on the other hand, were found to be interconnected each by high (0.75–0.9) and very high (> 0.9) correlation coefficients).

To further characterize the temporal behavior of the functionally associated brain areas, a voxel-wise comparison of the measured BOLD signal changes

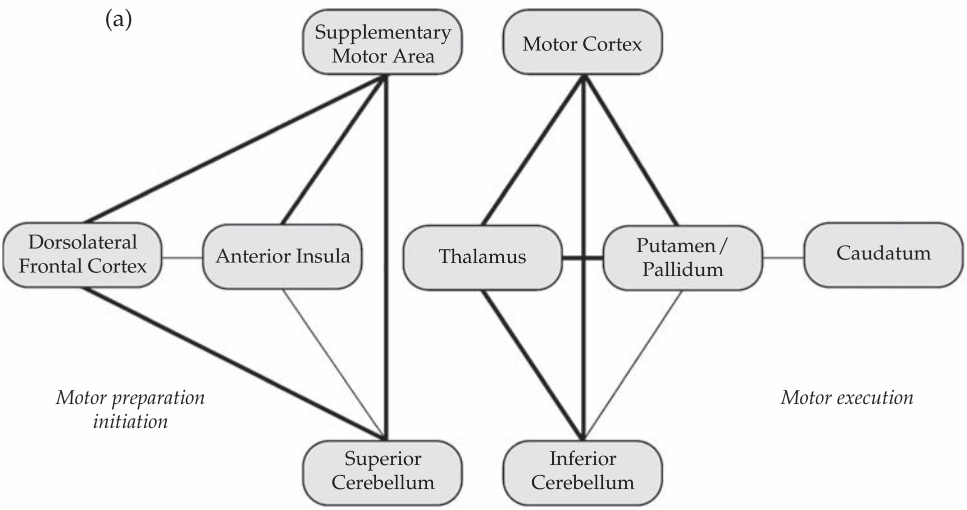


Figure 6.10 Quantitative functional connectivity analyses: computed correlation coefficients across the time series of the BOLD signal within the volumes of interest considered, i.e., the areas of a hemodynamic main effect (bold lines = correlation coefficient > 0.9, thin lines = 0.75–0.9, low and intermediate correlations not depicted).

with the canonical hemodynamic response function was conducted. SMC of either hemisphere displayed the expected time course of hemodynamic activation. As a rule, the onset of cerebral hemodynamic responses to a perceptual, motor or cognitive task is delayed by about two seconds, and the decline of the elevated BOLD signal towards rest level starts at a similar interval after offset of stimulation or behavioral performance (Figure 6.11). Furthermore, visual analysis revealed a similar temporal pattern within the basal ganglia (left putamen/ pallidum and left caudatum), left thalamus as well as inferior aspects of the cerebellar hemispheres. Thus, this network seems to be engaged in motor execution processes in terms of either the generation of motor control signals steering the "speech apparatus" or in the transmission and processing of reafferent input.

In contrast to this "motor execution loop," the hemodynamic responses of the various components of the other cluster were characterized by an earlier onset and a shorter duration. Tentatively, this second network might be linked to preparatory activities or movement initiation mechanisms ("preparation/initiation network"). Based upon these data, the two separate cerebellar activation spots appear to participate in different aspects of speech motor control. These findings are in accord with neuroanatomic tracer studies which revealed the cerebro-cerebellar pathways to be organized into several parallel closed loops (Kelly & Strick, 2003).

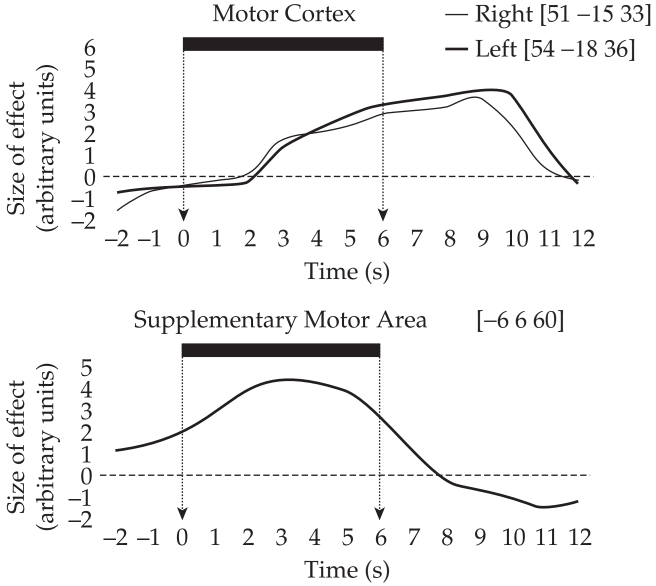


Figure 6.11 Upper panel: At the level of the motor cortex (bold line = left hemisphere, regular line = right hemisphere), the hemodynamic response shows a characteristic delay of 2 to 3 seconds with respect to task onset (horizontal bar = duration of the syllable synchronization task, extending across 6 seconds), then develops in a near-plateau, and finally extends beyond the offset of syllable repetitions for several seconds, i.e., the well-established and characteristic temporal pattern of hemodynamic activation in response to a motor task or to sensory stimulation. At the level of the motor cortex, it can be expected that neural activity starts some milliseconds in advance of the onset of syllable repetitions. Lower panel: Most notably, a different pattern can be observed at the level of SMA, indicating an engagement of this area in preparatory activities or in movement initiation. (From Riecker et al., 2005)

9.4 *Covert generation of verbal utterances (“inner speech”)*

Internal speech has been considered a pre-articulatory, but otherwise fully parsed speech code (Levelt et al., 1999). As a consequence, the investigation of covert verbal utterances might provide a “window” into preparatory activities preceding actual performance of movement sequences and allow for the delineation of the cerebral correlates of motor programming or planning processes. These suggestions must, however, be considered with some caution, since inner speech can be associated with subliminal activity of the tongue musculature (Sokolov, 1972).

In order to delineate the cerebral network mediating covert or imagined verbal utterances, an earlier fMRI study asked subjects to produce automatic speech, i.e., the ongoing recitation of the months of the year. By contrast to the production of single nouns, this task provides an opportunity to address fluent or connected verbal output, in the absence of any relevant demands on lexical retrieval or syntactic processing. And since automatic speech can be expected to elicit a rather monotonous mode of speech production, no significant confounding impact of intonational suprasegmental patterns must be expected. As a control condition, participants reproduced a nonlyrical tune drawn from a serenade in order to avoid any retrieval of propositional materials. Whereas covert speech was found to elicit exclusive activation of left-hemisphere precentral gyrus as well as the contralateral cerebellar structures silent singing resulted in an opposite response pattern encompassing right premotor cortex and superior aspects of the left cerebellar hemisphere (Ackermann et al., 1998; Riecker et al., 2000a).

High-frequency syllables, by definition, pertain to the most exercised motor activities and, therefore, should represent highly overlearned movement patterns. These syllable-sized articulatory programs ("mental syllabary") are assumed to be stored within the premotor cortex of the language-dominant hemisphere **(**Levelt, 2001). Conceivably, the observed hemodynamic response of left-sided precentral areas during inner speech reflects activation of the assumed mental syllabary. Based upon this model, word forms comprising two or more syllables or phrases, extending across several lexical items, are not stored as prespecified motor routines, and the sequencing of syllabic units into larger utterances thus must be performed online during speech. Besides left-hemisphere premotor cortex, "inner speech" was found to be associated with significant hemodynamic activation of the contralateral cerebellar hemisphere. Acoustic and kinematic analyses of ataxic dysarthria suggest the cerebellum provides a platform for the sequencing of syllables and, thus, subserves the concatenation of the retrieved syllable templates into smooth, i.e., coarticulated utterances at a speaker's habitual speech tempo (for a review see Ackermann, Mathiak et al., 2004). As a consequence, the cerebellum might engage into the temporal syllabic organization of spoken verbal utterances even at a pre-articulatory level.

A recent fMRI study found silent single-word repetition to be associated with a more widespread network of cerebral structures, including right-hemisphere areas, even in contrast to overt production of the same items (Shuster & Lemieux, 2005). Presumably, the much more restricted activation pattern during automatic speech reflects the simplicity of this task, lacking, e.g., significant demands upon lexical retrieval or prosodic modulation.

9.5 Comparison of verbal and nonverbal orofacial movements

Early functional imaging studies of word processing had used a compound subtraction procedure in order to separate cognitive (lexical retrieval) and motor aspects (articulation/phonation) of spoken language production (e.g., Petersen et al., 1989). As a probe of the validity of this approach, a more recent PET study applied the principles of "decomposition logic" on speech motor control itself (Sidtis et al., 1999). On these grounds, syllable repetitions might be conceived of as the "sum" of lip closure movements and laryngeal activity. Thus, the activation pattern bound to phonation tasks (minus baseline) must be expected, e.g., to equal the difference in hemodynamic responses to lip gestures and syllable repetitions. Besides a baseline condition, i.e., subjects being quiet and awake, three activation tasks were considered for analysis: (a) iteration of the syllables /pa/, /ta/, and /ka/ as fast as possible, (b) sustained production of the vowel /a/, and (c) repetitive lip closures. The PET measurements failed to document any task additivity under these conditions. Most notably, the allegedly less demanding tasks of sustained phonation and lip closure movements yielded more spacious activation spots within several regions than the ostensibly higher motor control demands linked to syllable repetitions. Two fMRI studies also found nonspeech orofacial movements to yield more extensive hemodynamic responses than linguistic test materials (Wildgruber et al., 1996: vertical nonspeech tongue movements, the lips being closed, versus highly overlearned word strings; Riecker et al., 2000b: horizontal tongue movements, the mouth open, versus reiteration of nonsense test materials differing in articulatory/phonetic complexity). Presumably, the enlarged activation foci during horizontal tongue movements might reflect the increased effort bound to nonspeech orofacial tasks as compared to coarticulated mono- and polysyllabic items (“movement fractionation” versus “nonindividualized” motor control; see Riecker et al., 2000b).

9.6 Summary

As a rule, functional imaging techniques found hemodynamic activation of SMA, ventral premotor and intrasylvian areas, primary sensorimotor cortex, and subcortical central-motor structures (basal ganglia, thalamus, cerebellum) during the production/repetition of, e.g., nonlexical mono- or polysyllabic items. These findings are in line with clinical data, since dysfunctions of these structures are known to impair higher-order processes of speech motor control (movement planning or initiation) as well as the execution of articulatory/phonatory vocal tract functions. Beyond topographic information, furthermore, imaging techniques begin to provide insights into the functional organization of the various components of the cerebral network of speech motor control such as different rate–response relationships or a dual role of the cerebellum within this domain.

10 Conclusions

10.1 Cerebral network of speech motor control

Disorders of the sensorimotor cortex, the basal ganglia, and the cerebellum, including the respective afferent and/or efferent fiber tracts, are characterized by distinct central-motor deficits that compromise, besides other domains, the innervation of vocal tract muscles during speech production, giving rise to dysarthric deficits. By contrast, damage to SMA of the language-dominant hemisphere

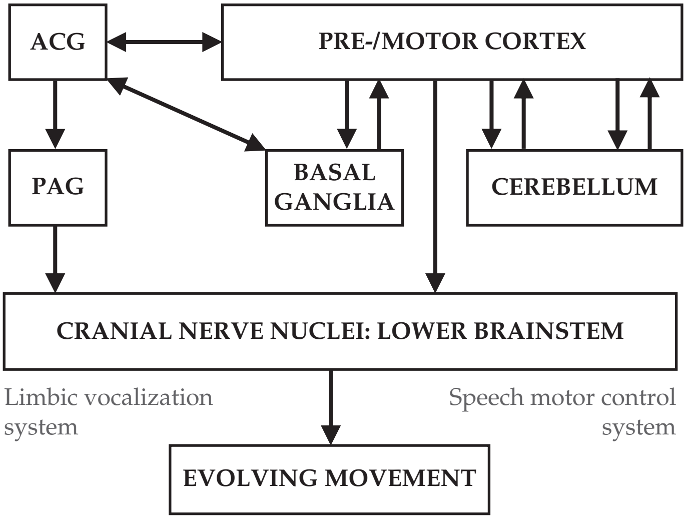

ACG
PRE-/MOTOR CORTEX
PAG
BASAL GANGLIA
CEREBELLUM
CRANIAL NERVE NUCLEI: LOWER BRAINSTEM
Limbic vocalization system
Speech motor control system
EVOLVING MOVEMENT

Figure 6.12 Dual-pathway model of the cerebral network subserving acoustic communication in humans: this schematic display aims at an integration of clinical data, findings from electrical cortical stimulation, and functional imaging studies.

spares articulatory and phonatory functions, but has been observed to result in a significant decrease of speech production. In consideration of the available electrophysiological data, this constellation presumably reflects a disruption of speech initiation mechanisms. Finally, lesions of left-hemisphere rostral perisylvian cortex (premotor areas of the precentral and the inferior frontal gyrus) and/or ipsilateral anterior insula may give rise to verbal apraxia, a syndrome assumed to reflect a dysfunction of higher-order levels of speech motor control such as the compilation of a phonetic plan.

In line with these clinical data, functional imaging allowed for a separation of the various cerebral components of speech motor control into two subsystems. Calculation of the time course of hemodynamic activation in response to syllable repetitions indicates SMA, premotor, and intrasylvian cortex as well as superior aspects of the cerebellar hemispheres to be engaged in pre-articulatory activities of speech production (“preparatory loop”), whereas, by contrast, sensorimotor cortex, basal ganglia, thalamus, and the inferior cerebellum exclusively seem to participate in movement execution, i.e., the online control of vocal tract movement patterns (“executive loop”). Furthermore, preliminary evidence suggests the superior parts of the cerebellum to contribute to the sequencing of syllable strings during generation of a pre-articulatory verbal code (“inner speech”). Though the available data so far do not allow for an unambiguous differentiation of cerebral processing stages, the clinical, electrophysiological, and functional imaging findings, taken together, suggest the brain network of speech motor control to decompose into at least three functional subsystems:

1 Starting mechanisms of speech production, initiating and maintaining an ongoing and fluent verbal stream, seem to depend upon mesiofrontal SMA of the language-dominant hemisphere.

2 Premotor areas of the precentral and inferior frontal gyrus of the left hemisphere (rostral perisylvian or frontal opercular cortex, respectively), eventually including the ipsilateral anterior insula, participate in the construction of the phonetic makeup of an utterance prior to innervation of vocal tract musculature. Furthermore, these structures appear to operate in concert with superior aspects of the cerebellum during the generation of a pre-articulatory verbal code ("inner speech").

3 Motor execution, i.e., the online innervation of respiratory, laryngeal, and supralaryngeal muscles during speech production, seems to be bound to the bilateral corticobulbar system as well as the cortico-subcortical motor loops traversing the basal ganglia and the cerebellar hemispheres.

10.2 Speech production and vocal expression of emotions (affective vocalizations and affective prosody)

Pathways arising in ACG and projecting via midbrain structures and a pontine "pattern generator" to cranial nerve nuclei have been found to mediate vocal behavior of subhuman primates. Clinical data indicate this phylogenetically old "limbic vocalization system" to be still instrumental for the emotional expression in our species. For example, patients with anarthria and/or aphonia subsequent to bilateral damage to the corticobulbar system nevertheless are often able to laugh and to cry. In spite of a paralysis of vocal tract muscles, these movement patterns emerge unimpaired within the context of emotional expression ("automatic-voluntary movement dissociation"; Mao et al. 1989). As a consequence, two "channels" subserving acoustic communication target the cranial nerve nuclei in humans: a limbic network subserving affective vocalizations, and a corticobulbar system bound to speech motor control ("dual pathway" model; Figure 6.12).

Most probably, the medial wall of the frontal lobes provides the interface between the limbic system of vocal behavior and speech motor control. First, there is evidence for reciprocal connections between ACG, the cortical control instance of vocal affective behavior, and SMA, supporting speech initiation processes. As a consequence, these mesiofrontal pathways might transform motivational drive for verbal communication into a mechanism releasing phonetic plans at precise times, maintaining ongoing verbal output. Second, ACG is assumed to represent the major cortical origin of the limbic basal ganglia loop. By contrast, the respective motor circuit arises in premotor areas of the frontal lobe. Assuming "information funneling" at the level of the basal ganglia, these subcortical nuclei could provide the platform for the integration of emotional and information sound structure during speech production, enabling the implementation of affective prosody onto verbal utterances.

NOTE

[1] According to its traditional definition, the frontal operculum (literally "frontal lid") encompasses the lateral portion of the frontal lobe covering insular cortex, i.e., the most ventral aspect of the precentral gyrus and the caudal part of IFG. It should be noted, however, that often – as in this chapter – this term is restricted to the opercular, i.e., most posterior, component of the third frontal convolution.

REFERENCES

Abbs, J. H. & Cole, K. J. (1982) Consideration of bulbar and suprabulbar afferent influences upon speech motor coordination and programming. In S. Grillner, B. Lindblom, J. Lubker, & A. Persson (eds.), *Speech Motor Control* (pp. 159–86). Oxford: Pergamon Press.

Ackermann, H. (2006) Neurobiologische Grundlagen des Sprechens [Neurobiological bases of speech motor control]. In H. O. Karnath & P. Thier (eds.), *Neuropsychologie*, 2nd edn. (pp. 333–9). Heidelberg: Springer.

Ackermann, H., Daum, I., Schugens, M. M., & Grodd, W. (1996) Impaired procedural learning following damage to the left supplementary motor area (SMA). *Journal of Neurology, Neurosurgery, and Psychiatry*, 60, 94–7.

Ackermann, H., Gröne, B. F., Hoch, G., & Schönle, P. W. (1993) Speech freezing in Parkinson's disease: A kinematic analysis of orofacial movements by means of electromagnetic articulography. *Folia Phoniatrica*, 45, 84–9.

Ackermann, H. & Hertrich, I. (2000) The contribution of the cerebellum to speech processing. *Journal of Neurolinguistics*, 13, 95–116.

Ackermann, H., Hertrich, I., Daum, I., Scharf, G., & Spieker, S. (1997) Kinematic analysis of articulatory movements in central motor disorders. *Movement Disorders*, 12, 1019–27.

Ackermann, H., Hertrich, I., Lutzenberger, W., & Mathiak, K. (2006) Zerebrale Hemisphärenlateralität der Sprachlautwahrnehmung: Klinische und funktionell-bildgebende Befunde [Functional neuroanatomy of speech sound perception: Clinical and imaging data]. *Aktuelle Neurologie*, 33, 218–31.

Ackermann, H., Hertrich, I., Ziegler, W., Bitzer, M., & Bien, S. (1996) Acquired dysfluencies following infarction of the left mesiofrontal cortex. *Aphasiology*, 10, 409–17.

Ackermann, H., Konczak, J., & Hertrich, I. (1997) The temporal control of repetitive articulatory movements in Parkinson's disease. *Brain and Language*, 56, 312–19.

Ackermann, H., Mathiak, K., & Ivry, R. B. (2004) Temporal organization of "internal speech" as a basis for cerebellar modulation of cognitive functions. *Behavioral and Cognitive Neuroscience Reviews*, 3, 14–22.

Ackermann, H., Mathiak, K., & Riecker, A. (2007) The contribution of the cerebellum to speech production and speech perception: Clinical and functional imaging data. *Cerebellum*, 6, 202–13.

Ackermann, H., Riecker, A., & Wildgruber, D. (2004) Functional brain imaging of motor aspects of speech production. In B. Maassen, R. D. Kent, H. F. M. Peters, P. H. M. M. van Lieshout, and W. Hulstijn (eds.), *Speech Motor Control* *in Normal and Disordered Speech* (pp. 85–111). Oxford: Oxford University Press.

Ackermann, H., Vogel, M., Petersen, D., & Poremba, M. (1992) Speech deficits in ischaemic cerebellar lesions. *Journal of Neurology*, 239, 223–7.

Ackermann, H., Wildgruber, D., Daum, I., & Grodd, W. (1998) Does the cerebellum contribute to cognitive aspects of speech production? A functional magnetic resonance imaging (fMRI) study in humans. *Neuroscience Letters*, 247, 187–90.

Ackermann, H., Wildgruber, D., & Grodd, W. (1997) Neuroradiologische Aktivierungsstudien zur zerebralen Organisation sprachlicher Leistungen: Eine Literaturübersicht [Neuroradiological activation studies on the cerebral organization of language capacities: A review]. *Fortschritte der Neurologie und Psychiatrie*, 65, 182–94.

Ackermann, H. & Ziegler, W. (1995) Akinetischer Mutismus: Eine Literaturübersicht [Akinetic mutism: A review]. *Fortschritte der Neurologie und Psychiatrie*, 63, 59–67.

Ackermann, H., Ziegler, W., & Petersen, D. (1993) Dysarthria in bilateral thalamic infarction: A case study. *Journal of Neurology*, 240, 357–62.

Aichert, I. & Ziegler, W. (2004) Syllable frequency and syllable structure in apraxia of speech. *Brain and Language*, 88, 148–59.

Aitken, P. G. (1981) Cortical control of conditioned and spontaneous vocal behavior in Rhesus monkeys. *Brain and Language*, 13, 171–84.

Alexander, M. P. & LoVerme, S. R. (1980) Aphasia after left hemisphere intracerebral hemorrhage. *Neurology*, 30, 1193–1202.

Alexander, M. P., Naeser, M. A., & Palumbo, C. (1990) Broca's area aphasias: Aphasia after lesions including the frontal operculum. *Neurology*, 40, 353–61.

Allman, J. M. (2000) *Evolving Brains*. New York: Scientific American Library.

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B. M., & Zilles, K. (1999) Broca's region revisited: Cytoarchitecutre and intersubject variability. *Journal of Comparative Neurology*, 412, 319–41.

Aydogdu, I., Ertekin, C., Tarlaci, S., Turman, B., Kiylioglu, N., & Secil, Y. (2001) Dysphagia in lateral medullary infarction (Wallenberg's syndrome): An acute disconnection syndrome in premotor neurons related to swallowing activity? *Stroke*, 32, 2081–7.

Barlow, S. M. & Farley, G. R. (1989) Neurophysiology of speech. In D. P. Kuehn, M. L. Lemme, and J. M. Baumgartner (eds.), *Neural Bases of Speech, Hearing, and Language* (pp. 146–200). Boston, MA: College-Hill Press.

Bhatia, K. P. & Marsden, C. D. (1994) The behavioural and motor consequences of focal lesions of the basal ganglia. *Brain*, 117, 859–76.

Bohland, J. W. & Guenther, F. H. (2006) An fMRI investigation of syllable sequence production. *NeuroImage*, 32, 821–41.

Botez, M. I. & Barbeau, A. (1971) Role of subcortical structures, and particularly of the thalamus, in the mechanisms of speech and language: A review. *International Journal of Neurology*, 8, 300–20.

Broca, P. (1861) Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletins de la Société Anatomique de Paris*, 36, 330–57.

Brooks, V. B. (1986) *The Neural Basis of Motor Control*. New York: Oxford University Press.

Butler, A. B. & Hodos, W. (2005) *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation*, 2nd edn. Hoboken, NJ: John Wiley.

Cancelliere, A. E. B. & Kertesz, A. (1990) Lesion localization in acquired deficits of emotional expression and comprehension. *Brain and Cognition*, 13, 133–47.

Cantalupo, C. & Hopkins, W. D. (2001) Asymmetric Broca's area in great apes: A region of the ape brain is uncannily similar to one linked with speech in humans. *Nature*, 414, 505.

Caplan, D. (1987) *Neurolinguistics and Linguistic Aphasiology: An Introduction*. Cambridge: Cambridge University Press.

Caplan, L. R. (1990) Caudate infarcts. *Archives of Neurology*, 47, 133–43.

Cheney, D. L. & Seyfarth, R. M. (1996) Function and intention in the calls of non-human primates. In W. G. Runciman, J. Maynard Smith, & R. I. M. Dunbar (eds.), *Evolution of Social Behaviour Patterns in Primates and Man: A Joint Discussion Meeting of the Royal Society and the British Academy* (pp. 59–76). Oxford: Oxford University Press.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975) *Motor Speech Disorders*. Philadelphia, PA: Saunders.

Deacon, T. W. (1992) The human brain. In S. Jones, R. Martin, & D. Pilbeam (eds.), *The Cambridge Encyclopedia of Human Evolution* (pp. 115–23). Cambridge: Cambridge University Press.

Deecke, L., Engel, M., Lang, W., & Kornhuber, H. H. (1986) Bereitschaftspotential preceding speech after holding breath. *Experimental Brain Research*, 65, 219–23.

Déjérine, J. (1891) Contribution à l'étude de l'aphasie motrice sous-corticale et de la localisation cérébrale des centres laryngés (muscles phonateurs). *Comptes Rendus Hebdomadaires des Séances et Mémoires de la Société de Biologie*, 43, 155–62.

DeLong, M. R. & Wichmann, T. (2007) Circuits and circuit disorders of the basal ganglia. *Archives of Neurology*, 64, 20–4.

Dronkers, N. F. (1996) A new brain region for coordinating speech articulation. *Nature*, 384, 159–61.

Duffy, J. R. (2005) *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 2nd edn., St Louis, MO: Elsevier Mosby.

Duffy, J. R. & Folger, W. N. (1996) Dysarthria associated with unilateral central nervous system lesions: A retrospective study. *Journal of Medical Speech-Language Pathology*, 4, 57–70.

Elias, S. A. (1990) Trigeminal projections to the cerebellum. In A. Taylor (ed.), *Neurophysiology of the Jaws and Teeth* (pp. 192–236). Houndmills, UK: Macmillan.

Fitch, W. T. (2000) The evolution of speech: A comparative review. *Trends in Cognitive Sciences*, 4, 258–67.

Fox, P. T., Huang, A., Parsons, L. M., et al. (2001) Location-probability profiles for the mouth region of human primary motor-sensory cortex: Model and validation. *NeuroImage*, 13, 196–209.

Freedman, M., Alexander, M. P., & Naeser, M. A. (1984) Anatomic basis of transcortical motor aphasia. *Neurology*, 34, 409–17.

Geschwind, N. (1969) Problems in the anatomical understanding of the aphasias. In A. L. Benton (ed.), *Contributions to Clinical Neuropsychology* (pp. 107–28). Chicago, IL: Aldine.

Gracco, V. L., Tremblay, P., & Pike, B. (2005) Imaging speech production using fMRI. *NeuroImage*, 26, 294–301.

Greenlee, J. D. W., Oya, H., Kawasaki, H., et al. (2004) A functional connection between inferior frontal gyrus and orofacial motor cortex in humans. *Journal of Neurophysiology*, 92, 1153–64.

Grinevich, V., Brecht, M., & P. Osten (2005) Monosynaptic pathway from rat vibrissa motor cortex to facial motor neurons revealed by lentivirus-based axonal tracing. *Journal of Neuroscience*, 25, 8250–8.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006) Neural modelling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96, 280–301.

Hage, S. R. & Jürgens, U. (2006) On the role of the pontine brainstem in vocal pattern generation: A telemetric single-unit recording study in the squirrel monkey. *Journal of Neuroscience*, 26, 7105–15.

Halpern, M. E., Güntürkün, O., Hopkins, W. D., & Rogers, L. J. (2005) Lateralization of the vertebrate brain: Taking the side of model systems. *Journal of Neuroscience*, 25, 10351–7.

Hartelius, L. & P. Svensson (1994) Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: A survey. *Folia Phoniatrica et Logopaedica*, 46, 9–17.

Hayes, T. L. & Lewis, D. A. (1995) Anatomical specialization of the anterior motor speech area. *Brain and Language*, 49, 289–308.

Hertrich, I. & Ackermann, H. (1994) Acoustic analysis of speech timing in Huntington's disease. *Brain and Language*, 47, 182–96.

Hertrich, I. & Ackermann, H. (1997a) Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures. *Journal of the Acoustical Society of America*, 102, 523–36.

Hertrich, I. & Ackermann, H. (1997b) Acoustic analysis of durational speech parameters in neurological dysarthrias. In Y. Lebrun (ed.), *From the Brain to the Mouth: Acquired Dysarthria and Dysfluency in Adults* (pp. 11–47). Dordrecht: Kluwer.

Hesselink, J. M. K., van Gijn, J., & Verwey, J. C. (1987) Hyperkinetic mutism. *Neurology*, 37, 1566.

Hiiemae, K. M. (2000) Feeding in mammals. In K. Schwenk (ed.), *Feeding: Form, Function and Evolution in Tetrapod Vertebrates* (pp. 411–48). San Diego, CA: Academic Press.

Hillis, A. E., Work, M., Barker, P. B., Jacobs, M. A., Breese, E. L., & Maurer, K. (2004) Re-examining the brain regions crucial for orchestrating speech articulation. *Brain*, 127, 1479–87.

Hopkins, W. D. & Cantalupo, C. (2004) Handedness in chimpanzees (*Pan troglodytes*) is associated with asymmetries of the primary motor cortex but not with homologous language areas. *Behavioral Neuroscience*, 118, 1176–83.

Ikeda, A., Lüders, H. O., Burgess, R. C., & Shibasaki, H. (1992) Movement-related potentials recorded from supplementary motor area and primary motor area. *Brain*, 115, 1017–43.

Indefrey, P. & Levelt, W. J. M. (2004) The spatial and temporal signatures of word production components. *Cognition*, 92, 101–44.

Jürgens, U. (2002) Neural pathways underlying vocal control. *Neuroscience and Biobehavioral Review*s, 26, 235–58.

Jürgens, U. & Alipour, M. (2002) A comparative study on the cortico-hypoglossal connections in primates, using biotin dextranamine. *Neuroscience Letters*, 328, 245–8.

Jürgens, U. & Ploog, D. (1981) On the neural control of mammalian vocalization. *Trends in Neuroscience*, 4, 135–7.

Jürgens, U. & von Cramon, D. (1982) On the role of the anterior cingulate cortex in phonation: A case report. *Brain and Language*, 15, 234–8.

Jungers, W. L., Pokempner, A. A., Kay, R. F., & Cartmill, M. (2003) Hypoglossal canal size in living hominoids and the evolution of human speech. *Human Biology*, 75, 473–84.

Kelly, R. M. & Strick, P. L. (2003) Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *Journal of Neuroscience*, 23, 8432–44.

Kent, R. D., Kent, J. F., & Rosenbek, J. C. (1987) Maximum performance tests of speech production. *Journal of Speech and Hearing Disorders*, 52, 367–87.

Kent, R. D., Kent, J. F., Weismer, G., & Duffy, J. R. (2000) What dysarthrias can tell us about the neural control of speech. *Journal of Phonetics*, 28, 273–302.

Kent, R. D., Martin, R. E., & Sufit, R. L. (1990) Oral sensation: A review and clinical perspective. In H. Winitz (ed.), *Human Communication and Its Disorders: A Review – 1990* (pp. 135–91). Norwood, NJ: Ablex.

Kim, J. S., Kwon, S. U., & Lee, T. G. (2003) Pure dysarthria due to small cortical stroke. *Neurology*, 60, 1178–80.

Kluin, K. J., Gilman, S., Markel, D. S., Koeppe, R. A., Rosenthal, G., & Junck, L. (1988) Speech disorders in olivopontocerebellar atrophy correlate with positron emission tomography findings. *Annals of Neurology*, 23, 547–54.

Kirzinger, A. & Jürgens, U. (1982) Cortical lesion effects and vocalization in the squirrel monkey. *Brain Research*, 233, 299–315.

Kuypers, H. G. J. M. (1958a) An anatomical analysis of cortico-bulbar connexions to the pons and lower brain stem in the cat. *Journal of Anatomy*, 92, 198–218.

Kuypers, H. G. J. M. (1958b) Some projections from the peri-central cortex to the pons and lower brain stem in monkey and chimpanzee. *Journal of Comparative Neurology*, 110, 221–55.

Kuypers, H. G. J. M. (1958c) Corticobulbar connexions to the pons and lower brain-stem in man: An anatomical study. *Brain*, 81, 364–88.

Larson, C. R. (1988) Brain mechanisms involved in the control of vocalization. *Journal of Voice*, 2, 301–11.

Larson, C. R., Sutton, D., & Lindeman, R. C. (1978) Cerebellar regulation of phonation in rhesus monkey (Macaca mulatta). *Experimental Brain Research*, 33, 1–18.

Lechtenberg, R. & Gilman, S. (1978) Speech disorders in cerbellar disease. *Annals of Neurology*, 3, 285–90.

Lee, B. C., Hwang, S. H., & Chang, G. Y. N. (1999) Isolated dysphagia due to a medullary infarction: A new lacunar syndrome. *European Neurology*, 41, 53–4.

Lesser, R. P., Lueders, H., Dinner, D. S., Hahn, J., & Cohen, L. (1984) The location of speech and writing functions in the frontal languge area: Results of extraoperative cortical stimulation. *Brain*, 107, 275–91.

Levelt, W. J. M. (2001) Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, 98, 13464–71.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–75.

Lotze, M., Erb, M., Flor, H., Huelsmann, E., Godde, B., & Grodd, W. (2000) fMRI evaluation of somatotopic representation in human primary motor cortex. *NeuroImage*, 11, 473–81.

Loucks, T. M. J. & De Nil, L. F. (2001) The effects of masseter tendon vibration on nonspeech oral movements and vowel gestures. *Journal of Speech, Language, and Hearing Research*, 44, 306–16.

MacNeilage, P. F. (1998) The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499–546.

Mao, C. C., Coull, B. M., Golper, L. A., & Rau, M. T. (1989) Anterior operculum syndrome. *Neurology*, 39, 1169–72.

Martin, R. (1992) Classification and evolutionary relationships. In S. Jones, R. Martin, & D. Pilbeam (eds.), *The Cambridge Encyclopedia of Human Evolution* (pp. 17–23). Cambridge: Cambridge University Press.

McGuinness, E., Sivertsen, D., & Allman, J. M. (1980) Organization of the face representation in macaque motor cortex. *Journal of Comparative Neurology*, 193, 591–608.

Mega, M. S. & Alexander, M. P. (1994) Subcortical aphasia: The core profile of capsulostriatal infarction. *Neurology*, 44, 1824–9.

Mohr, J. P., Pessin, M. S., Finkelstein, S., Funkenstein, H. H., Duncan, G. W., & Davis, K. R. (1978) Broca aphasia: Pathologic and clinical. *Neurology*, 28, 311–24.

Morecraft, R. J., Louie, J. L., Herrick, J. L., & Stilwell-Morecraft, K. S. (2001) Cortical innervation of the facial nucleus in the non-human primate: A new interpretation of the effects of stroke and related subtotal brain trauma on the muscles of facial expression. *Brain*, 124, 176–208.

Muellbacher, W., Artner, C., & Mamoli, B. (1999) The role of the intact hemisphere in recovery of midline muscles after recent monohemispheric stroke. *Journal of Neurology*, 246, 250–6.

Muellbacher, W., Boroojerdi, B., Ziemann, U., & Hallett, M. (2001) Analogous corticocortical inhibition and facilitation in ipsilateral and contralateral human motor cortex representation of the tongue. *Journal of Clinical Neurophysiology*, 18, 550–8.

Müller, J., Wenning, G. K., Verny, M. A., et al. (2001) Progression of dysarthria and dysphagia in postmortem-confirmed Parkinsonian disorders. *Archives of Neurology*, 58, 259–64.

Murphy, J. P. & Gellhorn, E. (1945) Multiplicity of representation versus punctate localization in the motor cortex. *Archives of Neurology and Psychiatry*, 54, 256–73.

Murray, G. M. & Sessle, B. J. (1992a) Functional properties of single neurons in the face primary motor cortex of the primate, I: Input and output features of tongue motor cortex. *Journal of Neurophysiology*, 67, 747–58.

Murray, G. M. & Sessle, B. J. (1992b) Functional properties of single neurons in the face primary motor cortex of the primate, II: Relations with trained orofacial motor behavior. *Journal of Neurophysiology*, 67, 759–74.

Naeser, M. A., Palumbo, C. L., Helm-Estabrooks, N., Stiassny-Eder, D., & Albert, M. L. (1989) Severe nonfluency in aphasia: Role of the medial subcallosal fasciculus and other white matter pathways in recovery of spontaneous speech. *Brain*, 112, 1–38.

Naeser, M. A., Martin, P. I., Nicholas, M., et al. (2005) Improved picture naming in chronic aphasia after TMS to part of right Broca's area: An open protocol study. *Brain and Language*, 93, 95–105.

Nestor, P. J., Graham, N. L., Fryer, T. D., Williams, G. B., Patterson, K., & Hodges, J. R. (2003) Progressive non-fluent aphasia is associated with hypometabolism centred on the left anterior insula. *Brain*, 126, 2406–18.

Nishikawa, M., Komiyama, M., Sakamoto, H., Yasui, T., & Nakajima, H. (1998) Cerebellar mutism after basilar artery occlusion: Case report. *Neurologia Medico-Chirurgica*, (Tokyo) 38, 569–73.

Ojemann, G. A. (1979) Individual variability in cortical localization of language. *Journal of Neurosurgery*, 50, 164–9.

Ojemann, G. A. (1994) Cortical stimulation and recording in language. In A. Kertesz (ed.), *Localization and Neuroimaging in Neuropsychology* (pp. 35–55). San Diego, CA: Academic Press.

Okuda, B. & Tachibana, H. (2000) Isolated dysarthria. *Journal of Neurology, Neurosurgery, and Psychiatry*, 68, 119–20.

Ozimek, A., Richter, S., Hein-Kropp, C., et al. (2004) Cerebellar mutism: Report of four cases. *Journal of Neurology*, 251, 963–72.

Parent, A. & Hazrati, L. N. (1995) Functional anatomy of the basal ganglia, I: The cortico-basal ganglia-thalamo-cortical loop. *Brain Research Brain Research Reviews*, 20, 91–127.

Paus, T., Petrides, M., Evans, A. C., & Meyer, E. (1993) Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses: A positron emission tomography study. *Journal of Neurophysiology*, 70, 453–69.

Penfield, W. & Roberts, L. (1959) *Speech and Brain-Mechanisms*. Princeton, NJ: Princeton University Press.

Penfield, W. & Welch, K. (1951) The supplementary motor area of the cerebral cortex: A clinical and experimental study. *Archives of Neurology and Psychiatry*, 63, 289–317.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1989) Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience*, 1, 153–70.

Pinto, S., Ozsancak, C., Tripoliti, E., Thobois, S., Limousin-Dowsey, P., & Auzou, P. (2004) Treatments for dysarthria in Parkinson's disease. *Lancet Neurology*, 3, 547–56.

Pinto, S., Gentil, M., Krack, P., et al. (2005) Changes induced by levodopa and subthalamic nucleus stimulation on Parkinsonian speech. *Movement Disorders*, 20, 1507–15.

Postuma, R. B. & Dagher, A. (2006) Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. *Cerebral Cortex*, 16, 1508–21.

Preuss, T. M., Stepniewska, I., & Kaas, J. H. (1996) Movement representation in the dorsal and ventral premotor areas of owl monkeys: A microstimulation study. *Journal of Comparative Neurology*, 371, 649–76.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001) A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98, 676–82.

Riecker, A., Ackermann, H., Wildgruber, D., Dogil, G., & Grodd, W. (2000a) Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. *NeuroReport*, 11, 1997–2000.

Riecker, A., Ackermann, H., Wildgruber, D., et al. (2000b) Articulatory/phonetic sequencing at the level of the anterior perisylvian cortex: A functional magnetic resonance imaging (fMRI) study. *Brain and Language*, 75, 259–76.

Riecker, A., Kassubek, J., Gröschel, K., Grodd, W., & Ackermann, H. (2006) The cerebral control of speech tempo: Opposite relationship between speaking rate and BOLD signal changes at striatal and cerebellar structures. *NeuroImage*, 29, 46–53.

Riecker, A., Mathiak, K., Wildgruber, D., et al. (2005) fMRI reveals two distinct cerebral networks subserving speech motor control. *Neurology*, 64, 700–6.

Riecker, A., Wildgruber, D., Grodd, W., & Ackermann, H. (2002) Reorganization of speech production at the motor cortex and cerebellum following capsular infarction: A follow-up fMRI study. *Neurocase*, 8, 417–23.

Robinson, B. W. (1967) Vocalization evoked from forebrain in Macaca mulatta. *Physiology and Behaviour*, 2, 345–54.

Rödel, R. M. W., Laskawi, R., & Markus, H. (2003) Tongue representation in the lateral cortical motor region of the human brain as assessed by transcranial magnetic stimulation. *Annals of Otology, Rhinology and Laryngology*, 112, 71–6.

Rubens, A. B. (1975) Aphasia with infarction in the territory of the anterior cerebral artery. *Cortex*, 11, 239–50.

Sasaki, K., Nambu, A., Tsujimoto, T., Matsuzaki, R., Kyuhou, S., & Gemba, H. (1996) Studies on integrative functions of the human frontal association cortex with MEG. *Cognitive Brain Research*, 5, 165–74.

Sawczuk, A. & Mosier, K. M. (2001) Neural control of tongue movement with respect to respiration and swallowing. *Critical Reviews in Oral Biology Medicine*, 12, 18–37.

Schiff, H. B., Alexander, M. P., Naeser, M. A., & Galaburda, A. M. (1983) Aphemia: Clinical-anatomic correlations. *Archives of Neurology*, 40, 720–7.

Semendeferi, K. (2000) Advances in the study of hominoid brain evolution: Magnetic resonance imaging (MRI) and 3-D reconstruction. In D. Falk and K. R. Gibson (eds.), *Evolutionary Anatomy of the Primate Cerebral Cortex* (pp. 257–89). Cambridge: Cambridge University Press.

Sherwood, C. C. (2005) Comparative anatomy of the facial motor nucleus in mammals, with an analysis of neuron numbers in primates. *The Anatomical Record*, 287A, 1067–79.

Sherwood, C. C., Hof, P. R., Holloway, R. L., et al. (2005) Evolution of the brainstem orofacial motor system in primates: A comparative study of trigeminal, facial, and hypoglossal nuclei. *Journal of Human Evolution*, 48, 45–84.

Sherwood, C. C., Holloway, R. L., Erwin, J. M., Schleicher, A., Zilles, K., & Hof, P. R. (2004) Cortical orofacial motor representation in Old World monkeys, great apes, and humans, I: Quantitative analysis of cytoarchitecture. *Brain, Behavior and Evolution*, 63, 61–81.

Shuren, J. E., Schefft, B. K., Yeh, H.-S., Privitera, M. D., Cahill, W. T., & Houston, W. (1995) Repetition and the arcuate fasciculus. *Journal of Neurology*, 242, 596–8.

Shuster, L. I. & Lemieux, S. K. (2005) An fMRI investigation of covertly and overtly produced mono- and multisyllabic words. *Brain and Language*, 93, 20–31.

Sidtis, J. J., Gomez, C., Groshong, A., Strother, S. C., & Rottenberg, D. A. (2006) Mapping cerebral blood flow during speech production in hereditary ataxia. *NeuroImage*, 31, 246–54.

Sidtis, J. J., Strother, S. C., Anderson, J. R., & Rottenberg, D. A. (1999) Are brain functions really additive? *NeuroImage*, 9, 490–6.

Sidtis, J. J., Strother, S. C., & Rottenberg, D. A. (2003) Predicting performance from functional imaging data: Methods matter. *NeuroImage*, 20, 615–24.

Sokolov, A. N. (1972) *Inner Speech and Thought*. New York: Plenum.

Sörös, P., Sokoloff, L. G., Bose, A., McIntosh, A. R., Graham, S. J., & Stuss, D. T. (2006) Clustered functional MRI of overt speech production. *NeuroImage*, 32, 376–87.

Sutton, D., Larson, C., & Lindeman, R. C. (1974) Neocortical and limbic lesion effects on primate phonation. *Brain Research*, 71, 61–75.

Sutton, D., Trachy, R. E., & Lindeman, R. C. (1985) Discriminative phonation in macaques: Effects of anterior mesial cortex damage. *Experimental Brain Research*, 59, 410–13.

Tanji, K., Suzuki, K., Yamadori, A., et al. (2001) Pure anarthria with predominantly sequencing errors in phoneme articulation: A case report. *Cortex*, 37, 671–8.

Tokuno, H., Takada, M., Nambu, A., & Inase, M. (1997) Reevaluation of ipsilateral corticocortical inputs to the orofacial region of the primary motor cortex in the macaque monkey. *Journal of Comparative Neurology*, 389, 34–48.

Ungerleider, L. G., Doyon, J., & Karni, A. (2002) Imaging brain plasticity during motor skill learning. *Neurobiology of Learning and Memory*, 78, 553–64.

Urasaki, E., Uematsu, S., Gordon, B., & Lesser, R. P. (1994) Cortical tongue area studied by chronically implanted subdural electrodes – with special reference to parietal motor and frontal sensory responses. *Brain*, 117, 117–32.

Urban, P. P., Hopf, H. C., Fleischer, S., Zorowka, P. G., & Müller-Forell, W. (1997) Impaired cortico-bulbar tract function in dysarthria due to hemispheric stroke: Functional testing using transcranial magnetic stimulation. *Brain*, 120, 1077–84.

Urban, P. P., Rolke, R., Wicht, S., et al. (2006) Left-hemispheric dominance for articulation: A prospective study on acute ischaemic dysarthria at different localizations. *Brain*, 129, 767–77.

Urban, P. P., Wicht, S., Fitzek, C., Vukurevic, G., Stoeter, P., & Hopf, H. C. (2000) Topodiagnostik ischämisch bedingter Dysarthrophonien [Topodiagnostics of dysarthrias due to ischemic stroke]. *Klinische Neuroradiologie*, 10, 35–45.

Urban, P. P., Wicht, S., Hopf, H. C., Fleischer, S., & Nickel, O. (1999) Isolated dysarthria due to extracerebellar lacunar stroke: A central monoparesis of the tongue. *Journal of Neurology, Neurosurgery, and Psychiatry*, 66, 495–501.

Urban, P. P., Wicht, S., Vukurevic, G., et al. (2001) Dysarthria in acute ischemic stroke: Lesion topography, clinico-radiologic correlation, and etiology. *Neurology*, 56, 1021–7.

Uylings, H. B. M., Malofeeva, L. I., Bogolepova, I. N., Amunts, K., & Zilles, K. (1999) Broca's language area from a neuroanatomical and developmental perspective. In C. M. Brown & P. Hagoort (eds.), *The Neurocognition of Languge* (pp. 319–36). Oxford: Oxford University Press.

Vigneau, M., Beaucousin, V., Hervé, P. Y., et al. (2006) Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, 30, 1414–32.

Von Cramon, D. (1981) Bilateral cerebellar dysfunctions in a unilateral meso-diencephalic lesion. *Journal of Neurology, Neurosurgery, and Psychiatry*, 44, 361–3.

Weiller, C., Willmes, K., Reiche, W., et al. (1993) The case of aphasia or neglect after striatocapsular infarction. *Brain*, 116, 1509–25.

West, R. A. & Larson, C. R. (1995) Neurons of the anterior mesial cortex related to faciovocal activity in the awake monkey. *Journal of Neurophysiology*, 74, 1856–69.

Wildgruber, D., Ackermann, H., & Grodd, W. (2001) Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: Effects of syllable repetition rate evaluated by fMRI. *NeuroImage*, 13, 101–9.

Wildgruber, D., Ackermann, H., Klose, U., Kardatzki, B., & Grodd, W. (1996) Functional lateralization of speech production at primary motor cortex: A fMRI study. *NeuroReport*, 7, 2791–5.

Wise, R. J. S., Greene, J., Büchel, C., & Scott, S. K. (1999) Brain regions involved in articulation. Lancet, 353, 1057–61.

Woolsey, C. N., Erickson, T. C., & Gilson, W. E. (1979) Localization in somatic sensory and motor areas of human cerebral cortex as determined by direct recording of evoked potentials and electrical stimulation. *Journal of Neurosurgery*, 51, 476–506.

Ziegler, W. (2002) Task-related factors in oral motor control: Speech and oral diadochokinesis in dysarthria and apraxia of speech. *Brain and Language*, 80, 556–75.

Ziegler, W. (2005) A nonlinear model of word length effects in apraxia of speech. *Cognitive Neuropsychology*, 22, 603–23.

Ziegler, W. (2008) Apraxia of speech. In G. Goldenberg and B. Miller (eds.), *Handbook of Clinical Neurology*, vol. 88 (3rd series) (pp. 269–85). London: Elsevier.

Ziegler, W., Kilian, B., & Deger, K. (1997) The role of the left mesial frontal cortex in fluent speech: Evidence from a case of left supplementary motor area hemorrhage. *Neuropsychologia*, 35, 1197–208.

Zimmermann, E. (1992) Vocal communication by non-human primates. In S. Jones, R. Martin, & D. Pilbeam (eds.), *The Cambridge Encyclopedia of Human Evolution* (pp. 124–7). Cambridge: Cambridge University Press.

7 Development of Neural Control of Orofacial Movements for Speech

ANNE SMITH

1 Introduction

The purpose of this chapter is to provide an integrative overview of studies of the development of the neuromotor processes involved in controlling articulatory movements for speech. Such a review is not currently available and seems warranted given the appearance over the past decade of a number of studies of prespeech and speech motor processes in infants, preschoolers, school-age children, and adolescents. There has been considerable interest in the processes underlying speech motor development in infancy and childhood for many years, but until recently, most investigators relied solely on measurements of the speech acoustic signal to infer underlying physiological processes. While many important insights into speech motor development arose from this work (e.g., Kent, 1976; Nittrouer, 1993), direct measurements of articulatory movements in children and infants provide a new avenue to expand our knowledge of speech motor development. Fortunately over the past decade, a variety of technologies have become available which allow the noninvasive transduction of articulatory movements in children and adults. There also have been numerous advances in examining the development of respiratory and laryngeal behaviors in infants and children (e.g., Stathopoulos & Sapienza, 1993, 1997; Boliek et al., 1996, 1997; Huber et al., 1999; Moore et al., 2001; Connaghan et al., 2004). However, a comprehensive review that includes these components of the speech motor system is beyond the scope of the present chapter.

When a scientist or indeed a casual observer contemplates the act of speaking, the result is usually amazement at how speakers produce this complicated and multilayered output in such an apparently effortless and rapid manner. A conceptual approach that helps to simplify the problem of speech motor control is illustrated in Figure 7.1. In a lower layer of this diagram, the groups of motorneuron pools (the neurons that innervate muscles) are shown. At this level, it is clear that for speech (or any other motor behavior) to be produced, the nervous system must generate sets of commands to drive these motorneuron pools. These command signals must be coordinated in time and space for the appropriate sequences

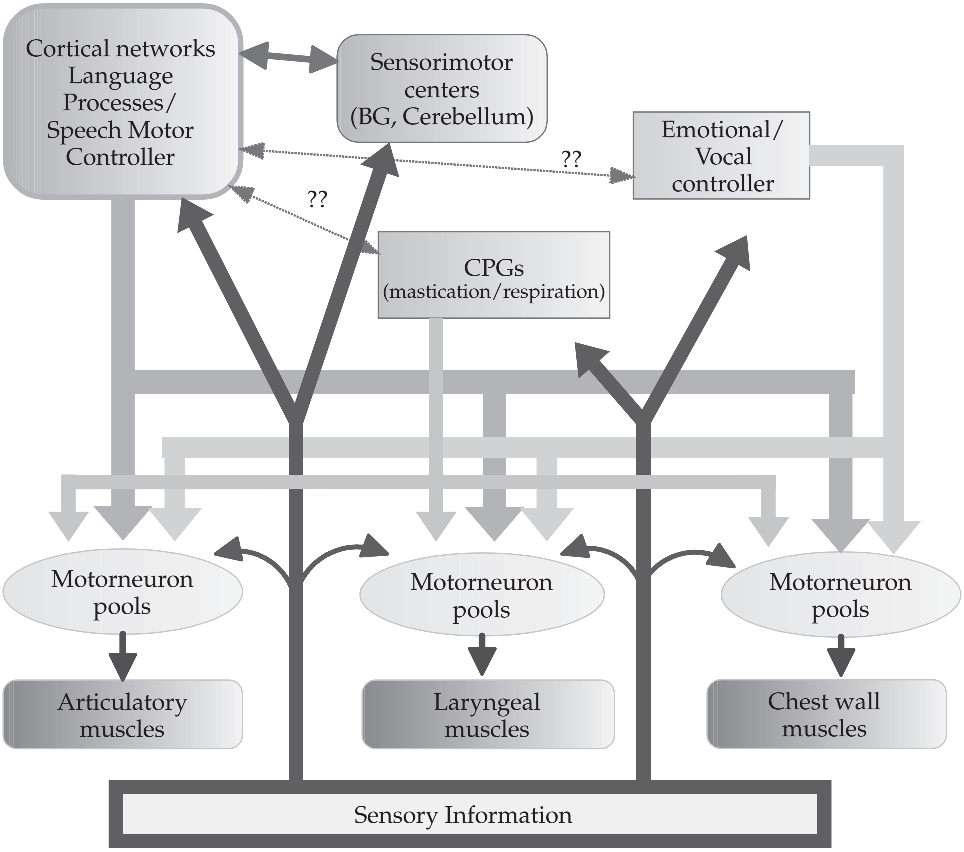


Figure 7.1 A schematic diagram of the many control pathways operating on the motorneuron pools that we use in speaking (BG = basal ganglia; CPG = central pattern generator).

of muscle activation to occur. Therefore, we can attempt to understand speech motor control processes by investigating where in the nervous system these neural commands to muscles are generated and how they are modified to achieve a variety of linguistic and metalinguistic goals.

Also included in Figure 7.1 are the centers involved in emotional vocalization, and another box representing central pattern generators (CPGs), networks of neurons in the brainstem that drive phylogenetically older behaviors such as mastication, respiration, and swallowing. These neural centers also have relatively direct connections to the motorneuron pools that we use for speaking. Thus, if we think of the motorneuron pools as the soldiers that are put into action to control muscle contraction, there are several different "generals" that can command them to produce the quite distinctive behaviors of speaking, chewing, quiet breathing, and laughing. Thus the speech control systems in the cortex must interact with these older neural control centers in some way. How these interactions occur is a matter of some debate (thus the question marks about these connections in Figure 7.1). Some authors suggest that the older circuits, for example, for emotional vocalization, mastication, and respiration, are engaged by the speech control system, which takes advantage of these pre-existing neural connections (MacNeilage, 1998; Lund & Kolta, 2006). The speech controller is hypothesized to bias these circuits in a way that produces the muscle activation patterns for speech, rather than for chewing or laughing. Other authors (von Euler, 1982; Moore & Ruark, 1996) suggest that the cortical networks involved in speech completely bypass these brainstem centers, arguing that the muscle activation patterns are radically distinctive for speech compared to the other behaviors and that speech is organized around entirely different (linguistic) goals. Therefore they hypothesize that the emotional vocalization center and other CPGs are simply suppressed and bypassed when cortical networks are engaged for speech. Anyone who has tried to carry on a conversation while jogging will know, however, that when speech goals interfere with metabolic demands for oxygen, the two systems compete, and metabolic breathing wins the battle for control of the motorneuron pools. Finally, this issue of the relationship between speech cortical networks and the older CPG neural networks has also been an important one in the study of the development of speech motor control. Relevant to the present chapter, there is debate about whether speech processes develop out of pre-existing oral motor behaviors, such as sucking and chewing, or whether the development of speech motor control takes an entirely independent course (Moore & Ruark, 1996; Ruark & Moore, 1997; MacNeilage, 1998).

2 Components of Articulatory Motor Control

2.1 Central mechanisms for articulatory control

The cortical and subcortical networks involved in language formulation and the planning and production of speech in adults have been investigated in a large literature starting with clinical lesion studies and now greatly expanded by neuroimaging studies. Interestingly, much more imaging work has focused on language processing, for example in verb generation tasks, rather than specifically on the prespeech planning and execution of speech movements (reviewed by Indefrey & Levelt, 2000). In a recent functional magnetic resonance imaging (fMRI) study, Bohland and Guenther (2006) focused their investigation on prespeech motor planning and on execution. They investigated the preparation and production of nonsense syllable sequences varying in length and complexity. Both "go" (response preparation and execution) and "no go" (response preparation only) conditions were included, so the differences in networks involved in speech motor preparation could be distinguished from those involved in overt speech production. Activations were observed in the areas of the brain that we would predict to be activated on the basis of earlier investigations, including bilateral activation of pre- and postcentral gyri, ventral motor and sensory cortical areas, anterior superior temporal cortex, medial premotor areas, supplementary motor area (SMA), the basal ganglia, the cerebellum, and the thalamus. Left lateralized activations emerged in the inferior frontal sulcus and anterior insula. Compared to the "no go" condition, the "go" condition resulted in significantly more bilateral activation in the primary sensorimotor areas representing lips, tongue, jaw, and larynx. This finding points to the critical role for these areas in generating the commands to the musculature for speech and integrating somatosensory feedback for online control of coordination of the articulators. Bilateral rather than a left lateralized activation is observed, because both hemispheres are involved in ongoing sensorimotor control. In contrast in the preparation only "no-go" trials, a left lateralized response was observed in the ventral motor and premotor cortices. Bohland and Guenther hypothesized that preparation for speaking "primes" motor cortical cells primarily in the left hemisphere, while overt speaking requires bilateral sensorimotor control.

Thus it is clear from studies of adults that, as shown in Figure 7.1, widely distributed cortical and subcortical networks are activated in the planning and production of speech. How and when do these widely distributed, highly specialized neural networks for speech production develop? The infant is not born with these networks in place. While functional imaging studies are more methodologically challenging, many structural imaging studies have been completed in infants and children. Advances in neuroimaging methods, especially MRI, have opened up the area of pediatric imaging. Lenroot and Giedd (2006) provided an excellent review of brain development in children and adolescents as revealed by morphometric measures of gray matter and white matter from MRI scans. They include results from their own large-scale longitudinal study of typically developing children and adolescents at the National Institute of Mental Health. Important for the present chapter is the finding that the growth of gray matter volume follows a regionally specific inverted U-shaped developmental curve. For example, in the frontal lobe, gray matter volume reaches its maximum at 11 years in girls and 12 years in boys, and temporal gray matter volume peaks at 16 years in boys and girls. In contrast, the amount of white matter in the brain generally increases throughout childhood and adolescence. Myelination and dendritic and axonal arborization continue well into middle age.

These volumetric studies have been very helpful in mapping the protracted development of the brain and the regional differences in developmental trajectories. Furthermore, these methods are now being employed to map the neural bases of a variety of developmental disorders. Jancke et al. (2007) found decreased white matter volumes and anomalous anatomy in a left-hemisphere fronto-temporal network that included both language and motor areas in a group of children (aged 4–10 years) with developmental language disorder.

Another recently developed technique based on magnetic resonance scans is diffusion tensor imaging (DTI), which allows mapping of fiber tracts essential for interregional communication in the central nervous system (CNS); (Paus et al., 2001). Many different laboratories are undertaking studies of children with normal developmental histories and children with a variety of atypical histories. Using DTI, Deutsch et al. (2005) studied children with a wide range of reading performance levels. They reported that in the left temporo-parietal region the white matter structure measured using fractional anisotropy (an index of the coherent structure of fiber tracts) correlated with behavioral indices of reading, spelling, and rapid naming performance. In adults who stutter there is evidence, based on DTI, that left-hemisphere fiber tracts that connect pre- and primary cortical motor areas are disrupted (Sommer et al., 2002). This methodology is extremely promising for future studies mapping the development of white matter tracts in normally developing children and in those with a variety of speech motor problems.

2.2 Peripheral mechanisms in articulatory control and coordination

2.2.1 Motorneuron pools and muscles As shown schematically in Figure 7.1, groups of motorneurons (motorneuron pools) located in columns in the brainstem innervate orofacial muscles. Muscles of the lips, jaw, and tongue are innervated by different cranial nerves: the facial, trigeminal, and hypoglossal nerves, respectively. In his review of the muscles involved in speech production, Kent (2004) emphasized the uniqueness of these muscles and their distinctiveness from limb muscles. Even within the articulatory muscle system, there is a remarkable variety of muscle types and sensory receptors. For example, the jaw-opening and closing muscles operate around a joint, while the muscles of the lips form a sphincter and do not have any bony attachments via tendons. The muscles of the human tongue operate as a muscular hydrostat (Smith & Kier, 1989), a class of muscles that includes trunks, tongues, and tentacles. Kent provided an excellent review of the histochemical properties of the muscles involved in speech and noted that many of these muscles have heterogeneous fiber types and varying regional distributions of the different classes of fibers. Stal et al. (2003) investigated the fiber composition of three intrinsic muscles of the human tongue – longitudinalis, verticalis, and transversus – in four anterior to posterior regions using morphological, enzyme, and immunohistochemical techniques. All three muscles show a mix of slow, intermediate, and fast fibers, but small, fast fibers predominate in the anterior region, while larger diameter, slow and intermediate fibers are predominant in the posterior region. Many cranial muscles show this reversal in pattern compared to limb muscles, in which slow fibers typically are the smallest in diameter, and fast fibers are the largest. Stal et al. note, "This muscle fibre composition of the tongue differs from those of limb, orofacial and masticatory muscles, probably reflecting genotypic as well as phenotypic functional specialization in oral function" (p. 147). Given the differences between humans and other primates in fiber types and distribution in the muscles involved in speech (e.g., Sciote et al., 2003), it seems highly likely that many of the features of the human orofacial, lingual, and mandibular muscles evolved uniquely to meet the complex, rapid, low-force, positional demands of speech (Kent, 2004).

2.2.2 Orofacial sensory information Many different kinds of sensory receptors are found in the muscles, skin, and connective tissues of the articulatory system. As with the histochemical properties of the cranial muscles, the nature of sensory innervation across lips, jaw, and tongue varies remarkably. Muscle spindles are densely supplied in the jaw-closing muscles (but are absent or sparse in jaw-opening muscles) and in the intrinsic tongue muscles (see A. Smith, 1992, for a review). Lip muscles lack muscle spindles (Stal et al., 1990). The vermillion borders, the intraoral mucosa, and the hairy skin of the face are densely innervated with a variety of mechanoreceptors. Direct recording using microneurography in humans has demonstrated that perioral and intraoral cutaneous receptors are activated during speaking (Trulsson & Johansson, 2002). While such recordings have not been made in human jaw-closing muscle spindle afferents, it seems highly likely, based on recordings of monkey jaw muscle spindle afferents during voluntarily controlled jaw movement (Larson et al., 1983), that these receptors provide very precise information to the CNS about jaw-opening velocity and position during speech. Recordings from the low-threshold cutaneous receptors of the lips and mucosa show that in addition to being activated by direct contact with external objects, they signal contact between the lips, changes in air pressure associated with speech sounds, and deformations of the tissues resulting from movement (Trulsson & Johansson, 2002). Therefore they provide both exteroceptive and proprioceptive information.

It is clear that the CNS is receiving highly specific and dynamic exteroceptive and proprioceptive information during speech. In addition to signaling sensory information along central pathways, afferent fibers from these low-threshold mechanoreceptors also make reflex connections with cranial motorneuron pools. These pathways are illustrated in Figure 7.1. In general, the importance of these reflex pathways has been underestimated in research on speech motor control, and the clinical literature often makes the assumption that cranial reflex pathways are arranged in the same patterns classically described for antagonistic pairs of limb muscles. However, the anatomical data indicate that analogous circuitry (e.g., the Ia inhibitory pathway to the antagonist) is unlikely, because antagonistic pairs of cranial muscles do not show the same sensory innervation patterns (e.g., jaw-closing muscles have spindles, while jaw-opening muscles do not). Another widespread clinical impression about oral motor reflexes is that they have powerful effects in the infant, and that these effects fade with development as reflex circuits are suppressed and overtaken by cortical originating control networks. As the review below will reveal, reflex effects of some classes of low-threshold mechanoreceptors in fact increase in gain during the childhood years.

2.2.3 Oral motor reflexes Muscle spindles in the human jaw-closing muscles have powerful excitatory effects on both their muscle of origin and bilaterally on all other jaw-closing muscles (Smith, Moore, & Pratt, 1985). Precise characteristics of the jaw-stretch reflex response in normal adults, including effects of varying frequency and amplitude of step and sinusoidal stretches, have been documented (Cooker et al., 1980). Very small step stretches to the mandible produce very large electromyographic (EMG) and jaw-closing force responses at a very short latency, 8–10 ms. In a recent study (Finan & Smith, 2005) the same techniques were applied to assess stretch reflex responses in two groups of typically developing children (age 5–6 and 9–10 years) and young adults. Latency of the responses increased with age, but the 9 to 10-year-old group showed the largest amplitude of responses and the highest reflex gain, which was significantly larger compared to both the younger children and the adult group. This suggests an inverted U-shaped growth curve for jaw-stretch reflex gain and the interesting conclusion that, as children are learning and refining speech motor skills, reflex circuits actually have higher gains.

Afferent fibers from low-threshold perioral and intraoral mechanoreceptors also make reflex connections with jaw-closing and lip muscles. We undertook a series of studies of the effects of intraoral stimulation on the jaw-closing muscle system (Smith, Moore, Weber, et al., 1985; Smith et al., 1991; Wood & Smith, 1992). Reflex responses were measured as stimulus-linked changes in masseter EMGs and jaw-closing force measured against a background static biting level. The mechanical stimuli were small (1 mm) displacements of a servo-controlled smooth metal probe, which were perceived as light, innocuous gliding changes in contact. In our initial study we applied the innocuous mechanical stimulus to eight sites on the tongue dorsum and palate in a large group of young adults. Stimulation of the palate produced primarily suppressions of EMG and drops in jaw-closing force, while stimulation of the tongue, especially in anterior placements, typically resulted in excitatory EMG responses and increased jaw-closing force. This study revealed a set of spatially organized reflex responses of the jaw-closing muscles in response to light cutaneous stimulation. Such responses could not be interpreted as primarily of protective significance.

In later studies of reflexes of the jaw-closing system produced by the same innocuous mechanical stimulus, we tested groups of 7 to 8-year-old (Smith et al., 1991) and 4 to 6-year-old children (Wood & Smith, 1992). The 7 to 8-year-old children had variable responses, with some showing adultlike EMG and force responses, while others showed extremely large and long-lasting jaw-closing responses to the stimulus. These large, long duration responses observed in some of the 7- and 8-year-old children were much larger than the reflex responses we had observed in our earlier study of adults. This led us to the hypothesis that 7–8 years is a transitional period in oral motor reflex development, with the very large responses at this age perhaps a sign of a less mature pattern present in some of the children. In a follow-up study of 4 to 6-year-old children (Wood & Smith, 1992), we tested the hypothesis that the younger children would have extremely high-gain cutaneous reflexes operating on the jaw-closing system. We were surprised to find that compared to the 7 to 8-year-olds, the younger children often had no response or responses that were much smaller than those of the 7- and 8-year-old children. This result is consistent with the findings for the jaw-stretch reflex, again suggesting an inverted U-shaped growth curve and that cutaneous oral motor reflexes develop along with the acquisition of speech motor skills.

Another reflex circuit that has received attention in relation to speech production is the perioral reflex, the response of lip muscles to innocuous mechanical stimulation of the vermilion border and/or perioral hairy skin (McClean & Clay, 1994; Smith, Moore, McFarland, et al., 1985; Smith et al., 1987; Barlow & Bradford, 1996). This short-latency response is typically a multi-component excitation of orbicularis oris to light mechanical stimulation. It is present in newborn infants (Barlow et al., 2001), and like the jaw-stretch and cutaneous jaw reflexes described above, a preliminary study of a small number of school-age children suggests that the perioral reflex response grows in amplitude over childhood (Barlow et al., 1993). In adulthood, the perioral reflex shows a highly localized spatial organization with responses highly dependent on the site of stimulation (Smith et al., 1987; Barlow & Bradford, 1996).

This review clearly leads to the conclusion that the motorneuron pools controlling the muscles involved in articulation can be powerfully affected by excitatory and inhibitory short-latency synaptic inputs arising from afferent feedback. Furthermore, rather than reflecting a set of primitive reflex circuits whose potency decreases over the childhood years, many of these brainstem reflex circuits have been shown to increase dramatically in gain in the school-age years. What role do these reflex pathways play in the control and coordination of speech movements in developing and mature oral motor systems? Unless a particular reflex pathway is suppressed during an ongoing movement, synaptic inputs arising from the circuit would be part of the synaptic drive affecting motorneuron pool excitability during that behavior. It might be useful to think of the synaptic drive to motorneuron pools during speech as arising from an "orchestra" of neural pathways, and the final "symphony" produced depends upon the mix of sources selected (E. Luschei, personal communication). Sherrington (1906) called the motorneuron pool "the final common pathway," because it is the site of integration of all the synaptic inputs from the various sources, and ultimately determines the activity of muscles. Returning to the question posed above about the nature of these reflex effects during speech, we must conclude that the answer is not known. One study in adults suggested that perioral reflex pathways can be activated by light mechanical stimulation of the perioral skin during speaking, and thus this pathway could contribute to the excitability of orbicularis oris (Smith, Moore, McFarland, & Weber, 1985); however another study suggested that the perioral reflex pathway is suppressed prior to the onset of speech (McClean & Clay, 1994). The gain of the jaw-stretch reflex has not been studied prior to or during speech, however it seems likely that the high-velocity opening movements involved in speaking would produce intense activation of muscle spindle afferents, which in turn would excite jaw-closing muscles. This reflex excitation of the closing muscles during the opening phase of movement would be functionally appropriate, because jaw-opening and closing muscles show highly co-activated patterns of activity during speech (Moore et al., 1988).

In our study of cutaneous reflexes of the jaw-closing system in 4- to 6-year-olds, we included a small group of children with speech delays (Wood & Smith, 1992). Interestingly, this group had smaller, less mature reflexes compared to their typically developing peers. Further work is needed to determine the significance of reflex pathways arising in low-threshold mechanoreceptors that are activated during speech, their developmental course, and their potential as indices of atypical oral motor and/or speech development. Finally, while I have chosen to emphasize the nature of brainstem reflex circuits, because they are often neglected in the speech motor control literature, there is also very little information generally about how children use somatosensory and/or auditory feedback centrally to shape the neural commands for speech arising from cortical networks.

2.2.4 Anatomical restructuring of the vocal tract during development Finally, no review of peripheral components of speech motor control and its development would be complete without mentioning the structural changes the vocal tract undergoes from infancy through childhood. These changes are so dramatic that the term "anatomic restructuring" has been used to describe these physical changes (Vorperian et al., 2005). While the nature and significance of these physical changes have been generally described for many years (e.g., Bosma, 1975; Kent, 1981), more recent studies using MRI have provided detailed, quantitative descriptions of the anatomic changes in the bony and soft tissue structures of the vocal tract from infancy through adulthood. Vorperian et al. (2005) used MRI to measure lip thickness, hard- and soft-palate length, mandibular depth and length, and overall vocal tract length. They analyzed scans from 63 infants and children, from birth to 81 months, and from 12 adults. The results revealed no sexual dimorphism in the infants and children and an extremely accelerated growth rate between birth and 18 months. The various soft and bony tissues they measured showed distinctive growth patterns, with increasing vocal tract length predominantly due to growth of oral/anterior structures during the first 18 months, while predominantly driven by growth of pharyngeal/posterior structures in later development. This anatomic restructuring with development means that the peripheral structures to be controlled via articulatory muscle activations are changing over time, and the neural circuits providing the input to the cranial motorneuron pools must be slowly changing and adapting as these anatomic changes occur. Clearly this source of variation must be considered when interpreting results of physiologic studies of developing speech motor control processes (Vorperian et al., 2005).

3 Development of Speech Motor Processes

3.1 The emergence of early vocalizations and their relationship to pre-existing behaviors

The newborn possesses a behavioral repertoire that includes breathing, sucking, crying, and a variety of spontaneous movements. The basic patterns for repetitive movements, such as those in sucking, chewing, and breathing, arise in brainstem CPGs. In a special volume of the *Journal of Communication Disorders*, Lund and Kolta (2006) and Barlow and Estep (2006) provided extremely helpful tutorials on CPGs, prepared especially for readers interested in the relationship between the various CPGs and speech production. Lund and Kolta (2006, p. 382) state, "The systems that control innate repetitive movements in humans and other animals have two basic characteristics: they contain assemblies of neurons that are capable of generating a fundamental rhythm, and they include feedback systems that adapt the rhythm to the state of the internal and external environments." This simple statement clarifies what was once a highly debated issue, whether there are behaviors that are solely under central control versus those that require sensory information in order to switch from one phase of movement to another (Gallistel, 1980). For many years it has been clear that circuits located in the brainstem can generate basic rhythmic movements in the absence of feedback. For example, Dellow and Lund (1971) demonstrated in the rabbit that the fundamental pattern of mastication, including coordinated movement of the jaw, lips, and tongue, can be generated by a brainstem CPG after all sensory inputs have been removed. However in a variety of studies of many different species and behaviors, it also has become clear that CPG-generated behaviors would not be adaptive unless their output was highly sensitive to the changing demands of the task. Thus, even relatively "simple" behaviors such as breathing, chewing, and sucking in the infant arise from the ongoing, dynamic interaction of phylogenetically old brainstem circuits with sensory information.

Recently, the critical importance of sensory information in modulating the activity of CPG-driven motor output has been shown in both full term and premature infants (Finan & Barlow, 1998; Barlow et al., 2004). These investigators developed a servo-controlled system to produce pressure changes in the nipple of a pacifier or "motorized nipple" (also called an "actifier," Barlow & Estep, 2006). The pattern of pressure changes was programmed to be similar to those characteristic of normal infant suck-cycle timing and amplitude. When the servo system was turned on, infants' natural sucking behaviors became entrained to the rhythm of the actifier. It seems reasonable to assume that in the newborn, this innocuous intraoral mechanical stimulation produces activity in a wide population of mechanoreceptors located in the lips, tongue, and jaw. This highly salient, patterned sensory inflow to the CNS modulates the activity of the brainstem CPG generating the suck behavior, resulting in the entrainment phenomenon. A basic principle of neural systems is that "cells that fire together wire together" (this catchy phrase is often used by modern neuroscientists to capture Hebb's (1949) proposal of a fundamental neural mechanism underlying associative learning). We can postulate that this patterned, repetitive bursting behavior of the oral sensorimotor systems serves to develop and sculpt neural connections. The discovery by Barlow and his colleagues that infants respond and entrain their sucking to intraoral mechanical input has had important clinical implications. They also studied premature infants who had long periods of oral sensory deprivation because of respiratory problems (Barlow & Estep, 2006). Within a relatively short time after these premature infants experience controlled oral stimulation via the actifier, an increase in natural sucking behavior occurs, which facilitates their ability to feed normally, a result which obviously has critical implications for their survival.

Many of the earliest behaviors of the human infant, then, reflect the presence at birth of brainstem CPGs. As infants develop, so do their abilities to chew, and the transition to more solid food occurs during the 5–8 month period (Sheppard & Mysak, 1984). The development of chewing has been studied both cross-sectionally (Gisel, 1991; Kiliaridis et al., 1991) and longitudinally (Green et al., 1997). The basic pattern of chewing is already well established in infants at 12 months (Green et al., 1997), but the patterns of activation of synergistic and antagonistic muscle pairs become more consistent with development. One of the most remarkable findings from the developmental chewing literature is the that the duration of the masticatory cycle remains virtually unchanged (mean cycle duration is approximately 0.7 seconds) from the age of 12 months into adulthood (Kiliaridis et al., 1991; Green et al., 1997). The details of the chewing cycle in terms of movements and muscle activity depend upon the nature of the food being chewed, evidence that the CPG output is modified to provide a pattern of activity appropriate for the food to be ingested.

Of course the period of transition into chewing soft and more solid foods at 5–8 months is also a time of great change in infant vocal development (Oller, 1980; Stark, 1980). It is during this time that infants show a dramatic drop in nonspeech-like vocalizations, such as cries and vegetative sounds, and an increase in speech-like vocalizations, including babbling (Nathani et al., 2006). One controversial issue concerning the motor processes underlying the emergence of babbling and more advanced forms of vocalizations in infants at this stage is the role played by pre-existing neural circuitry for ingestive behaviors. MacNeilage and Davis (MacNeilage & Davis, 1990, 2000; MacNeilage, 1998), in their frame/content theory of the evolution of speech production, propose that the fundamental organizational property of speech is the repetitive oral open–close cycle. They propose that the total open–close cycle represents a syllable, while the open and close phases represent segments, and vowels and consonants respectively. Furthermore, MacNeilage (1998) argues that these communication-related frames evolved from phylogenetically old ingestive behaviors involving mandibular oscillation, such as sucking and mastication. Resisting the argument that chewing is too simple to serve as a basis for the emergence of speech production, MacNeilage notes the complexities of the masticatory cycle and its highly adaptive nature related to ongoing task demands. Thus, he proposes that the masticatory CPG is the perfect candidate for "tinkering with" by evolutionary processes to produce the articulatory open–close cycles of human speech.

This view also has been supported by a number of neuroscientists who have studied centrally patterned behaviors, including mastication (Lund & Kolta, 2006) and locomotion (Grillner, 1982). Lund and Kolta note that the brainstem CPG for mastication receives input from cortical areas, especially from the inferior lateral region of the motor cortex. The CPG itself includes a core group of neurons with intrinsic bursting properties, and reorganization of subpopulations of these neurons, which can be produced by changes in sensory feedback and/or central drive, results in highly specific, adaptive movement patterns. Furthermore the subpopulations of neurons that constitute the CPG supply controlling inputs not only to muscles of the jaw, but also to muscles of the tongue and face; the CPG also biases reflex circuits to optimize the masticatory properties to ongoing demands. In sum, all of these features would seem to make this neural circuitry a perfect candidate for biasing by cortical networks to modulate its output to produce the rhythmic oral movements needed for dynamic control of the vocal tract during speech.

Seeking evidence in support of the frame/content theory, MacNeilage et al. (2000) compared the serial organization of infant babbling and early speech across 10 languages. Their analysis revealed four movement-related design features reflecting a "deep evolutionary heritage" operating on the pattern of infant vocalizations. These included the cyclical consonant–vowel alternation underlying the syllable (the "frame"), three within-cycle consonant–vowel co-occurrence preferences that were presumed to reflect biomechanical coupling properties of the articulators, and two other features related to consonant repetition and ease of production. In summary, MacNeilage and his colleagues have been proponents of a model of the transition from prespeech to speech vocalizations in infants in which strong evolutionary influences provide biological constraints on the fundamental units and structure of human languages. The underlying neural substrate is hypothesized to make use of brainstem pattern generation circuitry which can be flexibly biased to produce a range of adaptive behaviors, possibly including speech (see dotted lines in Figure 7.1).

On the other hand, as shown in Figure 7.1, an alternative hypothesis is that the activity of the motorneuron pools involved in speech is driven directly from motor cortex, via pathways that bypass the brainstem CPGs for respiration and mastication. This point of view was argued strongly by von Euler (1982), a neurophysiologist who made major contributions to delineating properties of the brainstem respiratory pattern generating circuitry. He argued strongly for complete separation of the control pathways for metabolic breathing and those involved in "voluntary" breathing, including speech breathing.

The view that the neural pathways for speech motor control do not arise from nor engage evolutionarily well-established brainstem networks for ingestive behaviors is also argued by Moore and colleagues (Moore & Ruark, 1996; Green et al., 1997; Ruark & Moore, 1997; Moore, 2004). This group was among the first to tackle the very difficult experimental task of obtaining physiological data, including both kinematic and electromyographic recordings, from infants and toddlers during early vocalizations and nonspeech oral motor behaviors. Their publications provide many new insights concerning oral motor development.

Green and Wilson (2006) studied spontaneously produced orofacial movements of infants. Using a video-based system, they captured the motion of passive reflective markers attached to infants' faces to track lip and jaw movements. Their study is particularly impressive, because of the relatively large number (n = 29) of infants aged 1, 5, 7, 9, and 12 months who were studied cross-sectionally. Only spontaneous facial movements *without any accompanying vocalization* were analyzed. Kinematic parameters computed included movement space, movement speed, movement duration, and spatial and temporal coupling between pairs of markers.

All of the infants produced spontaneous facial movements during the recording sessions with the 5-month-olds producing, on average, the most (approximately 60). Thus these investigators had a wide repertoire of spontaneous orofacial movements to examine for most of the infants they recorded. During the first year of life, these spontaneous orofacial movements showed some systematic changes. The speed of movements increased, while the duration of the spontaneous movement epochs decreased. Based on cross-correlational analysis, coupling of movements across pairs of facial markers increased. Perhaps the most interesting finding from the study, however, was that there was no evidence for stereotypic, repetitive, spontaneous oral movements, such as a rhythmic opening and closing of the jaw. The spatial and temporal characteristics of the spontaneous movements were not patterned or rhythmic but instead were highly variable. The movement spaces also were highly variable across epochs of spontaneous behaviors. This is in contrast to, for example, the highly stereotypic, rhythmic spontaneous leg movements observed in infants, labeled "stereotypies" by Thelen (1979; Thelen & Smith, 1994).

This result might be interpreted as very compelling evidence upon which to reject the frame/content theory of early infant vocal development. However, as emphasized with italics above, it is important to note that Green and Wilson (2006) only analyzed spontaneous oral movements not accompanied by vocalization. The older infants in their study would have been babbling, and presumably would have produced vocalized orofacial behaviors during the recording session. These behaviors were not analyzed in this report, but these oral movement sequences accompanied by vocalization most likely would show repetitive, rhythmic cycles of opening and closing movement. One of the most well-documented and salient features of babbling is its rhythmic syllabic structure (Kent et al., 1991). This leads to the interesting hypothesis that repetitive oral open–close movement sequences emerge only when orofacial, laryngeal, and respiratory systems are coactivated in a coordinated manner. Thus the nonvocalized, spontaneous oral motor behaviors described by Green and Wilson might be a distinctive class of spontaneous behaviors which bears little relationship to prespeech vocalizations.

Moore and Ruark (1996) recorded jaw-opening and closing muscle activity during spontaneous episodes of chewing, sucking, babbling, and speech in seven 15-month-olds. After rectifying and smoothing the EMGs to create muscle activity envelopes for each task, they computed cross-correlations among pairs of antagonistic and synergistic jaw muscles. Surprisingly, they found that the coupling between muscle pairs was greater for later appearing behaviors such as variegated babbling and early word production. Coupling of muscle pairs for chewing and sucking was less strong. Furthermore, a qualitatively different pattern of jaw-muscle coordination was characteristic of chewing, reciprocal activation of opening and closing muscles, compared to speech, which involves a high degree of co-activation of antagonistic jaw-muscle pairs. Thus, these authors found task-specific organization of jaw-muscle coordination in 2-year-olds, which resembled patterns observed in adults in an earlier study (Moore et al., 1988). Using a similar analysis applied to lip muscles in 2-year-olds, Ruark and Moore (1997) found that speech and nonspeech coordinative patterns of activation of lip muscles also were highly distinctive. They stated, "This level of coordinative specialization is consistent with . . . the accumulation of findings suggesting that children develop speech-specific coordinative mechanisms very early in life. Although conclusive results are yet to be obtained, the present findings support the suggestion that speech emerges separately from extant oral motor behaviors, and failed to support the existence of redundancy in control mechanisms across tasks (p. 1384)."

As Ruark and Moore stated above, the evidence is not conclusive regarding the role of brainstem CPGs in the emergence of speech. Differences in activation patterns of muscles across tasks could arise from distinctive biasing of the neural assemblies that produce rhythmical jaw-, facial-, and tongue-muscle activity during mastication. In fact, as already noted above, a cardinal feature of CPGs is their adaptability in the face of different task demands. Obviously, we cannot record from brainstem neurons to determine whether the neural networks of the CPG active during mastication are also active during speaking. However, with improved imaging techniques that are providing more and more spatial resolution, investigators may be able in the future to provide functional imaging data that can address this question.

Another approach to the issue of whether common control processes are engaged across distinctive motor behaviors, which has not been applied in pediatric populations, is to analyze the frequency content of EMGs of muscle pairs. In the respiratory system of humans (and rabbits, cats, and dogs) during metabolic breathing, there is a signature frequency of activity in brainstem neurons involved in pattern generation. This activity has been referred to as high-frequency oscillations or HFOs and occurs in the 60–110 Hz band. Activity of respiratory nerves recorded in experimental animals is highly coherent (coherence values are computed as the cross-correlation between two signals in the frequency domain) in this frequency band, and pairs of respiratory muscles in humans also show this highly coherent pattern during inspiration for metabolic breathing (Ackerson & Bruce, 1983). Smith and Denny (1990) recorded right- and left-diaphragm activity during speaking and metabolic breathing. We also recorded right- and left-masseter activity during speaking and chewing. We found that the signature band of coherent diaphragmatic muscle activity was present in metabolic breathing as expected, but that coherence in this band was greatly reduced during speaking. We interpreted this result as an indication that the respiratory CPG still contributed some synaptic drive to the respiratory motorneuron pools during speech, but that this drive was greatly reduced compared to its amplitude during metabolic breathing. We also observed a highly coherent frequency band in right- and left-masseter activity in the 40–60 Hz range during chewing. This coherent frequency band was also greatly suppressed during muscle activation for speaking. Again, results for both respiration and mastication indicated that while the CPGs for respiration and mastication are not completely bypassed, their activity is greatly reduced. Surface EMG electrodes were used in these experiments and thus they might be applicable in young children.

Finally, the unresolved issue of whether the neural control mechanisms for speech take advantage of earlier existing ingestive and/or respiratory CPG circuitry is an important one for the field of speech/language pathology. There is an intense debate about the use of nonspeech oral motor tasks as part of therapy for speech disorders in both children and adults (Luschei, 1991; Weismer & Liss, 1991; Forrest, 2002; Weismer, 2006). If the neural circuitry for speech is highly specialized and completely independent of neural control networks controlling nonspeech behaviors, some scientists and clinicians conclude that nonspeech-oriented therapies provide no rehabilitative benefit for speech, which is the target of the therapy. On the other hand, if the neural pathways involved in speech overlap those involved in nonspeech behaviors, a stronger case is made for the use of nonspeech oral motor approaches to facilitate reaching speech and/or language therapeutic goals. In any case, all neural systems affecting the timing and amplitude of muscle activity must operate through "the final common pathway," to the motorneurons and out to the muscles. Therefore, any task used to activate a muscle, whether it is a speech or nonspeech task, contributes to the health of the motorneurons and the muscle cells.

3.2 *Speech motor development in children*

Around 18 months of age, toddlers have a vocabulary of about 50 words, and they typically begin to produce two-word utterances at about this age (Brown, 1973). By 5 years, children are producing thousands of words in multiword utterances. The astonishingly rapid growth in the capacity of the speech production system to produce a variety of words, phrases, and sentences in the preschool years arises from a vast array of developing regional and inter-regional connections in the brain. The acquisition of new words continues well into adolescence, and studies of language perception using event-related potentials (ERPs) have demonstrated that some aspects of the neural networks underlying language processing are not adultlike until late adolescence (Holcomb et al., 1992; Neville et al., 1992). Before studies of the control of articulatory movements in late childhood and adolescence were available, it was often assumed that speech motor control processes were fully mature by age 10–12 years (e.g., Tingley & Allen, 1975). Recent studies, however, in which kinematic parameters of the articulatory system were measured in older children and adolescents, have demonstrated that the developmental time course for achieving mature, adult levels of speech motor control processes extends into late adolescence (Walsh & Smith, 2002; Cheng et al., 2007). Certainly the rate of change is much slower in the later years of the growth curve compared to the dramatically rapid rates of change in the preschool years (Smith & Zelaznik, 2004); yet there are still significant differences in some motor aspects of the speech of 16-year-olds compared to that of young adults (Walsh & Smith, 2002). Here we consider the protracted developmental course to adult speech motor control processes and some of the underlying factors that contribute to it.

3.2.1 Development of basic parameters of articulatory movement Amplitude, duration, and velocity are three fundamental parameters that can be measured for any movement trajectory. These parameters have been assessed in studies of speech production in children and adults. Intuitively, one would expect that smaller speakers, that is, children, would produce smaller speech movements compared to adults. Also, based on earlier acoustic and perceptual evaluations of children's speech, it is well known that children are slower speakers compared to adults. Therefore we also would predict that children would be moving at relatively low velocities, such that their speech movements are longer in duration.

Children do, in fact, produce articulatory movements for speech with lower velocities compared to adults (B. Smith and Gartenberg, 1984; B. Smith and McLean-Muse, 1986; A. Smith & Goffman, 1998), but a few early studies of small numbers of young children aged 4–7 years, who obviously have smaller orofacial structures compared to adults, suggested that young children produce oral movements for speech of equal amplitude to those of adults (B. Smith & Gartenberg, 1984; Sharkey & Folkins, 1985). Riely and Smith (2003) explored this issue directly by asking if a size principle operates in speech (e.g., that smaller speakers produce smaller articulatory movements). A size principle has been documented in locomotion, such that stride length is directly related to limb length (e.g., Beck et al., 1981), suggesting that biomechanical factors play a major role in determining the kinematic parameters of gait. Such a relationship has been reported for speech by Kuehn and Moll (1976); they found a positive relationship between oral structure size and the amplitude and velocity of speech movements for a small number of adult speakers.

In a study involving a relatively large number of participants, we (Riely & Smith, 2003) collected lip and jaw kinematic data from thirty 5-year-olds and thirty young adults (15 males and females in each group). Following the guidelines of Farkas (1994), we also made anthropometric measures of a number of orofacial structures. A measure of the range of amplitude and velocity of oral movements was calculated from the entire movement trajectory for two short sentences, and standard peak measures of amplitude and velocity were made from selected single movements within the movement sequences for the sentences. The results of the study clearly indicated that there is not a size principle operating in speech production. There were the expected significant differences between 5-year-olds and young adults in the size of oral structures. While there was a trend for adult speech movement amplitudes to be larger than those of the 5-year-olds, this difference was not significant for the amplitude range measure, nor for the amplitude measures from single movements. There were no differences in movement amplitude between men and women or girls and boys. Furthermore, in within-group analyses, we did not find significant correlations in each age group between the speech kinematic variables and oral structure size measures. The velocity of the 5-year-olds' speech movements was much lower than that of adults, about 50–70 percent of adult values. Thus, we concluded that 5-year-olds' speech movements reflect a large-amplitude, low-velocity style, which would be consistent with a motor control system that requires more time to plan movement sequences and one that has greater reliance on sensory feedback. We can also infer that biomechanical properties of the articulatory system do not account for fundamental differences in speech movement characteristics of young children and adults.

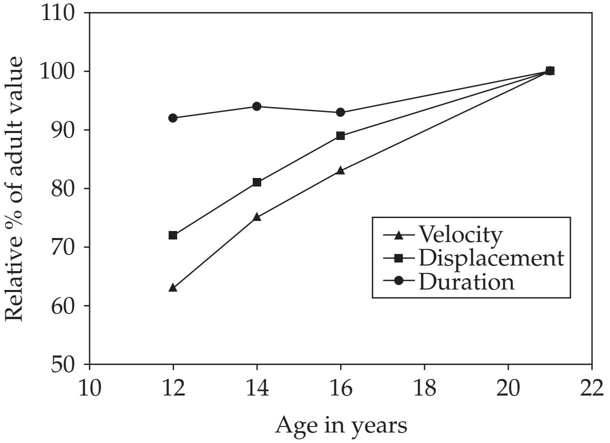


Figure 7.2 The adult values for each measure are arbitrarily set to 100 percent so that we can compare the growth curves of three different variables, duration (average duration computed from two syllables), displacement, and velocity (both averaged across four lower-lip movements). Speech rate, as reflected in the syllable duration measure, is almost adultlike by age 12 years, but the speed and extent of oral movements for speech become adultlike much later.

How long during development do children continue to use this low-velocity, relatively high-amplitude speech movement style? Earlier studies of small numbers of school-age children suggested that speech movements continue to be slow and relatively large during these years (A. Smith & Goffman, 1998), although the data are mixed, with B. Smith and McLean-Muse (1986) reporting that young children and adults had equal amplitudes and velocities of lip and jaw movements for speech. We (Walsh & Smith, 2002) recorded upper-lip, lower-lip, and jaw motion in four subject groups (n = 30/group, 15 females and 15 males in each group) of 12-, 14-, 16-year-olds, and young adults (aged 21–22).[1] Participants produced a six-syllable sentence in a repetition task. We observed significant trends for increasing velocity and displacement of articulator movement beyond 16 years (see Figure 7.2), and there was also a significant decrease in total utterance duration, as well a decrease in the durations of syllables within the sentence with increasing age.

To facilitate comparison of the developmental trajectories for the various measures we made, we plotted normalized growth curves for amplitude, velocity, and duration measures. In these plots, adult values for each measure were arbitrarily assigned a value of 100 percent, and younger subjects' means were plotted as a percentage of the adult value. In Figure 7.2 (adapted from Walsh & A. Smith, 2002) the relative growth curves are plotted for duration, velocity, and displacement of lower-lip (plus jaw) movements for groups of 30 participants aged 12 years to young adult. These measures were taken from two syllables ("Bob" and "pup") produced within a sentence context. The measures were averaged across the opening and closing movements for each of the two syllables. From this plot, it is clear that adolescents have higher speech rates compared to young children, because by age 12 years, durational measures were already 90 percent of the adult value. In contrast, at age 12, velocity of articulatory movement was only about 60 percent of the young adults' value, and movement amplitude was approximately 70 percent of the young adults' value. In real terms, for example, 12-year-olds had a mean velocity of approximately 75 mm/s for the movements we measured, while for young adults, the mean velocity was approximately 120 mm/s. Average amplitude of movement increased approximately 2 mm over the period from 12 years to young adult. These results show that at different points in the course of development, varying trade-offs between speech movement amplitude and velocity occur, and we have hypothesized that these are driven by the goal of increasing speech rate (Walsh & Smith, 2002; Smith & Zelaznik, 2004).

A number of authors have discussed changes in speech motor control processes after age 10–12 years as "refinements" of basic patterns that already have been well established (Green et al., 2002; Cheng et al., 2007). Our study, however, showing as much as 30–40 percent increases in velocity and amplitude of articulatory movements between the ages of 12 and 21 years, suggests that rather dramatic changes are occurring during this late developmental period. Furthermore, the use of the term "refinement" suggests that basic patterns of behavior are the same, and that the developmental curve is a slowly changing trajectory always moving toward the adult state. In contrast to this view, our results suggest that movement amplitude for speech follows a U-shaped developmental trajectory. Speech movement amplitude is relatively large in very young children, decreases in adolescence, and then increases again toward the young adult values. As noted above, one possible factor driving this pattern of amplitude change may be that achieving adult speech rate is a high priority of the speech motor control system, but adolescents apparently cannot produce a higher rate by simply increasing the velocity of articulatory movement as the 21–22-year-olds do. Sixteen-year-olds are still producing significantly lower velocity movements compared to young adults. Thus in order to achieve a higher speech rate, they appear to be reducing articulatory displacements compared to the displacements they produced as younger children. Presumably teenagers age 12–16 years are capable of producing velocities of oral movements in the adult range in nonspeech tasks (but this should be tested empirically), and therefore there is no biomechanical or neuromuscular reason for the reduced velocities of speech in these age groups. This suggests that the lower velocities of speech movements are a result of immature cortical networks involved in language formulation, prespeech planning and/or execution. Obviously the work to date has employed limited speech samples, and future studies will need to replicate these findings in additional utterances and across additional articulators. It should be noted, however, that these recent studies of large numbers of participants have shown rather convincing developmental trends, which likely will require a change in our thinking about speech motor development during late childhood and adolescence.

Our investigation of basic lip and jaw speech movement parameters and that of Cheng et al. (2007) of relative tongue and jaw motion in adolescence also reveal another interesting result: no sex differences have been found. We had predicted, based on anatomical growth curves showing that adolescent girls reach maturity in orofacial structural growth before boys (Farkas, 1994), that girls would show adultlike movement parameters and variability on these articulatory kinematic measures before boys. This was not the case. Cheng et al. (2007) and we (Walsh & Smith, 2002) found no differences between the adolescent girls and boys, nor were there any differences between the young adult men and women. These findings speak to the speculation of B. Smith and McLean-Muse (1986) that the development of adult speech output is delayed by physical growth and by the continued development of the neural systems involved in the formulation and planning of speech. The lack of differences in articulatory kinematic parameters and variability between adolescent boys and girls again suggests that the prolonged developmental course to mature adult speech production systems is primarily driven by the prolonged development of the neural networks involved in cognitive and linguistic processing in the brain (Walsh & Smith, 2002; Smith & Zelaznik, 2004), rather than by peripheral growth factors.

Also relevant to this point is that clear sex differences in articulatory kinematics have been documented in 4- and 5-year-old children (Smith & Zelaznik, 2004). We found that boys lag girls in the consistency of their inter-articulator coordination in the production of short sentences. At this age there is no sexual dimorphism in craniofacial growth patterns (Vorperian et al., 2005) which again points to a role for central rather than peripheral factors driving this difference. These findings are, of course, in contrast to those for speech acoustic, laryngeal, and respiratory measures, because many sex-related differences, which are clearly related to anatomical size and growth factors, have been documented in these output measures (e.g., Hoit et al., 1990; Huber et al., 1999).

3.2.2 Nonuniform maturational profiles across articulatory structures? Another basic question about the development of articulatory movement control for speech is whether control of the various structures involved in speech follows a uniform developmental course. In section 2 above, the pronounced differences in anatomical, biomechanical, and neural innervation of the various articulators were noted. Given the remarkable differences in these characteristics for the jaw, tongue, lips, and velum, it seems reasonable to hypothesize that the development of the control of structures might be nonuniform. In other words, control of one articulator might be more adultlike earlier than control of another. From the frame/content theory (MacNeilage & Davis, 1990, 2000), the prediction would be made that the mandible would show more mature movement patterns earlier than other structures such as the lips and tongue.

Green et al. (2002) addressed the question of sequential development of articulatory control by recording upper-lip, lower-lip, and jaw motion in 1- and 2-year-olds, 6-year-olds, and adults. All groups of speakers produced simple two-syllable utterances such as "mama" and "baba." Green et al. employed an innovative within- and between-group movement pattern analysis which involved time and amplitude normalization of the displacement trajectories for each articulator. The normalized trajectories were then averaged to produce templates, which were compared by computing cross-correlations between pairs of templates on a within-subject basis, as well as within age groups, and finally across the three age groups. These analyses and measures of the variability of movement trajectories within groups clearly demonstrated that in the infants and children, control of the jaw is much more adultlike than control of the lips. Thus these results support the frame/content proposal that jaw open–close cycles provide a basis for the subsequent development of the precise control of all the articulators needed to produce the full repertoire of sounds in the language (MacNeilage & Davis, 1990, 2000).

There is also some evidence from studies of preschool and school-age children that jaw movements for speech are less variable compared to upper lip and lower lip movements (Sharkey & Folkins, 1985; B. Smith, 1995). We (Walsh & Smith, 2002) examined the issue of nonuniform maturation of articulatory control in adolescence. Using a measure of the composite spatial and temporal variability of sets of normalized movement trajectories of upper lip, lower lip, and jaw for a short sentence, we found that movement variability was lower for young adults compared to all of the younger age groups, and that jaw trajectory variability was lower than that of upper lip or lower lip. However, we did not find evidence of a nonuniform rate of maturation across the three articulators during adolescence. The growth curves toward adult performance were parallel for the three structures during the period from 12–22 years, but it should be noted the endpoints for the three articulators were not equal. Composite spatiotemporal variability was lowest for the jaw and highest for the upper lip.

When considering the issue of nonuniform maturational profiles for specific articulators and the possibility that the jaw plays a key role early in prespeech vocalizations, it is essential to keep in mind that the lips have higher degrees of freedom of movement compared to the jaw. Jaw movements for speech are primarily in the vertical dimension and do not occupy a large percentage of the potential working space of the mandible (Ostry et al., 1997). Lip (and tongue) movements and shape goals and the underlying muscle contractions that produce them are extremely complex and multidimensional for speech (Honda et al., 1995; Honda, 1996; Gerard et al., 2003). This may explain why jaw-motion trajectories, from infancy on, display less variability in patterning compared to lip-motion trajectories. In other words, the lower trajectory variability for the jaw may reflect its inherently fewer degrees of freedom and the lower complexity (compared to, for example, the shape requirements of the tongue and lips) of the demands placed upon it for speech.

Another relevant point is that in the face area of the primary motor cortex, which presumably plays a major role in generating the motor commands to control articulatory muscle activity during speech, there is a mosaic of repeated representations of muscles of the lips, jaw, and tongue (Huang et al., 1988). Huang and colleagues reported that on a single electrode penetration, microstimulation typically activates muscles of each of these structures. Given this kind of interleaved representation, which is clearly ideal for the coordinated activities required of the articulators in speech, the idea that the maturation of control of individual articulators follows very distinctive courses seems unlikely to be correct. Perhaps an alternative hypothesis would be that in infants, pre-existing neural circuits, such as those involved in sucking and chewing, generate cyclic open–close jaw movements that provide a stable foundation for prespeech vocalizations. As infants begin to babble, the form of their vocalizations changes to become more speechlike (Kent et al., 1991). It seems reasonable to hypothesize that at this point, cortical networks, possibly associated with syllable-sized units, are being formed. These cortical networks ultimately will be the predominant source of neural control for the speech musculature. This hypothesis is consistent with the data of Moore and Green and their colleagues showing the jaw to be dominant in early vocalizations of 1- and 2-year-olds (Green et al., 2000, 2002). As toddlers begin to produce more speechlike vocal output in babbling and single words, the coordinative patterns of the muscles involved in speech are quite distinctive from those used in sucking or chewing (Moore & Ruark, 1996; Ruark & Moore, 1997). This finding is consistent with the idea that different sources of control (e.g., cortically originating networks) are beginning to be established.

3.2.3 Understanding the sources of variability in articulatory movements and coordination

Higher variability in younger speakers: an epiphenomenon of their slower speech rates? From the earliest studies in which direct measurements of children's articulatory movements were made by Bruce Smith and his colleagues (Smith & Gartenberg, 1984; Smith & McLean-Muse, 1986) and in earlier acoustic studies of children's and adults' speech, the issue of how to interpret differences in variability between child and adult speakers has been debated. One suggestion, repeatedly mentioned, is that the higher variability observed in many measures of children's speech is simply an epiphenomenon (or statistical artifact) arising from children's slower speech rates (B. Smith et al., 1983; Crystal & House, 1988); thus reflecting a general principle that slower speakers tend to be more variable speakers. This parsimonious explanation for differences in variability between immature and mature speakers should be rejected on the basis of results from many recent studies, which have shown clear dissociations between speech rate and variability measures (B. Smith, 1992; Maner, Smith, and Grayson, 2000; Smith & Zelaznik, 2004).

One example of such a dissociation is plotted in Figure 7.3. The data used to generate these plots are from our large-scale study of children, adolescents, and young adults aged 4–22 years (Smith & Zelaznik, 2004). If differences in speech rate fully accounted for differences in speech variability measures, plots of speech rate as a function of age should parallel plots of variability measures. In the example in Figure 7.3, a speech rate measure, the mean duration (and SEM) of two short sentences ("Mommy bakes pot pies" and "Buy Bobby a puppy") for 30 speakers in each age group is plotted (triangles). Also plotted for each age group is a consistency of coordination measure, the lip aperture variability

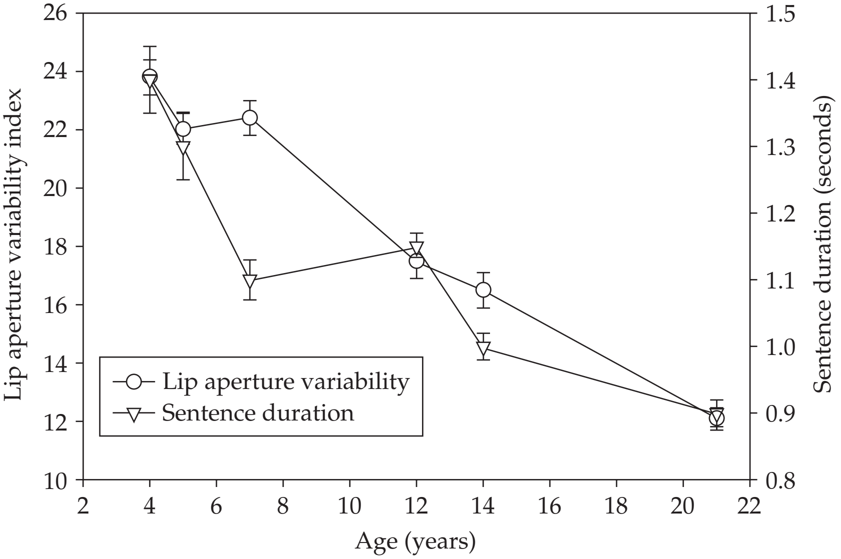


Figure 7.3 The lip aperture variability index (a composite measure of spatial and temporal variability computed for 10 repetitions of a sentence) averaged across two sentences ("Mommy bakes pot pies" and "Buy Bobby a puppy") and the average duration of the two sentences are plotted as a function of age. As children mature, their variability on repeated productions of the sentences and the duration of the sentences drop dramatically (i.e., speech rate increases). The two growth curves show very different slopes in various developmental periods; thus demonstrating that changes in speech movement variability with maturation are not simply an epiphenomenon of increasing speech rates.

index (circles, mean and SEM for each group computed across the two sentences). This index reflects the consistency in the pattern of upper-lip, lower-lip and jaw coordination for 10 productions of each sentence. The two sentences were produced in a repetition task.

From the plot in Figure 7.3, it can be seen that from age 4 years to young adulthood, the average duration of the two sentences decreases from approximately 1.45 seconds to 0.9 seconds. If speech rate in syllables per second is computed from these measures, this dramatic reduction in sentence duration translates into an increase in rate from 3.8 syllables/second in 4-year-olds to 6.1 syllables/second in young adults. Interestingly, there is a plateau in the speech rate function, with no increase in rate (no decrease in average sentence duration) occurring from age 7–12 years. This plateau in speech rate is very surprising, given the dramatic changes in a variety of cognitive abilities, including motor abilities, that occur over this developmental period. Returning to the graph of Figure 7.3, it is clear that oral motor coordination for speech, as reflected by the lip aperture variability index, becomes much more adultlike in the period from 7–12 years. This is a compelling example of a dissociation between variability and speech rate. There are periods of time when the two plots are parallel, for example in the 4- to 5-year-old data, there are parallel increases in speech rate and decreases in variability. However, the nonparallel segments of the growth curves clearly demonstrate that changes in variability cannot be explained as an epiphenomenon or statistical artifact of overall speech rate. As a caveat, we note that these data are derived from a repetition task for two short sentences, and that as such, they may not reflect naturally produced, spontaneous speech. This, however, is a necessary limitation of any study in which physiological or acoustic parameters are measured for identical utterances produced by different speakers. In addition, we note that this variability measure is based on the average of coordination indices computed for the entire movement sequences for 10 repetitions of each of two sentences, which would seem to be an improvement over earlier studies, in which kinematic measures typically were made for single movements, and duration or rate measures were often reported for single words, and in some cases single speech segments. In fact, we have made extensive use of the method of time- and amplitude-normalizing sets of single articulator or inter-articulator trajectories produced for a single utterance over multiple repetitions (method described in Smith, et al., 1995 and 2000). An index of the composite spatial and temporal variability computed for these sets of trajectories has proved to be a useful indicator of within-subject and between-group differences in speech motor performance.

In the above example, the question was whether differences in speech movement variability between groups of speakers of varying ages could be accounted for by differences between groups in average speech rates. One can also ask whether within a given age group, the slower speakers tend to be more variable in output. To address this question, correlations between the average duration and the average lip aperture variability index for the two sentences were computed for each of the six age groups whose data were plotted in Figure 7.3. Scatter plots of lip aperture variability and duration are shown for two age groups, the 12-year-olds and 4-year-olds, in Figure 7.4. As these plots suggest, there was not a significant correlation within groups between mean sentence duration and mean lip aperture variability. Correlations between speech rate (sentence duration) ranged from a low of 0.02 for the 4-year-olds to the highest value of −0.23 for the 14-year-old group. The correlation between the two measures was not significant for any of the age groups.

Decreasing movement variability with age: an index of neuromotor maturation? Having, hopefully, helped to put to rest the assertion that developmental changes in measures of speech variability simply reflect changes in speech rate, I return to the main issue at hand, which is: many investigators, including the present author, have interpreted the greater variability often observed in children's data to be a sign of the operation of immature motor control systems. Thus, the general phenomenon of decreasing variability observed with age, described in many studies in the sections above, commonly has been interpreted as a sign of the maturation of the speech motor control systems. Stathopoulos (1995) strongly objected to this

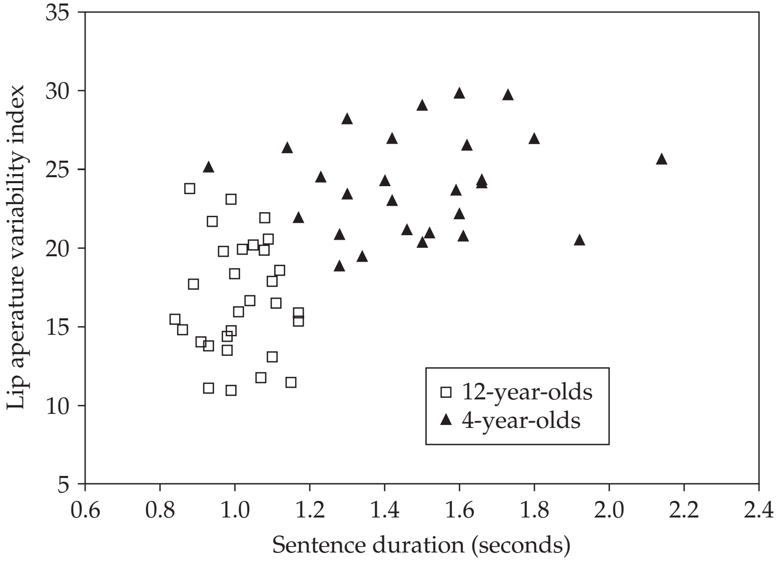


Figure 7.4 Scatter plots of a speech coordination variability index, lip aperture variability, computed for 10 repetitions of a sentence (averaged across two short sentences) plotted as a function of sentence duration (again averaged across the two sentences) for thirty 4-year-olds and thirty 12-year-olds. It is apparent that there is no correlation between these two measures. Thus within an age group, more variable speakers do not tend to be the slower speakers of the group. Note the large range in sentence duration, from about 1.0–2.2 seconds in the 4-year-olds and the much smaller range in sentence duration in the 12-year-olds, about 0.8–1.2 seconds.

general interpretation of decreasing variability with increasing age; she presented acoustic, aerodynamic, and respiratory kinematic data from 72 participants ranging in age from 4 years to adults. She reported a very large number of physiological and acoustic measures and found that significant age differences were present for only a subset of the measures. Furthermore, within that subset, statistical analysis revealed that the 4-year-olds primarily accounted for the age effect. Stathopoulos argued that, given the generally accepted view that variability is an index of neuromotor maturation, she would have to conclude that 6-year-olds possess adultlike speech production systems. She therefore rejected the idea that declining variability is always or even usually a hallmark of maturation toward the adult state and instead suggested that when studying variability as an index of maturation, different subcomponents of the system may show very different maturational profiles. It should be noted, however, that the experimental data on which Stathopoulos based her argument were derived solely from measurements made on repeated trains of a single syllable (/pa, pa, pa/). Therefore one might argue that she studied a "speech" production task that would be least likely to reveal age-related differences.

In any case, Stathopoulos made a useful theoretical argument, and as pointed out above, significant differences in developmental trajectories have been reported for the different subsystems involved in speech and for different measures within a single subsystem. When studying variability in speech output as an index of neuromotor maturation, observed between- or within-group differences in variability must be placed within the appropriate context. Regardless of these points, declining variability and increasing accuracy classically have been viewed as hallmarks of maturing motor systems and of successful motor learning in the limb motor control literature (Schmidt, 1988). The influential work on the application of dynamical systems theory to motor development by Esther Thelen and her colleagues (Thelen & Smith, 1994) has also emphasized the necessity of contextualizing interpretations of higher or lower levels of variability. For example, higher variability often has been viewed as an undesirable feature of motor control systems, but as Thelen and Smith pointed out, over the course of development, variability may provide a flexible substrate from which new patterns of behavior may emerge. Furthermore, variability is not a unidimensional construct, and as we will discuss in more detail below, many sources of variability contribute to the observed output variations. Recent models and experimental approaches to motor development in the limb literature have emphasized the need to uncover the relative importance of multiple contributing sources to movement output variability over the course of development (Davids et al., 2006).

This is the point of view that we have taken in interpreting our kinematic studies of the development of articulatory control and coordination (Walsh & Smith, 2002; Smith & Zelaznik, 2004). It would not be desirable for children to produce speech movements with the almost machinelike consistency of adults. Children need flexibly organized motor control systems, so that they can acquire new patterns, e.g., new words and/or new languages. In addition, while we have argued that peripheral biomechanics are not the only factors driving the prolonged maturation of speech motor control processes in adolescence, certainly children's speech motor control systems must make adaptations as craniofacial growth occurs (Vorperian et al., 2005).

To interpret the higher variability we have observed in children and adolescents, we (Smith & Zelaznik, 2004) have also relied upon Bernstein's definition of the development of motor coordination as "the process of mastering redundant degrees of freedom of the moving organ, in other words its conversion to a controllable system" (1967, p. 127). He proposed that the degrees of freedom for movement are reduced through the soft assembly of muscle synergies, also referred to as coordinative and/or functional synergies. Functional synergies (the term we have used) are fundamental units of the control of movement, and they consist of collectives of muscles or motorneurons that in turn control muscle contraction (Bernstein, 1967; Gelfand et al., 1971). In speech, given the complexity of the movements to be produced and the necessity of recruiting specific subpopulations of the motorneurons within a motorneuron pool (Honda et al., 1995), the idea of motorneurons (or motor units – a motor unit is a single motorneuron plus the muscle fibers that it innervates) rather than muscles as the basic elements constituting functional synergies is appealing. The repeated coactivation of collectives of motor units results in the formation of synergies, which are organized to achieve functional goals. For example in babbling, motor units of upper-lip, lower-lip, and jaw muscles would be repeatedly coactivated, for example, for the syllable /ba/. With repeated activation, the synergistic group of motor units would be linked to the goal of producing the acoustic output for /ba/. As children mature, these functional synergies become more stable, and there is less variability in the pattern of recruitment of motor units to achieve the behavioral goal.

In the developmental limb motor control literature, many investigators (e.g., Crossman & Szafran, 1956; Schmidt et al., 1979; Van Galen et al., 1993) have suggested that one source of variability in the recruitment of motorneurons is a global factor, "neuromotor noise." Neuromotor noise is postulated to arise from a variety of sources, including a background of unpatterned synaptic inputs to motorneuron pools and from the motor commands generated to achieve movement goals. Both sources of neuromotor noise, the background, unpatterned synaptic inputs and the variability of motor commands generated by the CNS, are hypothesized to be greater in young children. With maturation, neuromotor noise is hypothesized to decrease, which contributes to the increased consistency of motor output seen in adults (Smits-Engelsman & Van Galen, 1997; Yan et al., 2000).

The nature and sources of variability in motor output have been extensively discussed in the limb motor control literature (Davids et al., 2006). With regard to motor development, another important perspective is that sources of movement output variability operate over different time scales (Newell et al., 2001). Neuromotor noise is hypothesized to operate over a long developmental time scale, such that the level of variability contributed by neuromotor noise is relatively constant on a day-to-day or even week-to-week basis. In other words, 7-year-olds on average will show higher movement variability than 12-year-olds. Other sources of variability operate over very short time scales, for example, within a single experimental session (Newell et al., 2001, 2006; Deutsch & Newell, 2004). Furthermore, during development the operation of the distinct sources of variability may change. For example, Deutsch and Newell (2004) demonstrated that children can exhibit short-term improvements in motor performance, becoming more accurate and faster within a single experimental session. Age-related differences in short-term changes in movement output variability are hypothesized to reflect differences in the way children and adults use a variety of feedback loops and potential differences in the way the systems' degrees of freedom are controlled to achieve movement goals (Newell et al., 2006). Finally, these investigators have also used spectral analysis to explore the possibility that the structure of movement variability arises from both stochastic and deterministic processes (Deutsch & Newell, 2003).

These issues are just beginning to be addressed within the developmental speech motor control literature. To our knowledge, our laboratory was the first to explore experimentally the possibility that short-term motor learning effects would be observed in a speech production task in children and adults. We (Walsh et al., 2006) assessed the potential role of short-term plasticity by examining performance on a novel nonword learning task. Learning effects were measured by computing the consistency in coordination over repeated productions of a higher-level (lip aperture) and a lower-level (lower-lip–jaw) functional synergy. This experimental approach was derived from our earlier study (Smith & Zelaznik, 2004), in which we hypothesized that lip aperture is a higher-order synergy compared to the lower-lip–jaw synergy, because lip aperture control has important acoustic effects, while the relative lower-lip–jaw motions do not. This hypothesis was supported: for all age groups studied, the higher-order synergy showed less variability across repeated sentence productions compared to the lower-order synergy. This was true, despite the fact that the lip aperture synergy involves the relative motions of the upper lip, lower lip, and jaw, while the lower-lip–jaw synergy involves only two articulators. (To understand the idea of higher- and lower-level functional synergies, consider the analogy of clapping your hands 10 times. For each clap cycle, we plot the trajectory over time of the inter-hand distance, which is analogous to lip aperture, and we plot the trajectory of relative motion of the right wrist and elbow, alogous to lower lip–jaw. We would expect that the inter-hand difference trajectories would be much more consistent from cycle the cycle than the within-arm wrist–elbow difference trajectory.)

In the novel nonword learning task, participants heard novel nonwords in random order and were instructed to repeat the word after hearing it. There were five novel nonwords and they ranged from one (“mab”) to four syllables (“mabshaytiedoib”). With increasing length, phonological complexity of the nonwords also increased. The analysis was designed to determine whether the early repetitions of the novel nonwords were more variable in inter-articulator coordination compared to the later repetitions. For each of 10 repetitions of each word, the lip aperture and the lower-lip–jaw difference signals were computed. The sets of early (first five) and late (last five) trajectories were compared on variability and duration. All participants (twenty young adults and twenty 9–10-year-olds) correctly produced all of the novel nonwords.

Our results confirmed that speakers of different ages have different speech motor learning characteristics. The young adults showed no changes in lip aperture or lower-lip–jaw coordinative patterns over the course of experiment. Variability for both upper- and lower-level synergies and nonword duration did not change over early and late productions for young adults. As predicted, the higher-order synergy showed less variability across all nonwords in both groups. Unlike the adults, the 9- and 10-year-olds showed a pronounced practice effect within the experimental session. Coordination variability for the lip aperture signal was significantly lower for the children’s last five productions. This effect was dramatic and was most pronounced for the longer and more complex nonwords. A parallel result was observed with regard to duration of the nonword productions for the children: the duration of the last five trials was significantly shorter than the duration of the first five. These results demonstrate that young children learning novel words show rapid and dramatic decreases in movement variability and increases in the speed of the sequential movements necessary for articulating the novel nonwords. In this case, children were able to simultaneously improve in consistency of motor execution, while speeding up the execution process. Interestingly, with regard to our hypothesized higher- and lower-order synergies, the motor learning effect was observed only for the higher-order synergy, lip aperture. Lower-lip–jaw coordination, the lower synergy, did not become more consistent over the early to late trials. This result provides additional evidence that lip aperture is a higher-order synergy, and that lip–jaw coordination is adjusted to more consistently achieve a higher order-control variable, that is, the distance between the lips.

The results of our experiment support a role of neuromotor noise, which operates over a relatively long time scale, in speech production. The 9- and 10-year-olds were more variable than the young adults on all measures, and their level of performance on the improved later trials did not reach adult values. This suggests that there are sources of variability that, even with practice, prevent younger speakers from attaining adultlike performance levels. The question could be raised whether, if given enough practice trials, the children would reach adult levels of consistency and speed. We suggested that this is unlikely, because even when children and adults repeat familiar well-practiced utterances, such as "Mommy bakes pot pies," children are more variable and slower speakers. Our results also support the idea that there are sources of movement output variability that operate on shorter time scales compared to neuromotor noise. The improvement the children showed in the consistency of coordination of the upper lip, lower lip, and jaw and in the rate of nonword production occurred over the 30-minute experimental session. The five nonwords were randomized, so this was not simply a result of an improvement over repeated, sequential productions of the words. We would suggest that this improvement reflected systematic changes in cortically originating motor commands to the motorneuron pools. With just five practice trials, children were already becoming more consistent in generating the motor commands necessary to produce this novel sequence. These observations are consistent with the model proposed by Newell and his colleagues, suggesting that there are sources of movement variability that can be observed to change over short time scales (Newell et al., 2001).

Another interesting issue that arises from the study of short-term motor learning in speech is whether adults would show short-term motor learning effects if they produced nonword stimuli that were more difficult. We are addressing this issue in a follow-up study (Sasisekaran et al., in press), and the results indicate that similar short-term motor learning effects are present in young adults when the novel nonwords are longer and phonologically more complex. It seems likely that the stimuli in our first study were easy, such that adults were at ceiling in the early trials, and therefore showed no improvement from early to late trials. Another interesting question concerning these short-term changes in speech production performance within a single experimental session is whether they represent speech motor learning. In other words, have changes in synaptic connections occurred, such that on retesting the next day, participants would retain the improved performance observed on the later trials of the day 1 testing. In our second experiment, participants returned for second-day testing, and it is clear that speech motor learning does occur. The early trials of participants, both 9–10-year-olds and young adults, on the second day show greater consistency compared to their early trials on day 1.

In summary, the recent literature on speech and limb motor development reveal that movement trajectory or inter-effector coordinative variability typically decreases as humans mature. Variability is not unidimensional and must be interpreted within the context appropriate for the task under study and the ages of the subjects. The time courses over which reductions in variability occur, short- and longer-term, can reveal many significant aspects of the underlying maturational processes. Recent studies demonstrating the extremely prolonged developmental course to adult levels of speech motor control and coordination are intriguing, and the relationship of this developmental trajectory to the growth curves characteristic of other skills, e.g., language processing abilities, will be significant areas for future investigation.

3.3 Theoretical issues and models of speech motor development

There is general agreement among those who have written about speech motor development that the process involves the formation of neural mappings among motor, somatosensory, and auditory systems (Kent et al., 1991; Callan et al., 2000; Smith & Goffman, 2004; Guenther, 2006; Smith, 2006). The earliest speechlike vocalization of infants is babbling, in which the canonical syllable appears, followed by repetitive canonical babble and variegated canonical babble (Oller, 1980; Stark, 1980; Kent & Bauer, 1985). Kent et al. (1991) proposed a model of early vocal development in which they applied Edelman's theory of neural group selection (Edelman, 1987, 1989) to postulate how these mappings might be generated. They suggested that the production of even a simple sound, such as a syllable, would activate a variety of sensory "receptor sheets," including static and dynamic intraoral mechanoreceptors, pressure and flow receptors, and auditory pathways. For each sensory modality, a sensory map would be formed, and with repetition of the syllable, the various sensory maps would be correlated with one another and with the motor map that produced the behavior. As we noted earlier, "neurons that fire together, wire together" – thus neural connections would develop to form these functionally linked maps arising from the infant's vocal behavior. Kent et al. suggested that, in addition, sounds made by others, such as parents, are also represented in the auditory receptor sheets and associated map. "Re-entrant" or repetitive signaling of this type would ultimately lead to the establishment of phonetic categories, which would be defined, not by the specific sensory information generated by the category of behavior, but by the correlations among the various maps. This kind of model, the authors noted, avoids the problem of postulating invariant motor or sensory representations of speech sounds. This is an important feature of any model attempting to account for developing or adult speech motor control, because the ubiquity of variability is a cardinal feature of speech production (e.g., MacNeilage, 1970).

Written in 1991, the model of early vocal development proposed by Kent and his colleagues presaged later, more formal neural network models of speech. Guenther (1995) proposed the DIVA (D = directions in orosensory space, I = into, V = velocities of, A = articulators) model for speech sound acquisition. Like the Kent et al. model, the DIVA model posits that babbling is an action–perception cycle, and that with repetition, cyclic babbling behaviors tune the speech production system by establishing mappings among reference frames (orosensory, acoustic, and motor). Like the speech production theory of Perkell and his colleagues, which focuses on adult speech motor contol (Perkell et al., 1995, 2000), the DIVA model posits a major role for auditory targets, and the phoneme is the basic unit of production. After the babbling phase, the DIVA model can produce phoneme strings entered as input by the user (Guenther, 1995).

In our discussion of linguistic units and models of speech motor development (Smith & Goffman, 2004; Smith, 2006), we proposed a preliminary model of speech motor development which also included the idea that mappings between various neural systems must occur during speech acquisition. In addition to orosensory, acoustic, and motor linkages, however, we proposed that bidirectional linkages from motor to linguistic representations must be formed as toddlers begin to produce first, words, then longer utterances. In general then, there appears to be agreement that speech motor development entails the establishment of a variety of connections among the various neural centers involved in language formulation, speech motor control, and sensory representations. Beyond this basic premise, the speech motor development literature is extremely diverse in the theories, models, and/or frameworks used to generate experimental questions and to discuss the resulting data. Aside from the DIVA model, there are no formal models that attempt to elucidate the course of speech motor development. Therefore, in the sections that follow, I consider critical issues likely to be relevant to future theoretical approaches to speech motor development.

3.3.1 Units of production: the language–motor interface There is one period during human development when the basic unit of speech production seems clear. When the infant begins to babble, the unit of production is clearly the syllable (in my view, but note that the DIVA model of Guenther originally assumed phonemes as the input unit during the babbling stage, though later descriptions of the DIVA model indicate the input is a "speech sound," which can be a phoneme, a syllable, or a word (Guenther, 2006)). One can return to the chapter by Kent and his colleagues (1991) for an insightful discussion of the role of the syllable in infant vocal development. They note that all units of speech production, including syllables, "present interpretive difficulties across and within levels of observation. A major factor in these difficulties is the attempt to impose segmentation on what is often a continuous motor pattern. But of all the candidates for behavioral analysis of vocal development, the syllable appears to be the most practical and the most commonly used" (Kent et al., 1991, p. 136). They consider the syllable to be the fundamental unit of early infant vocalization, and that the generation of sequences of syllables gives rise to the rhythmic structure of early infant vocalizations. Within this context, they defined the syllable as "a grouping of motor adjustments, highly variable in composition from one syllable to another, that is associated with the auditory perception of the fundamental prosodic unit" (p. 137). In the babbling stage then, the unit that serves as a basis for the mappings among the oral sensory, auditory, and motor systems is proposed to be the syllable.

We have hypothesized (Smith & Goffman, 2004; Smith 2006) that the nature of mappings between linguistic, auditory, and motor networks changes over the course of development. As suggested above, the syllable is the most likely unit of babbling. As the toddler enters the single word and multiword period of development, we proposed that multiple units emerge as the basis for neural mappings. In the toddler, syllables, words, and word combinations would serve as bases for mappings. In 4- and 5-year-olds, we proposed that phonemes, syllables, words, and phrases would serve as units of interface among the systems. In adults, we hypothesized multilayered mappings between linguistic units and the motor system. In other words, there is no privileged unit of production in the adult system, and as the child matures, he or she acquires these multilayered mappings.

The data to support this proposal comes from experiments from our laboratories and others to suggest that motor output is intimately shaped by the linguistic goals of the speaker. We have examined the relationship between motor output and linguistic units for many different sizes of units, including the phoneme, syllable, word, phrase, and sentence levels (reviewed in Smith & Goffman, 2004; Smith, 2006). Goffman and her colleagues have also demonstrated clear effects of prosodic goals on the details of speech movement output (Goffman, 1999; Goffman et al., 2006, 2007). One compelling example, which supports the idea of the parallel operation of many units of production, comes from our study of coarticulation in 5-year-old children and adults (Goffman et al., submitted). Participants produced three pairs of sentences (10–15 repetitions of each) that varied only in an utterance internal vowel (e.g., "Mom has the *goose/geese* in the box"). We measured the timing and amplitude of the lip rounding gesture for /u/ relative to the duration of the lip movement sequence for the entire utterance. For both children and adults, the lip rounding gesture had broad influences on the lip movement sequence for the entire sentence, with the rounding gesture continuing for 50–60 percent of the sentence duration. Adults showed less variable rounding gestures, but the influence on the entire utterance was similar across age groups.

Returning to Figure 7.1, we note that in order to produce the sentence, "Mom put the goose in the box," the brain has to generate motor commands to activate the appropriate muscles with exquisite control of the timing of their activation. What this experiment reveals is that the neural commands to the lip muscles involved in the rounding gesture are modified across a large portion of the sentence in relation to the identity of a single vowel in the middle of the sentence. Adults and 5-year-olds make a similar modification in terms of the temporal organization of the gesture. Thus these findings suggest that by 5 years of age, children are already using multiunit speech production planning strategies. In order to produce this long-lasting change in the motor commands for the sentence specifically attached to a single segment, the speaker would need to have at least a phrase-level motor plan. This result, in combination with an earlier study (Goffman & Smith, 1999) showing that children as young as 4 years produce phonetically specific oral movement patterns for consonants that vary only by a single feature (e.g., "ban"/"pan"), supports the claim that by 4–5 years, children are using multiple planning units in speech production.

Our multiunit view may seem at odds with models of speech production that propose a privileged or basic unit of production planning, often the phoneme (e.g., Perkell et al., 2000), or the syllable, for example, the syllabary of Levelt and Wheeldon (1994). On the other hand, the suggestion that some units may be more prominent, "more fundamental" than others does seem reasonable. In my view the syllable is a good candidate as the "most basic" unit, because of its connection to the open–close oral movement cycle and its appearance as the first speech-like vocalization of infants. In any case, from the point of view we have taken (Smith & Goffman, 2004; Smith, 2006), as the child matures, stored commands can be of varying lengths, from syllable, to word, to phrase, to sentence length. I am certain that many Americans have extremely stable, stored motor commands for the phrases, such as "Hi, how are you," and "Have a nice day."

In summary, future theoretical work on speech motor development must address the gap between language production models that ignore the motor system and speech motor control models that ignore the language system. This is a difficult gap to bridge, but by studying the unfolding relationship between motor output and the linguistic and metalinguistic goals of the speaker as children mature, future theorists will have a better chance to understand how linguistic constructs are transformed into muscle contraction and movement.

3.3.2 Factors driving the protracted developmental course to mature speech motor control A comprehensive account of speech motor development must consider the long developmental trajectory for attaining adultlike speech motor control processes. In addition to the protracted developmental course for speech motor control, some periods in development are marked by very rapid changes toward the adult end product, while in other intervals, for example, the period from 7–12 years, plateaus in some aspects of performance are observed. It seems likely that in various developmental periods, the factors that drive either rapid or relatively slow speech motor development vary.

Compared to girls, 4- and 5-year-old boys are less mature in the consistency of inter-articulator coordination (Smith & Zelaznik, 2004). This sex difference disappears by age 7, and in terms of consistency of coordination and speech rate, we observed no sex differences in any of the other age groups we studied. The fact that preschool boys lag girls in articulatory coordinative development is not surprising, given that girls are often shown to be better in verbal tasks, but the reasons for this difference in development are unknown. In the burgeoning neuroimaging literature, in which large groups of children are being imaged in cross-sectional and longitudinal studies, many sex differences in brain development have been documented. These include different developmental trajectories for global measures, such as white and gray matter volumes, and differences in specific regional measures (Lenroot & Geidd, 2006). In their study of 200 normal children, Wilke et al. (2007) reported that girls have a proportionally higher gray matter volume in a very distinct area in the left inferior frontal gyrus, a difference not observed for the right homologous area. This finding is consistent with a number of anatomical and functional imaging studies pointing to neurophysiological bases for sex differences in verbal tasks (Harasty et al., 1997; Plante et al., 2006).

The idea that speech motor control processes continue to mature post puberty would have been surprising some years ago. Now it is widely recognized that adolescence is a time of very significant development behaviorally and cognitively, and there is clear evidence that brain development continues well into the twenties (see review by Blakemore & Choudhury, 2006). For speech motor control, it is important to note that the frontal lobe continues to show increased myelinization throughout adolescence, which would contribute to faster conduction speeds allowing more rapid inter-regional communication among neural populations (Giedd et al., 1996, 1999). Furthermore, there is an increase in white matter in the left arcuate fasciculus during adolescence, and the corpus callosum undergoes regionally specific changes until the mid-twenties (& and Choudhury, 2006). Given these results, it is not surprising that 16-year-olds are not yet adultlike in speech motor control processes. We suggested that there are trade-offs during adolescence, such that higher rates of speech are achieved at the expense of more variable coordinative patterns. Also, we noted that given that girls and boys do not differ in articulatory motor control during the adolescent years, the protracted course of articlulatory motor development apparently is not related to craniofacial growth. Rather, we hypothesize that the continued maturation of the brain is a primary factor delaying maturation of speech motor control processes to adult levels.

Another interesting issue that future models of developing speech motor control must address is individual differences. Much of the focus of this chapter has been on between-group effects, and the developmental growth curves under discussion reflect changes in group means and variability with maturation. As Figure 7.3 shows, however, there are dramatic differences between individuals within an age group as is evident for the thirty 4-year-olds and thirty 12-year-olds whose data are plotted. Some of the 12-year-olds have coordinative consistency (as measured by the lip aperture variability index) equal to adult levels (in the 10–15 range). Some 12-year-olds, however, have coordinative consistency indices equal to that of some of the 4-year-olds (19–25 range). The 4-year-olds have a remarkable dispersion of speech rate. What accounts for these individual differences? Future studies in which measures of speech motor control, language abilities, and general cognitive abilities are obtained in addition to neuroimaging data for the same subjects will be necessary to explore the potential factors contributing to these differences between individuals in speech motor performance.

3.3.3 The role of feedback and the nature of stored motor commands There is general agreement that auditory and somatosensory feedback play a critical role in the development of speech motor control processes (e.g., Kent et al. 1991; Guenther, 1995), and it is well known that normal speech production patterns cannot be established in the absence of auditory information. Furthermore, there is also general agreement that with maturation, speech motor control becomes primarily under feedforward control, that is, driven by stored neural commands for speech movement sequences. The nature of what is stored and how these stored motor commands relate to linguistic units is, as noted above, a matter of debate. Despite the general agreement about the importance of sensory information in developing speech motor control, almost nothing is known about the details of how and when somatosensory and/or auditory information is used to shape ongoing motor output, to build internal models of movement goals, and to tune feedforward commands.

In general, speech production systems of children are slower, and it has been noted that this is consistent with a feedback-based control system (e.g., Riely & Smith, 2003). As children mature, they become faster speakers and their movement output patterns become much more consistent from trial to trial. Presumably, during maturation children are establishing stored motor commands for speech production, and they are relying less and less on feedback. In order to understand how this process of shifting from feedback to feedforward control unfolds in development, experiments are needed to manipulate feedback and examine the effects on motor output. Adults can compensate automatically online to mechanical perturbation of the articulators as well as to alterations and auditory feedback (reviewed in A. Smith, 1992). While a few preliminary studies of children's responses to bite block perturbations have been completed using acoustic analyses, the results of the studies are mixed (Baum & Katz, 1988; Edwards, 1992). It would be useful in future studies to examine the effects of altered somatosensory and auditory feedback on the variability of movement output in children at different ages. In addition, studies of novel word learning could incorporate altered feedback conditions and address the question of the role of sensory information in establishing new patterns of output.

3.3.4 Neuroplasticity and sensitive periods for speech motor learning A primary source of evidence for the existence of sensitive periods for speech motor learning comes from the well-known fact that the ability of humans to learn new languages and to achieve near-native accents in them decreases as we mature into adulthood. In general the loss of capacity to acquire new languages with aging has been attributed to a sensitive or critical period for language acquisition, such that with maturation there is a loss of plasticity in the neural systems involved in language learning (see references cited by Flege et al., 1999). While this explanation seems intuitively appealing, the issue of age-related changes in second language (L2) performance is a complicated one. The sensitive period hypotheses for L2 performance changes have not been specific with regard to what particular abilities are lost (Flege et al., 1999). For example, the inability to achieve a near-native accent in L2 could be due to a loss of plasticity in speech motor output circuitry, a decrement in the ability to auditorily discriminate L2 sounds, a loss of ability to create new perceptual representations of sounds, and/or a decrement in the ability to translate stored auditory representations into speech motor commands (Flege et al., 1999). Flege and colleagues have completed many experiments designed to test the sensitive (or critical) period hypothesis, and with regard to accent, the data do not support a strictly maturational explanation. They therefore prefer an alternative explanation for the age-related decline in L2 pronunciation accuracy, which is that the greater the stability of the first language (L1) phonetic system, the more interference L1 has on L2 learning. As children mature into adolescence and young adulthood, the L1 phonetic system becomes extremely stable, and new patterns of phonetic output are more difficult to achieve. They also note that the difficulties in acquisition of L2 morphosyntax may be affected by a different set of factors compared to those that determine accent.

Perhaps, if one considers the sensitive period hypothesis within the framework proposed by Knudsen (2004), some of the issues suggested above become clearer, and one could propose the combined operation of both sensitive periods and L1 interference in the age-related decline in L2 performance. Knudsen notes that while we tend to think of sensitive periods in terms of behavior, they are actually properties of neural circuits. In his enlightening article, Knudsen provides a variety of examples of sensitive periods in development, including those of human language, birdsong, visual representation in monkeys, filial imprinting in ducks and chickens, and auditory processing of spatial information in owls. He defines the term sensitive period as one that "applies whenever the effects of experience on the brain are unusually strong during a limited period in development" (p. 1412). In some of the animal models mentioned above, changes in the neural circuitry that arise during sensitive periods, in other words the changes that underpin learning, have been mapped out. These involve axonal elaboration and synapse formation in addition to axon and synapse elimination. The metaphor Knudson uses to explain stable neural circuits is relevant for the present discussion of speech motor learning. He invokes a stability landscape in which experience shapes troughs or wells of stable behaviors over development (note the similarity to the stability landscapes of Thelen & Smith, 1994). Once a highly stable neural circuit has been formed, there is a loss of plasticity in that circuit. "After a sensitive period has ended, many independent mechanisms that support plasticity continue to operate. The amount of plasticity that persists in a mature circuit varies widely, depending on the circuit's function. The plasticity that remains enables mature circuits to modify their patterns of connectivity within the enduring constraints established as a result of experience during a sensitive period" (Knudsen, 2004, p. 1417). Adults have passed the sensitive period for speech motor learning, but they retain some degree of plasticity. Applying Knudsen's framework, learning new behaviors, such as a second language, requires more attention and effort after the sensitive period has passed, and the new behavior may be atypical (e.g., retention of an L1 accent) due to influences of previously established neural circuitry.

Also relevant to the observation that accent and morphosyntax may show different age-related performance trajectories in L2 learning (Flege et al., 1999) is Knudsen's observation that complex behaviors such as language result from the interactions of multiple hierarchies of neural circuits. Therefore, he suggests it will be difficult to identify critical or sensitive periods based on behavioral measures. As discussed above, widely distributed neural circuits contribute to language production, including cognitive, linguistic, motor, and sensory circuits. The different circuits are likely to undergo different developmental trajectories and thus display distinctive sensitive periods. He notes that while the neural circuitry involved in semantic analysis remains highly plastic throughout life, the L2 data strongly suggest that the neural hierarchies involved in phonetic processing lose a great deal of plasticity with maturation to adulthood. As another example, Weber-Fox and Neville (1996) concluded from studies of event-related brain potentials recorded in language processing tasks that there is a sensitive period for acquiring the syntax of a language. These ideas are important to bring to bear on interpreting data in speech motor control experiments. For example, in a recent study we hypothesized that speech coordination variability would be higher for L2 compared to L1 in bilingual Bengali–English speakers. At first, we were surprised to find that movement variability assessed in L1 (Bengali) and L2 (English) was equal and independent of age of immersion in L2 (Chakraborty et al., 2008). Despite the fact that many of the speakers had a pronounced accent in L2, their movement variability in L2 was low and typical of adult levels of performance in L1 for native speakers. This suggests that the strongly accented L2 production is also highly stable, and that the neural circuitry underlying it has lost plasticity. Furthermore, there are strong interference effects on L1 of L2.

The idea of sensitive periods for speech motor learning is also supported by the experience of young cochlear implant recipients in learning to produce speech. The younger the recipient, the more likely he or she is to produce speech that is highly intelligible, and after age 4–5 years, high degrees of intelligibility typically are not attained, even after many years of implant use (Peng et al., 2004). Ertmer et al. (2007) assessed vocal development in a prospective longitudinal study of seven children who received cochlear implants between 10 and 36 months of age. These infants and toddlers, for the most part, had passed the age at which babbling begins when they received their implant; however most of them proceeded from babbling through the normal stages of vocal development. Interestingly, though, relative to the sensitive period hypothesis, the older implant recipients in this study (30 and 36 months) achieved all the milestones of complete vocal development in the shortest time frame. Apparently, these older toddlers were still young enough to take maximum advantage of the new input, and they did so in a very rapid manner given their more mature overall cognitive developmental levels. Taken together, the results of these two studies suggest that there is a sensitive period which extends through 3 to 4 years of age when optimal gains in speech production abilities can be made in response to auditory inputs. As our technical ability to record movement and muscle activity during early vocal behavior improves, our understanding of the operation of sensitive periods for speech motor development should expand.

3.3.5 Other theoretical issues and conclusion There are many other issues that will be relevant for future modelers of speech motor development. Many of these have been discussed at some length in earlier sections of this chapter, for example, the need to understand the sources of variability in speech movement output and their differing time scales of operation, and the issue of the relationship between the neural systems that generate speech movements and those that generate other, earlier appearing motor behaviors using the same output pathways.

As a final note in this discussion of future theories of speech motor development, I would like to point out that one of the most astonishing conclusions one reaches after completing a review of this literature is that there is a real paucity of studies of oral motor development for speech. There are very few laboratories doing work in this area, and this is surprising given the importance of normal speech development in human experience. In contrast, there are many more investigators studying the development of a variety of limb movements, from gait, to finger tapping, to precision and power grip. The recent literature on power grip, in fact, would serve as a good model of what the future might hold for the study of the development of speech motor processes. My summary of the research on power grip is based on the review provided in Halder et al. (2007).

Early behavioral studies of the development of power grip performance under visual feedback revealed that younger participants were slower in reaction time to produce a target force, slower in rise times to the target, and more variable in achieving the target. The developmental course of the increasing abilities of young children and adolescents to produce power grip was mapped out in many studies of children from 3 years through adolescence. A series of functional imaging studies completed on adults in the period from the late 1990s to the early 2000s demonstrated the extensive neural network involved in producing power grip under visual control. It involves the contralateral primary sensorimotor cortex, the ipsilateral cerebellum, the superior parietal cortex, the ventrolateral thalamus, occipital, and premotor regions. In addition to mapping the network involved in visually guided power grip, imaging studies also confirmed a linear relationship between activity in motor regions and the level of the target force. Electroencephalographic (EEG) studies were completed to examine movement related potentials, which were also found to be sensitive to force parameters.

These studies provided the experimental foundation for the 2007 developmental study of Halder and colleagues, a large Swiss team. In this study power grip performance was studied in 17 participants in each of three age groups (9–11 years, 15–17 years, and young adults) in both functional magnetic resonance imaging and EEG recording sessions. Thus this team had detailed motor output measures (e.g., reaction time, target force achieved, rising slope), excellent spatial resolution of the activated neural networks from the fMRI data, and excellent temporal resolution of the neural activation patterns preceding and during performance of the power grip task from the EEG data. Their developmental study replicated the results of earlier studies in adults showing an extended neural network involved in producing a power grip under visual feedback. The spatial distribution of the network was consistent over all the age groups studied. The expected activation patterns in relation to increasing target force levels were also observed in all the age groups studied. They also replicated earlier developmental findings with regard to slower reaction times and shallower slopes in the force trajectories.

These authors were able to conclude that the spatial distribution of the power grip network matures early in development. Interestingly, however, a network that is robustly deactivated when performing the power grip task in adults showed little or no deactivation in the younger groups. The younger groups also showed substantially higher amplitude task-related premovement potentials, and the ERPs in the feedback interval were also much larger in children and adolescents. These results suggest that children and adolescents show less focused neural activity in relation to performance of the visually guided power grip task. Thus, as humans mature from adolescence to young adulthood, the neural circuits involved in power grip continue to undergo changes until the neural activity is much more task specific. From the details of this study, which are not reviewed here, we know the specific brain regions showing greater activation and lesser deactivation in the younger participants, and we know the time course of increased activation in the younger participants in relation to perrformance of the task.

It seems clear that much more is known about the neural control of power grip generation over development than is known about the development of neural control of orofacial movements for speech. We have begun the stage of making detailed behavioral observations of speech motor output over many developmental periods, and imaging studies in adults are beginning to map the neural networks involved in prespeech planning and production (Bohland & Guenther, 2006). Obviously, there are technical difficulties involved in imaging during performance of a motor task that takes place in the head, which makes it more difficult to study speech movements with imaging and EEG approaches. However, signal processing methods that reduce noise and artifact contamination of data are improving. In any case, the work by Halder and colleagues may be seen as a map of what the future holds for the study of speech motor development. In addition to the nature of the experiment, which allowed detailed assessment of the motor output at the periphery along with excellent spatial and temporal resolution of the accompanying central nervous system activity, an important aspect of this study is the research team itself. Advances in understanding the neural control of speech motor development will depend upon assembling similar research teams who can make multileveled observations of both peripheral speech motor output and the neuronal activity generating that output in young children, adolescents, and adults.

NOTE

This investigation, as well as those of Riely and Smith (2003) and Smith and Zelaznik, (2004) was part of a large study of the development of speech motor control in typically developing children and young adults in which 240 participants were tested on the same protocols. This and other work from our laboratory described in this chapter was supported by grants from the NIH's National Institute on Deafness and other Communication Disorders.

REFERENCES

Ackerson, L. M. & Bruce, E. N. (1983) Bilaterally synchronized oscillations in human diaphragm and intercostal EMGs during spontaneous breathing. *Brain Research*, 271, 346–8.

Barlow, S. M. & Bradford, P. (1996) Comparison of perioral reflex modulation in the upper and lower lip. *Journal of Speech and Hearing Research*, 39, 55–75.

Barlow, S. M., Dusick, A., Finan, D. S., Coltart, S., & Biswas, A. (2001) Mechanically evoked perioral reflexes in premature and term human infants. *Brain Research*, 899, 251–4.

Barlow, S. M. & Estep, M. (2006) Central pattern generation and the motor infrastructure for suck, respiration, and speech. *Journal of Communication Disorders*, 39, 366–80.

Barlow, S. M., Finan, D. S., Bradford, P. T., & Andreatta, R. D. (1993) Transitional properties of the mechanically evoked perioral reflex from infancy through adulthood. *Brain Research*, 623, 181–8.

Barlow, S. M., Finan, D. S., & Park, S. (2004) Sensorimotor entrainment of respiratory and orofacial systems in humans. In B. Maasen, R. D. Kent, H. Peters, P. H. H. M. van Lieshout, & W. Hulstijn (eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 211–24). Oxford: Oxford Universsity Press.

Baum, S. R. & Katz, W. F. (1988) Acoustic analysis of compensatory articulation in children. *Journal of the Acoustical Society of America*, 84, 1662–8.

Beck, R. J., Andriacchi, T. P., Kuo, K. N., Fermier, R. W., & Galante, J. O. (1981) Changes in the gait patterns of growing children. *Journal of Bone and Joint Surgery*, 63, 1452–6.

Bernstein, N. A. (1967) *The Coordination and Regulation of Movements*. Oxford: Pergamon Press.

Blakemore, S. & Choudhury, S. (2006) Development of the adolescent brain: Implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry*, 47, 296–312.

Bohland, J. W. & Guenther, F. H. (2006) An fMRI investigation of syllable sequence production. *Neuroimage*, 32, 821–41.

Boliek, C. A., Hixon, T. J., Watson, P. J., & Morgan, W. J. (1996) Vocalization and breathing during the first year of life. *Journal of Voice*, 10, 1–22.

Boliek, C. A., Hixon, T. J., Watson, P. J., & Morgan, W. J. (1997) Vocalization and breathing during the second and third years of life. *Journal of Voice*, 11, 373–90.

Bosma, J. F. (1975) Anatomic and physiologic development of the speech apparatus. In D. B. Tower (ed.), *The Nervous System, Part 3: Human Communication and its Disorders* (pp. 469–81). New York: Raven Press.

Brown, R. (1973) *A First Language: The Early Stages*. London: George Allen & Unwin.

Callan, D. E., Kent, R. D., Guenther, F. H., & Vorperian, H. K. (2000) An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research*, 43, 721–36.

Chakraborty, R., Goffman, L., & Smith, A. (2008) Physiological indices of bilingualism: Oral motor coordination and speech rate in Bengali-English speakers. *Journal of Speech, Language, and Hearing Research*, 51, 321–32.

Cheng, H. Y., Murdoch, B. E., Goozee, J. V., & Scott, D. (2007) Physiological development of tongue–jaw coordination from childhood to adulthood. *Journal of Speech, Language, and Hearing Research*, 50, 352–60.

Connaghan, K. P., Moore, C. A., Christopher, A., & Higashikawa, M. (2004) Respiratory kinematics during vocalization and nonspeech respiration in children from 9 to 48 months. *Journal of Speech, Language, and Hearing Research*, 47, 70–84.

Cooker, H., Larson, C., & Luschei, E. (1980) Evidence that the human jaw stretch reflex increases the resistance of the mandible to small displacements. *Journal of Physiology*, 308, 61–78.

Crossman, E. R. & Szafran, J. (1956) Changes with age in the speech of information-intake and discrimination. *Experientia*, 4, 128–34.

Crystal, T. H. & House, A. S. (1988) A note on the variability of timing control. *Journal of Speech and Hearing Research*, 31, 497–502.

Davids, K., Newell, K., & Bennett, S. (eds.) (2006) *Movement System Variability*. Champaign, IL: Human Kinetics.

Dellow, P. G. & Lund, J. P. (1971) Evidence for central timing of rhythmical mastication. *Journal of Physiology*, 215, 1–13.

Deutsch, G. K., Dougherty, R. F., Bammer, R., Siok, W. T., Gabrieli, J. D., & Wandell, B. (2005) Children's reading performance is correlated with white matter structure measured by diffusion tensor imaging. *Cortex*, 41, 354–63.

Deutsch, K. M. & Newell, K. M. (2003) Deterministic and stochastic processes in children's isometric force variability. *Developmental Psychobiology*, 43, 335–45.

Deutsch, K. M. & Newell, K. M. (2004) Changes in the structure of children's isometric force variability with practice. *Journal of Experimental Child Psychology*, 88, 319–33.

Edelman, G. M. (1987) *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

Edelman, G. M. (1989) *The Remembered Present*. New York: Basic Books.

Edwards, J. (1992) Compensatory speech motor abilities in normal and phonetically disordered children. *Journal of Phonetics*, 20, 189–207.

Ertmer, D. J., Young, N. M., & Nathani, S. (2007) Profiles of vocal development in young cochlear implant recipients. *Journal of Speech, Language, and Hearing Research*, 50, 393–407.

Euler, C. von (1982) Some aspects of speech breathing physiology. In S. Grillner, B. Lindblom, J. Lubker, & A. Persson (eds.), *Speech Motor Control* (pp. 95–103). New York: Pergamon Press.

Farkas, L. G. (1994) *Anthropometry of the Head and Face*. New York: Raven Press.

Finan, D. S. & Barlow, S. M. (1998) Intrinsic dynamics and mechanosensory modulation of non-nutritive sucking in human infants. *Early Human Development*, 52, 181–97.

Finan, D. S. & Smith, A. (2005) Jaw stretch reflexes in children. *Experimental Brain Research*, 164, 58–66.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999) Age constraints on second-language learning. *Journal of Memory and Language*, 41, 78–104.

Forrest, K. (2002) Are oral-motor exercises useful in the treatment of phonological/articulatory disorders? *Seminars in Speech and Language*, 23, 15–25.

Gallistel, C. R. (1980) *The Organization of Action: A New Synthesis*. Hillsdale, NJ: Lawrence Erlbaum.

Gelfand, I. M., Gurfinkel, V. S., Tsetlin, M. L., & Shik, M. L. (1971) Some problems in the analysis of movements. In I. Gelfand, V. Gurfinkel, S. Fomin, & M. Tsetlin (eds.), *Models of the Structural-Functional Organization of Certain Biological Systems* (pp. 329–45). Cambridge, MA: MIT Press.

Gerard, J. M., Wilhelms-Tricarico, R., Perrier, P., & Payan, Y. (2003) A 3D dynamical biomechanical tongue model to study speech motor control. *Recent Research Developments in Biomechanics*, 1, 49–64.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., et al. (1999) Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience*, 2, 861–3.

Giedd, J. N., Snell, J. W., Lange, N., et al. (1996) Quantitative magnetic resonance imaging of human brain development: Ages 4–18. *Cerebral Cortex*, 6, 551–60.

Gisel, E. G. (1991) Effect of food texture on the development of chewing of children between six months and two years of age. *Developmental Medicine and Child Neurology*, 33, 69–79.

Goffman, L. (1999) Prosodic influences on speech production in children with specific language impairment: Kinematic, acoustic, and transcription evidence. *Journal of Speech, Language, and Hearing Research*, 42, 1499–1517.

Goffman, L., Gerken, L., & Lucchesi, J. (2007) Relations between segmental and motor variability in prosodically complex nonword sequences. *Journal of Speech, Language, and Hearing Research*, 50, 444–58.

Goffman, L., Heisler, L., & Chakraborty, R. (2006) Mapping of prosodic structure onto words and phrases in children's and adults' speech production. *Language and Cognitive Processes*, 21, 25–47.

Goffman, L. & Smith, A. (1999) Development and differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 649–60.

Goffman, L., Smith, A., Heisler, L., and Ho, M. (2008) Coarticulation as an index of speech production units in children and adults. *Journal of Speech, Language, and Hearing Research*, 51, 1424–37.

Green, J. R., Moore, C. A., Higashikawa, M., & Steeve, R. W. (2000) The physiologic development of speech motor control: Lip and jaw coordination. *Journal of Speech, Language, and Hearing Research*, 43, 239–55.

Green, J. R., Moore, C. A., & Reilly, K. J. (2002) The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*, 45, 66–79.

Green, J. R., Moore, C. A., Ruark, J. L., Rodda, P. R., Morvée, W. T., & Vanwitzenburg, M. J. (1997) Development of chewing in children from 12 to 48 months: Longitudinal study of EMG patterns. *Journal of Neurophysiology*, 77, 2704–27.

Green, J. R. & Wilson, E. M. (2006) Spontaneous facial motility in infancy: A 3D kinematic analysis. *Developmental Psychobiology*, 48, 16–28.

Grillner, S. (1982) Possible analogies in the control of innate motor acts and the production of sound in speech. In S. Grillner, B. Lindblom, J. Lubker, & A. Persson (eds.), *Speech Motor Control* (pp. 217–30). New York: Pergamon Press.

Guenther, F. H. (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102, 594–621.

Guenther, F. H. (2006) Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39, 350–65.

Halder, P., Brem, S., Bucher, K., et al. (2007) Electrophysiological and hemodynamic evidence for late maturation of hand power grip and force control under visual feedback. *Human Brain Mapping*, 28, 69–84.

Harasty, J., Double, K. L., Halliday, G. M., Krill, J. J., & McRitchie, D. A. (1997) Language-associated cortical regions are proportionally larger in the female brain. *Archives of Neurology*, 54, 171–6.

Hebb, D. O. (1949) *The Organization of Behavior*. New York: John Wiley.

Hoit, J. D., Hixon, T. J., Watson, P. J., & Morgan, W. J. (1990) Speech breathing in children and adolescents. *Journal of Speech and Hearing Research*, 33, 51–69.

Holcomb, P. J., Coffey, S. A., & Neville, H. J. (1992) Visual and auditory sentence processing: A developmental analysis using event-related brain potentials. *Developmental Neuropsychology*, 8, 203–41.

Honda, K. (1996) The organization of tongue articulation for vowels. *Journal of Phonetics*, 24, 39–52.

Honda, K., Kurita, T., & Kakita, Y. (1995) Physiology of the lips and modeling of lip gestures. *Journal of Phonetics*, 23, 243–54.

Huang, C. S., Hiraba, H., Murray, G. M., & Sessle, B. J. (1988) Organization of the primate face motor cortex as revealed by intracortical microstimulation and electrophysiological identification of afferent inputs and corticobulbar projections. *Journal of Neurophysiology*, 59, 796–818.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999) Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America*, 106, 1532–42.

Indefrey, P. & Levelt, W. J. M. (2000) The neural correlates of language production. In M. Gazzaniga (ed.), *The New Cognitive Neurosciences*, 2nd edn. (pp. 845–65). Cambridge, MA: MIT Press.

Jancke, L., Siegenthaler, T., Preis, S., & Steinmetz, H. (2007) Decreased white-matter density in a left-sided fronto-temporal network in children with developmental language disorder: Evidence for anatomical anomalies in a motor-language network. *Brain and Language*, 102, 9–8.

Kent, R. D. (1976) Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research*, 19, 421–77.

Kent, R. D. (1992) The biology of phonological development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (eds.), *Phonological Development: Models, Research, Implications* (pp. 65–90). Timonium, MD: York Press.

Kent, R. D. (2004) The uniqueness of speech among motor systems. *Clinical Linguistics and Phonetics*, 18, 495–505.

Kent, R. D. & Bauer, H. R. (1985) Vocalizations of one-year-olds. *Journal of Child Language*, 12, 491–526.

Kent, R. D., Mitchell, P. R., & Sancier, M. (1991) Evidence and role of rhythmic organization in early vocal development in human infants. In J. Faggard & P. H. Wolff (eds.), *The Development of Timing Control and Temporal Organization in Coordinated Action* (pp. 135–49). North Holland: Elsevier.

Kiliaridis, S., Karlsson, S., & Kjellberg, H. (1991) Characteristics of masticatory mandibular movements and velocity in growing individuals and young adults. *Journal of Dental Research*, 70, 1367–70.

Knudsen, E. I. (2004) Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, 16, 1412–25.

Kuehn, D. P. & Moll, K. L. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4, 303–20.

Larson, C., Finocchio, D., Smith, A., & Luschei, E. (1983) Jaw muscle afferent firing during an isotonic jaw-positioning task in the monkey. *Journal of Neurophysiology*, 50, 61–73.

Lenroot, R. K. & Giedd, J. N. (2006) Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews*, 30, 718–29.

Levelt, W. J. & Wheeldon, L. (1994) Do speakers have access to a mental syllabary? *Cognition*, 50, 239–69.

Lund, J. P. & Kolta, A. (2006) Brainstem circuits that control mastication: Do they have anything to say during speech? *Journal of Communication Disorders*, 39, 381–90.

Luschei, E. S. (1991) Development of objective standards of nonspeech oral strength and performance: An advocate's views. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (eds.), *Dysarthria and Apraxia of Speech: Perspectives on Management* (pp. 3–14). Baltimore: Paul H. Brookes.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. *Psychological Review*, 77, 182–96.

MacNeilage, P. F. (1998) The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499–511.

MacNeilage, P. F. & Davis, B. L. (1990) Acquisition of speech production: Frames then content. In M. Jeannerod (ed.), *Attention and Performance XII: Motor Representation and Control* (pp. 453–76). Hillsdale, NJ: Lawrence Erlbaum.

MacNeilage, P. F. & Davis, B. L. (2000) On the origin of internal structure of word forms. *Science*, 288, 527–31.

MacNeilage, P. F., Davis, B. L., Kinney, A., & Matyear, C. L. (2000) The motor core of speech: A comparison of serial organization patterns in infants and languages. *Child Development*, 71, 153–63.

Maner, K., Smith, A., & Grayson, L. (2000) Influences of length and syntactic complexity on speech motor performance of children and adults. *Journal of Speech, Language, and Hearing Research*, 43, 560–73.

McClean, M. D. & Clay, J. L. (1994) Evidence for suppression of lip muscle reflexes prior to speech. *Experimental Brain Research*, 97, 541–4.

Moore, C. A. (2004) Physiologic development of speech production. In B. Maasen, R. D. Kent, H. Peters, P. H. H. M. van Lieshout, and W. Hulstijn (eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 191–210). Oxford: Oxford University Press.

Moore, C. A., Caulfield, T. J., & Green, J. R. (2001) Relative kinematics of the rib cage and abdomen during speech and nonspeech behaviors of 15-month-old children. *Journal of Speech, Language, and Hearing Research*, 44, 80–94.

Moore, C. A. & Ruark, J. L. (1996) Does speech emerge from earlier appearing motor behaviors? *Journal of Speech and Hearing Research*, 39, 1034–47.

Moore, C. A., Smith, A., & Ringel, R. L. (1988) Task-specific organization of human jaw muscles. *Journal of Speech and Hearing Research*, 31, 670–80.

Nathani, S., Ertmer, D. J., & Stark, R. E. (2006) Assessing vocal development in infants and toddlers. *Clinical Linguistics and Phonetics*, 20, 351–69.

Neville, H., Mills, D., & Lawson, D. (1992) Fractionating language: Different neural subsystems with different sensitive periods. *Cerebral Cortex*, 2, 244–58.

Newell, K. M., Deutsch, K. M., Sosnoff, J. J., & Gottfried, M. (2006) Variability in motor output as noise: A default and erroneous proposition? In K. Davids, K. Newell, & S. Bennett (eds.), *Movement System Variability* (pp. 3–24). Champaign, IL: Human Kinetics.

Newell, K. M., Liu, Y-T., & Mayer-Kress, G. (2001) Time scales in learning and development. *Psychological Review*, 108, 57–82.

Nittrouer, S. (1993) The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. *Journal of Speech and Hearing Research*, 36, 959–72.

Oller, D. (1980) The emergence of speech sounds in infancy. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (eds.), *Child Phonology, vol. I: Production* (pp. 93–112). New York: Academic Press.

Ostry, D. J., Vatikiotis-Bateson, E., & Gribble, P. L. (1997) An examination of the degrees of freedom of human jaw motion in speech and mastication. *Journal of Speech, Language, and Hearing Research*, 40, 1341–51.

Paus, T., Collins, D. L., Evans, A. C., Leonard, G., Pike, B., & Zijdenbos, A. (2001) Maturation of white matter in the human brain: A review of magnetic resonance studies. *Brain Research Bulletin*, 54, 255–66.

Peng, S., Spencer, L. J., & Tomblin, J. B. (2004) Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience. *Journal of Speech, Language, and Hearing Research*, 47, 1227–36.

Perkell, J. S., Guenther, F. H., Lane, H., et al. (2000) A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, 28, 233–72.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1995) Goal-based speech motor control: A theoretical framework and some preliminary data. *Journal of Phonetics*, 23, 23–5.

Plante, E., Schmithorst, V. J., Holland, S. K., & Byars, A. W. (2006) Sex differences in the activation of language cortex during childhood. *Neuropsychologia*, 44, 1210–21.

Riely, R. & Smith, A. (2003) Speech movements do not scale by orofacial structure size. *Journal of Applied Physiology*, 94, 2119–26.

Ruark, J. L. & Moore, C. A. (1997) Coordination of lip muscle activity by 2-year-old children during speech and nonspeech tasks. *Journal of Speech and Hearing Research*, 40, 1373–85.

Sasisekaran, J., Smith, A., Sadagopan, N., & Weber-Fox, C. (In press) Nonword repetition in in children and adults: Effects on movement coordination. Developmental Science.

Schmidt, R. A. (1988) *Motor Control and Learning: A Behavioral Emphasis*, 2nd edn. Champaign, IL: Human Kinetics.

Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn, J. T. (1979) Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review*, 47, 415–51.

Sciote, J. J., Horton, M. J., Rowlerson, A. M., & Link, J. (2003) Specialized cranial muscles: How different are they from limb and abdominal muscles? *Cells Tissues Organs*, 174, 73–86.

Sharkey, S. G. & Folkins, J. W. (1985) Variability of lip and jaw movements in children and adults: Implications for the development of speech motor control. *Journal of Speech and Hearing Research*, 28, 8–15.

Sheppard, J. J. & Mysak, E. D. (1984) Ontogeny of infantile oral reflexes and emerging chewing. *Child Development*, 55, 831–43.

Sherrington, C. S. (1906) *The Integrative Action of the Nervous System*. New Haven: Yale University Press.

Smith, A. (1992) The control of orofacial movements in speech. *Critical Reviews in Oral Biology and Medicine*, 3, 233–67.

Smith, A. (2006) Speech motor development: Integrating muscles, movements, and linguistic units. *Journal of Communication Disorders*, 39, 331–49.

Smith, A. & Denny, M. (1990) High frequency oscillations as indicators of neural control mechanisms in human respiration, mastication, and speech. *Journal of Neurophysiology*, 63, 745–58.

Smith, A. & Goffman, L. (1998) Stability and patterning of speech movement sequences in children and adults. *Journal of Speech, Language, and Hearing Research*, 41, 18–30.

Smith, A. & Goffman, L. (2004) Interaction of language and motor factors in speech production. In B. Maasen, R. D. Kent, H. Peters, P. H. H. M. van Lieshout, & W. Hulstijn (eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 225–52). Oxford: Oxford University Press.

Smith, A., Goffman, L., Zelaznik. H., Ying, G., & McGillem, C. (1995) Spatiotemporal stability and patterning of speech movement sequences. *Experimental Brain Research*, 104, 493–501.

Smith, A., Johnson, J., McGillem, C., & Goffman, L. (2000) On the assessment of stability and patterning of speech movement sequences. *Journal of Speech, Language, and Hearing Research*, 43, 277–86.

Smith, A., McFarland, D., Weber, C., & Moore, C. (1987) Spatial organization of perioral reflexes. *Experimental Neurology*, 98, 233–48.

Smith, A., Moore, C. A., McFarland, D. J., & Weber, C. M. (1985) Reflex responses of human lip muscles to mechanical stimulation during speech. *Journal of Motor Behavior*, 17, 148–67.

Smith, A., Moore, C. A., & Pratt, C. A. (1985) Distribution of the human jaw stretch reflex response elicited by percutaneous, localized stretch of jaw-closing muscles. *Experimental Neurology*, 88, 544–61.

Smith, A., Moore, C., Weber, C., McFarland, D., & Moon, J. (1985) Reflex responses of the human jaw-closing system depend on the locus of intraoral mechanical stimulation. *Experimental Neurology*, 90, 489–509.

Smith, A., Weber, C. M., Newton, J., & Denny, M. (1991) Developmental and age-related changes in reflexes of the human oral motor system. *Electroencephalography and Clinical Neurophysiology*, 81, 118–28.

Smith, A. & Zelaznik, H. N. (2004) Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, 45, 22–33.

Smith, B. L. (1992) Relationships between duration and temporal variability in children's speech. *Journal of the Acoustical Society of America*, 91, 2165–74.

Smith, B. L. (1995) Variability of lip and jaw movements in the speech of children and adults. *Phonetica*, 52, 307–16.

Smith, B. L. & Gartenburg, T. E. (1984) Initial observations concerning developmental characteristics of labiomandibular kinematics. *Journal of the Acoustical Society of America*, 75, 1599–1605.

Smith, B. L. & McLean-Muse, A. (1986) Articulatory movement characteristics of labial consonant production by children. *Journal of the Acoustical Society of America*, 80, 1321–7.

Smith, B. L., Sugarman, M. D., & Long, S. H. (1983) Experimental manipulation of speaking rate for studying temporal variability. *Journal of the Acoustical Society of America*, 74, 744–9.

Smith, K. K. & Kier, W. M. (1989) Trunks, tongue, and tentacles: Moving with skeletons of muscle. *American Scientist*, 77, 29–35.

Smits-Engelsman, B. C. & Van Galen, G. P. (1997) Dysgraphia in children: Lasting psychomotor deficiency or transient developmental delay? *Journal of Experimental Child Psychology*, 67, 164–84.

Sommer, M., Koch, M. A., Paulus, W., Weiller, C., & Büchel, C. (2002) A disconnection of speech-relevant brain areas in developmental stuttering. *Lancet*, 360, 380–3.

Stal, P., Eriksson, P. O., Eriksson, A., & Thornell, L. E. (1990) Enzyme-histochemical and morphological characteristics of muscle fibre types in the human buccinator and orbicularis oris. *Archives of Oral Biology*, 35, 449–58.

Stal, P., Marklund, S., Thornell, L. E., De Paul, R., & Eriksson, P. O. (2003) Fibre composition of human intrinsic tongue muscles. *Cells Tissues Organs*, 173, 147–61.

Stark, R. E. (1980) Stages of speech development in the first year of life. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (eds.), *Child Phonology*, vol. 1, (pp. 73–90). New York: Academic Press.

Stathopoulos, E. T. (1995) Variability revisited: An acoustic, aerodynamic, and respiratory kinematic comparison of children and adults during speech. *Journal of Phonetics*, 23, 67–80.

Stathopoulos, E. T. & Sapienza, C. M. (1993) Respiratory and laryngeal measures of children during vocal intensity variation. *Journal of the Acoustical Society of America*, 94, 2531–43.

Stathopoulos, E. T. & Sapienza, C. M. (1997) Developmental changes in laryngeal and respiratory function with variations in sound pressure level. *Journal of Speech, Language, and Hearing Research*, 40, 595–614.

Thelen, E. (1979) Rhythmical stereotypies in normal human infants. *Animal Behavior*, 27, 699–715.

Thelen, E. & Smith, L. B., (1994) *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge, MA: MIT Press.

Tingley, B. M. & Allen, G. D. (1975) Development of speech timing control in children. *Child Development*, 46, 186–94.

Trulsson, M. & Johansson, R. S. (2002) Orofacial mechanoreceptors in humans: Encoding characteristics and responses during natural orofacial behaviors. *Behavioural Brain Research*, 135, 27–33.

Van Galen, G. P., Portier, S. J., Smits-Engelsman, B. C. M., & Schomaker, L. R. B. (1993) Neuromotor noise and deviant movement strategies as an explanatory ground for poor handwriting in children. *Acta Psychologica*, 82, 161–78.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005) Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, 117, 338–50.

Walsh, B. & Smith, A. (2002) Articulatory movements in adolescents: Evidence for protracted development of speech motor control processes. *Journal of Speech, Language, and Hearing Research*, 45, 1119–33.

Walsh, B., Smith, A., & Weber-Fox, C. (2006) Short-term plasticity in children's speech motor systems. *Developmental Psychobiology*, 48, 660–74.

Weber-Fox, C. & Neville, H. J. (1996) Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience*, 8, 231–56.

Weismer, G. (2006) Philosophy of research in motor speech disorders. *Clinical Linguistics and Phonetics*, 20, 315–49.

Weismer, G. & Liss, J. M. (1991) Reductionism is a dead-end in speech research: Perspectives on a new direction. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (eds.), *Dysarthria and Apraxia of Speech: Perspectives on Management* (pp. 15–28). Baltimore: Paul H. Brookes.

Wilke, M., Krägeloh-Mann, I., & Holland, S. K. (2007) Global and local development of gray and white matter volume in normal children and adolescents. *Experimental Brain Research*, 178, 296–307.

Wood, J. & Smith, A. (1992) Cutaneous oral motor reflexes of children with normal and disordered speech. *Developmental Medicine and Child Neurology*, 34, 797–812.

Yan, J. H., Thomas, J. R., Stelmach, G. E., & Thomas, K. T. (2000) Developmental features of rapid aiming arm movements across the lifespan. *Journal of Motor Behavior*, 32, 121–40.

Part III Modeling Speech Production and Perception

8 Speech Acquisition

BARBARA L. DAVIS

1 The Problem

The young child's ability to integrate physiology and cognition to achieve phonological competence for linguistic communication is a unique human achievement. Controversy continues regarding the nature of structural categories underlying mature language competence. Theories suggesting graded perception based on frequency and context (Bybee, 2001; Pierrehumbert, 2003) seek to challenge Universal Grammar (UG) based notions of universal underlying representations (e.g., Prince & Smolensky, 1993, 2004). Whatever the nature of the adult phonology, the child's acquisition of an ambient language phonological system implies two distinct necessities. The young child must marshal maturing production, perception, neural, and cognitive capacities to perceive and produce the sounds, sequences, and prosodic regularities of the ambient language. Co-temporally, a stable knowledge base for connecting speech forms with meaning in communicative exchanges must be mastered. The link of developing peripheral capacities with growth in ambient language knowledge, memory, and retrieval capacities enables children to master the necessity of conveying an ever broadening set of ideas to an increasingly diverse set of listeners. The central problem of modeling speech acquisition is focused on validly conceptualizing the nature of this process.

2 The Broader Context

The "nature"–"nurture" debate is a subject of a continuing controversy related to the origins of complex capacities in young humans; simply put, where does knowledge come from and how does it grow in childhood? "Nature" refers to an understanding of human infants as endowed at birth with categories of knowledge to assist them in developing understanding of their world. This perspective is in the nativist philosophical tradition of Descartes (1824) and Kant (1924). In the nativist view, adults and infants share the same capacities and view the world in largely the same way, although certain abilities clearly take time to mature. In contrast, the "nurture" philosophical tradition emerges from work by empiricist philosophers, including John Locke in the seventeenth century (Locke & St John, 1854) and William James in the nineteenth century (1890), who both proposed that development comes from the senses and reflects learning. Adults are mature and infants are considered naïve, so that all of the knowledge of the world must be gained through learning by the growing infant from the environment.

Debate on this fundamental topic of scientific inquiry forms a basic philosophical difference in epistemological study of the origins of knowledge in acquisition of complex systems. It has, as well, been consistently present in theoretical proposals on speech acquisition. These two quite diverse interpretations have been pursued in separate research programs considering acquisition of the complex system embodied in development of speech perception and production capacities to support phonological knowledge. There has been little philosophical overlap in underlying conceptualization of the process of acquisition between these two perspectives.

Linguistic study of the acquisition phase of synchronic sound patterns in languages has been related to the general issue of "learnability" (e.g., Pesetsky, 1999). Traditional linguistic inquiry adheres strongly to a nativist perspective on the origins of phonological knowledge. Here there is relatively more emphasis on acquisition as emblematic of the child's abstractly represented knowledge of phonology rather than consideration of peripheral production or perceptual structures and their function. Recent treatments considering the potential role of input and function (Newmeyer, 1998) depart from earlier, more strictly adhered to philosophical emphases on modularity (Chomsky & Halle, 1968; Fodor, 1983).

From a phonetic, or more empiricist perspective, Lindblom (2004) has noted the complex relationships between the child's physical capacities and the linguistic description of an utterance. He has emphasized the need to understand the unit of speech production in circumstances where speech is a dynamic event (Kohler, 2000) necessarily characterized by coarticulation (see Hardcastle & Hewlett, 1999, for a review). Lindblom (1992) voices the general tenor of phonetic approaches in suggesting that understanding of phonological acquisition should derive "from starting points that are motivated by knowledge independent of the facts to be explained" (Lindblom, 1992, p. 135). In contrast to phonological interpretations centered on abstract knowledge, phonetic perspectives emphasize early bio-behavioral capacities of the production, perception, and neural subsystems. Less emphasis is placed on connections between mature language forms and the process of acquisition of those forms. We will consider the implications of these diverse conceptualizations of speech acquisition in this chapter.

Study of speech acquisition as it is embodied in speech production skill marshaled for linguistic communication has applications across a variety of scientific disciplines. In recent treatments of the acquisition of speech production skill, connections have been made to questions about the evolution of speech capacities from both linguistic (e.g., Newmeyer, 1998; Blevins, 2004) and phonetic perspectives (e.g., Studdert-Kennedy, 1998; MacNeilage & Davis, 2000). The question of evolution of speech production capacity in early hominids leading to generative linguistic communication capacities in modern speakers is related to deeper levels of explanation for patterns found in modern languages. The acquisition phase of these capacities in young children enables a look at this system when it is in a simple form providing a site for consideration of origins. The acquisition paradigm has been taken into account within contemporary phonological, phonetic, syntactic, and general cognitive science perspectives on language origins.

Speech acquisition in human infants also provides a site for considerations of boundary values for human language capacities relative to nonhumans (Hauser, 1996; Hauser et al., 2002). Hauser (1996) asserts that ultimate understanding of why human communication systems exhibit unique design features differing from other animals necessitates that "language" be defined as a communication tool rather than as a formal symbolic structure organized around syntax. Hauser et al. (2002) have proposed that recursiveness forms the boundary that separates human from nonhuman phonological capacities. They propose an abstract linguistic computational system, and a broader faculty of language including both the abstract computational system and sensory-motor (i.e., phonetic) and conceptual (i.e., semantic and pragmatic) systems with which that system interacts. Their proposal illuminates the heart of the dialogue on what constitutes the boundary value for human language and illustrates the power of acquisition study as a tool in this area of scientific inquiry.

Applied concerns regarding underlying bases of developmental speech disorders also reflect the conceptual dichotomy manifest in understanding the typical course of speech acquisition (see Baker, 2006, for a review). Simply put, what is the appropriate focus of clinical remediation in conditions where children do not develop age-appropriate intelligibility for spoken language? A knowledge system; or a set of behavioral capacities; or both? Conceptualization of the typical course of speech acquisition provides a necessary backdrop for clinical intervention in providing a picture of behaviors appropriate for a child's chronological age.

3 Contemporary Theoretical Perspectives on Speech Acquisition

Consideration of the underlying nature of the speech acquisition process falls within the scope and interests of a number of distinct scientific disciplines. Exploring the claims and paradigms of available theoretical perspectives is basic to characterizing the contemporary scholarly landscape. In addition, this type of overview points to new directions for generating an integrative view of how young humans acquire the bio-behavioral action and the perception and neural-cognitive knowledge capacities to support intelligible speech production for linguistically based social communication functions.

Contemporary acquisition theories are predominantly found in phonology (e.g., Prince & Smolensky, 1993; Archangeli & Pulleyblank, 1994) and phonetics (e.g., Lindblom, 1992; MacNeilage & Davis, 1993; Vihman, 2000). An additional dimension of phonetic study strongly emphasizes the essential need to incorporate functional social pressures in conceptualizing the driving forces underlying speech acquisition (e.g., Locke, 1993; Oller & Greibel, 2008).

Approaches to the acquisition of knowledge available from cognitive science have also been applied to considering emergence of various complex knowledge systems in humans. Computer learning algorithms (e.g., de Boer, 2001, 2005) and neural net models (e.g., Rumelhart & McClelland, 1986) have fuelled the development of new ways of considering mechanisms underlying the process of acquisition. The burgeoning area of modeling studies centered in artificial intelligence (Kirby, 1999), and robotics (Steels, 2006) contributes additional new methodologies for understanding the course of change in a complex system as it is reflected in the human speech acquisition process. In these modeling paradigms, the "pressures" on an emerging complex system can be observed over countless "generations" of change. These paradigms have facilitated broadening the scope of inquiry to consideration of mechanisms underlying the process of speech acquisition. Self-organization (Kauffman, 1995), learning (Guenther, 1995; Guenther & Perkell, 2004), and linguistic processing variables including memory (Gathercole & Baddeley, 1993) and lexical retrieval (Edwards, Beckman, & Munson, 2004) have been considered. These diverse perspectives within cognitive science seek understanding of phonological acquisition via domain-general cognitive mechanisms available for supporting the emergence of complexity. Almost no notice is given to the production system. Perception mechanisms are largely equated with frequency of input as the major factor underlying learning of environmental regularities from the ambient phonology.

Paradigms employed for evaluation of historical and contemporary theoretical conceptualizations of acquisition illustrate the disparate and often quite philosophically dissimilar standards of evidence used to test hypotheses across intellectual disciplines. Both the quantity and quality of information considered as providing support within diverse theoretical perspectives differ. Corpora range from small sets of example data illustrating underlying rules (e.g., Goad & Rose, 2003) to extremely large bodies of data used to evaluate system-wide patterns and how they emerge over time (e.g., Davis & MacNeilage, 1990; de Boer, 2001).

3.1 Formalist phonological perspectives

Phonological theories represent a consistent strand of proposals on acquisition of sound patterns in languages. The conceptual basis for phonological theories rests on the Platonic premise of innate knowledge. Plato was the first to propose innateness as a solution to the problem of describing the origins of knowledge. Platonic philosophy counted "essences" as foundational building blocks necessary to construction of knowledge of the world. This approach is characterized in contemporary terms as "Essentialism" (e.g., Gelman, 2003). In this view, categories can be considered to be natural kinds by their "essence." These natural kinds are based in nature. They capture many of the regularities of their component elements. In the context of speech acquisition, nasal consonants /m/, /n/, and

/ng/ are considered to be a natural class or kind based on the congruence of their nasal production characteristics. Natural kinds are said to be discovered by the child. From the essentialist perspective, such categories are characterized as being highly stable in the environment. In the domain of language, natural kinds are said to enable children to more readily notice some types of environmental categories, such as living things.

Although quite diverse in a wide variety of aspects, phonological perspectives on acquisition share the view that a competence based system driven by a priori form underlies a child's expression of phonological knowledge. Abstract representations of language regularities are found in the child's "underlying representation" and are "parameterized" by experience (Stemberger & Bernhardt, 1999). Blevins (2004) suggests a set of "phonological primitives" in UG that include distinctive features, segments, length, and prosodic categories (mora, syllable, foot, etc.). However, the membership and nature of the set of these phonological primitives differs across available theories. These stored mental representations, whatever their nature, are said to have "psychological reality," as they are not precisely specified relative to neural instantiation.

Contemporary phonological theories have diversified from those of early perceptually based theorists (Jakobson, 1968) and the classic linear generative phonology proposed by Chomsky and Halle (1968). Current constructions of phonological theory center largely on approaches such as optimality theory (OT) (e.g., Prince & Smolensky, 1993; Bernhardt & Stemberger, 1998), grounded phonology (Archangeli & Pulleyblank, 1994), and prosodic phonology (McCarthy & Prince, 1993). Compared with the linear feature strings of classic linear generative phonology (Chomsky & Halle, 1968), these newer conceptualizations include syllable-level prosodic effects and incorporate units of of sizes different from the phoneme in underlying representations. Within OT (Prince & Smolensky, 1993), for example, a series of mentally coded "constraints" on phonology and morphosyntax are proposed as interacting to guide production of possible well-formed output. Phonological constraints such as "*complex onset" and "no coda" lead speakers to avoid complexity. These constraints are generally hypothesized to be innate and universal (e.g., Dekkers, van der Leeuw, & van de Weijer, 2000). However, some OT proposals suggest that constraints are phonetically grounded (e.g., Bernhardt & Stemberger, 1998). Constraints can be violated and rankings can vary across languages and over time within speakers leading to the variation observed across languages and in acquisition (Bernhardt & Stemberger, 1998).

Markedness forms a central construct of phonological theory from early proposals (Trubetzkoy, 1929; Greenberg, 1966) through contemporary theory (Prince & Smolensky, 1993). In this conceptualization, frequency of occurrence is the main underlying metric for designating a sound as marked. Unmarked members of a phonological system are more basic. They appear earlier in acquisition and are more frequent in language inventories. Presence of more marked phonemes hierarchically implies presence of less marked phonemes. Frequency is accorded an explanatory status, implying a circularity whereby frequency descriptions of posited underlying forms are said to provide explanation. Other recent treatments of frequency have explored perceptual processing of language regularities via rapid learning mechanisms (Rose, 2009).

Research paradigms evaluating phonological perspectives on acquisition have generally centered on relatively small corpora to illustrate either underlying representations or rules that mediate between underlying representations and observable output (e.g., Stemberger & Bernhart, 1999; Goad & Rose, 2003; Pater, Stager, & Werker, 2004). Generally, the size of the corpora is not as important as the ability to generate parsimonious rules for observable linguistic structures in the child's output. Prelinguistic behaviors are of far less interest within phonological theory, with the consequence that the starting point for phonological considerations of acquisition begins when children produce identifiable word-based forms.

3.2 Functionalist phonetic perspectives

Phonetic approaches have focused on biological characteristics of the developing child and the ways in which these capacities contribute to emergence of complex speech output patterns. These approaches have been extremely diverse in both comprehensiveness and in paradigms employed for evaluation of hypotheses. They are broadly contrasted with phonological approaches in looking for biological explanation for acquisition patterns within peripheral anatomy and physiology embedded in the social function of speech forms for the young child. Movements of the organism in time and space are proposed as an important potential source of input for building *eventual* mental knowledge categories to support linguistic communication rather than a *starting point* characterized by presence of underlying representations as in phonological approaches.

The status of acquisition research from functional phonetic perspectives could presently be characterized as a mosaic of data and information about peripheral subsystem capacities during the process of speech acquisition. Diverse perspectives are employed to consider the speech acquisition process. No single approach or paradigm represents phonetic approaches to emergence of phonology in the same way that the original linear and newer OT and prosodic approaches represent formalism.

Classic phonetically oriented transcription and acoustic studies of the first year of life (e.g., Oller, 1980; Stark, 1980; Holmgren et al., 1986; Koopmans-van Beinum & Van der Steldt, 1986; Kent & Bauer, 1985; Stoel-Gammon, 1985) and of phonetic output patterns in the early word period (Bickley et al., 1986; Vihman et al., 1986; Roug et al., 1989; Davis et al., 2002) have detailed the course of vocal development in the prelinguistic and early word periods. Studies of early vocal output have shown a remarkable continuity within as well as similarity across infants. In some conceptualizations, these behavioral patterns have been termed a "motor core," based on the links between production system characteristics and vocal patterns (Locke, 1983; MacNeilage et al., 2000) as well as in their relationships to patterns in languages (Maddieson, 1986; MacNeilage et al., 1999). Recent cross-language studies have confirmed the generality of these early vocal patterns in studies of infants in a variety of language environments (Teixeira & Davis, 2002; Lee, 2003; Kern & Davis, 2009). In contrast, other perspectives on this period have emphasized individual variation in acquisition profiles (Vihman & Velleman, 2000). Studies of later periods are far less well represented in phonetically oriented research from a production system perspective (although see Smith et al., 2000).

The diverse paradigms employed to evaluate peripheral capacities present a mosaic of findings. Motor speech-oriented kinematics (Smith & Goffman, 1998), EMG (Green et al., 1997) and MRI (e.g., Fitch & Giedd, 1999; Vorperian et al., 2005) have been employed to detail emergence of speech-related movements and infer emerging speech motor control processes. Transcriptional or acoustic analyses are very commonly employed to describe behavioral patterns, with the goal of making inferences from perceptually apparent speech forms about potential physiological and or cognitive explanations for the child's output repertoire (e.g., Davis et al., 2002).

Areas as diverse as prelinguistic oral motor development (e.g., Green et al., 2002; Moore et al., 2001), respiratory capacity (Boliek et al., 1996, 1997; Moore et al., 2001), and articulatory system structure and function (Kent & Vorperian, 1995) have been explored. Vocal tract models have also been employed to simulate the ways in which articulator control is properly characterized relative to the size and shape of the developing vocal tract across development (e.g., Ménard et al., 2004; Boë & Maeda, 1997; Callan et al., 2000). In somewhat older children, kinematic paradigms have been employed to link peripheral movement measurements with establishment of linguistic categories (e.g., Smith et al., 2000).

Social-functional components have also been integrated into phonetic conceptualizations relative to consideration of functional pressures driving speech acquisition (Locke, 1993; Oller, 2000). "Functional" refers broadly to the necessity for the child to use communication tools (here vocal forms recognizable to members of the speech community) to achieve needs in the environment. Both Locke and Oller emphasize an integrative approach including bio-behavioral capacities and the social function of emerging vocal forms for the young child. This emphasis is congruent with functionally oriented cognitive approaches focused on syntactic acquisition (e.g., Tomasello, 1998), emphasizing the importance of the communicative function of language in emergence of complexity in young children. In this social-cognitive view on the driving forces for acquisition, language is seen as founded within social communication. As applied to phonetic aspects of speech acquisition, the conceptual claim is that bio-behavioral capacities are necessary, but not sufficient to model the speech acquisition process in young children; social pressures are also a necessary component of the process.

Phonetic paradigms focusing on behavioral output across acquisition emphasize the earliest periods of vocal development relative to phonological theories which typically begin by consideration of language-based patterns. This relative emphasis on early stages of acquisition implies a need for consideration of differences between *precursor* behaviors and *pre-requisite* behaviors when evaluating early vocal forms proposed as leading toward ambient language complexity. *Precursors*, in this sense, are those behaviors proposed as occurring earlier in the process of acquisition, but for which there is no clear way of testing whether they may form necessary prerequisites for more complex language-based forms. For example, early vocalizations in the first months of life (e.g., prolonged vowel-like sounds) are certainly precursor behaviors to complex behavioral forms occurring later in development. However, there are no paradigms to falsify the hypothesis that they are necessary *pre-requisites* for later phonological development. At present, phonetic perspectives do not clearly address this important conceptual requirement for producing falsifiable hypotheses. One major challenge for phonetic science is to consider this issue as crucial for building valid and comprehensive hypotheses for acquisition of mature speech production capacities. Emphasis on initial phases of the process of acquisition may be a necessary step in the development of fruitful comprehensive hypotheses. A goal of phonetic science in considering acquisition as an emergent process must be to move toward a coherent conceptualization synthesizing across varied areas of development as well as across diverse paradigms. Several contemporary approaches can be described to illustrate ways that contemporary phonetically oriented theories address production, perception, and cognitive processes supporting speech acquisition.

One production-oriented phonetic perspective, the frame/content (F/C) theory (MacNeilage & Davis, 1993; Davis & MacNeilage, 1995; Davis et al., 2002), has attempted to generate a more comprehensive set of theoretical predictions about serial tendencies in speech acquisition. In this perspective, earliest vocal sequences are enabled by rhythmic jaw oscillations without independent movements of tongue or other active articulators from the onset of babbling (e.g., "bababa"). Within syllables, open and close aspects of the jaw produce consonant and vowel percepts without place or front–back change for consonants or vowels (e.g., /ba/, dae/, or /ku/). Across syllables, manner and height changes are predicted to predominate over place and front–back changes for consonants and vowels (e.g., /daedi/, not /daedu/, and /bawa/ over /baku/), based on rhythmic jaw cycles and little independent movement of other articulators. These patterns have been tested in English (Davis & MacNeilage, 1995) and a variety of languages (Kern & Davis, 2009) and have proved strongly, although not universally, characteristic of output patterns in babbling and early words. These jaw-cycle dependent patterns are predicted to differentiate into eventual "content" or segmental elements (e.g., "b" or "d") as the child's production mechanism matures, enabling individual movement patterns for phonemes in sequences consistent with increase in pressure to produce more complex language forms for intelligibility. The unit underlying output is considered to be a holistic syllable-like unit based on jaw oscillation that progressively differentiates (Fentress, 1984) into flexible capacities for programming individual segments. The F/C theory does not make explicit predictions related to perceptual or cognitive capacities and has been tested to date only in babbling and early words.

Lindblom (1992) has proposed a "re-use" hypothesis to explain how a phonological system may emerge resulting from early development of the lexicon. In Lindblom's view, when the child consistently uses a small number of spoken forms, holistic patterns are interpreted as *gestalt motor scores*. The segmentation of these holistic patterns into smaller units defines *articulatory scores*, which are said to be composed of anatomically distinct components. Each articulatory pattern is stored in a distinct neuronal space. Once stored, such patterns are not stored again, but marked for appropriate lexical access. New segments emerge as a result of pressure toward economy of the memory storage system. For example, suppose that a child produces speech forms like [didi], [meme], and [baba]. In articulatory space, these forms are represented by both levels of jaw opening and levels of tongue position. Each specification is linked to its own type of closure movement: d_d_, m_m_, and b_b_. Once these articulator spaces are saved, because of the memory constraint, additional potential re-use of patterns of jaw-tongue movement would result, making a number of forms accessible for production in combination with the available vowel types. Lindblom (1992 argues that these forms are derived not because there is central organization, but because they emerge in a self-organized way from the interactions of perception and mental storage capacity.

Vihman's "articulatory filter" model (1993) posits that babbling and early word patterns are intrinsically related to the infant's use of production capacities, evidenced in output forms termed "vocal motor schemes" (VMS), to select salient lexical output. Vihman and colleagues' experimental perception studies (dePaolis, 2006) have tested the hypothesis that when one or more VMS are established in infant output, it is possible to measure differential attentional responses to series of short sentences featuring nonwords which include an infant's own unique VMS. Early word output forms are based on the infant's continuing experience of the general match between VMS frequently produced in babbling and salient words available from input (Vihman & Kunnari, 2006). Selection of salient words to attempt based on available VMS resonates with the biomechanical basis of babbling postulated by Davis and MacNeilage (Davis et al., 2002) and adds a cognitive perceptual component to early word selections.

"Gestural" phonology (Browman & Goldstein, 1992) presents a different approach seeking to neutralize the competence–performance dichotomy of phonological theory. It is based on action theory (e.g., Kelso et al., 1986; Saltzman & Munhall, 1989), where speech articulators are viewed as co-coordinative structures operating within an interactive system to produce goal-directed action, each constrained by membership in the structure. Phoneme "targets" take the form of locations and degree of constriction of targets in the vocal tract. "Speech and phonology are low and high dimensional descriptions of a *single* (italics ours) complex system" (Browman & Goldstein, 1992 p. 180). No translation is necessary between aspects of the system, as cognitive and motor aspects are integral to one another. The well-known "gestural scores" represent gestures (e.g., velum open/closed, lips closed/open). Notes and phase relationships between gestures are representative of the time domain. Acquisition is characterized by undifferentiated syllables, followed by differentiation into individual articulatory gestures (the "particulate principle," Studdert-Kennedy, 1998; Abler, 1989). Via self-organization, the child's behavior is proposed as converging on oral, velic, and laryngeal constrictions shared by both child and communication partner as human speakers. Imitative visual/auditory "attunement" is seen as driving this process where children recover their communication partner's degree of articulator constriction from acoustic and visual input within communicative interactions. As in phonological approaches, gestural research has largely involved transcriptional or acoustic analysis (e.g., Nittrouer et al., 1989).

3.3 Auditory input perspectives

Auditory perceptual studies have a rich and varied history in establishing characteristics of the infant auditory system at birth as well as refinements toward the ambient environment by the second half of the first year. Categorical perception demonstrated by Eimas et al. (1971) as well as a large body of subsequent research demonstrates that languages differ in ways that infants can detect at birth. By the second half of the first year, infants demonstrate sophisticated learning abilities for ambient language sequences (e.g., Saffron et al., 1996) as well as reduction in abilities to process nonnative contrasts (Werker & Tees, 1984), indicating at least a phonetic level of readiness to process ambient language patterns. Stager and Werker (1997) showed a decrease in perceptual focus in the early word period relative to babbling, suggesting a general link between perception and production as the child begins to associate language forms with meanings. However, the link of this robust body of early perceptual results, demonstrating a rich complex of supportive capacities, to what infants *do* in terms of speech-related actions has rarely been considered.

Guenther's "DIVA" model (1994; Guenther & Perkell, 2004) represents a contemporary attempt to achieve comprehensive synthesis of auditory perceptual, production system, and neural correlates underlying earliest phases of speech acquisition. The primary focus of Guenther's model is in auditory planning of speech movements; goals of speech *movements* are founded in auditory temporal space according to his conceptualization. Movement planning represents a mapping between articulation movements and their neurally coded auditory consequences. Guenther employs computational modeling, with a set of neurons illustrating putative neural representations as model parameters. Parameters are said to be tuned during a "babbling" phase when random articulator movements provide auditory feedback used to train neural mappings. These neural mappings emerge over time as phoneme representations. One important aspect of Guenther's model is his conceptualization of convex region targets (adapted from Keating, 1990), where phoneme targets are seen as occurring in multidimensional regions rather than as points. He proposes that consideration of phoneme targets as regions rather than discrete points enables accommodation of important speech phenomena including coarticulation, and contextual variability characteristic of infant vocalizations as well as adult connected speech patterns. Guenther's model has been tested on babbling and produces phoneme-like regions. He does not address later speech acquisition stages. His perspective suggests that the locus of speech acquisition lies in neurally coded auditory representations and de-emphasizes the role of the infant production system in determining early vocal forms as is suggested in perspectives such as the F/C theory (MacNeilage & Davis, 1993).

3.4 Cognitive science perspectives

Contemporary theoretical perspectives from cognitive science have typically addressed mechanisms underlying acquisition of knowledge structures and centered on other aspects of language than speech or phonology. In contrast with phonological and phonetic perspectives, the emphasis is strongly placed on "how" acquisition occurs rather that what patterns of output are apparent. Connectionist modeling (e.g., Elman et al., 1996) represents the best-studied proposal within this genre relative to speech acquisition. The emphasis in neural network connectionist perspectives is on modeling of neural systems, proposed as being the seat of mental activities underlying instantiation of complex knowledge structures. Like Guenther's DIVA model, heavy conceptual prominence in understanding knowledge acquisition is placed on perceptually driven learning. There is a lack of emphasis both conceptually and in modeling structures on motor and social influences on the developmental process. All neural networks include units (defined variously) activated by input of varied kinds. Learning via activation of component units is the critical mechanism by which these networks are modified. Activation results in changing weights. Learners are assumed to weigh specific differing inputs, such as phonemes or words, according to their frequency of occurrence and to generate the neural underpinnings of a phonological system based on a critical mass of information. While neural net models do not produce a comprehensive picture of the course of acquisition, a strength of such paradigms is their potential power for prediction of learning mechanisms in acquisition of complex knowledge structures.

Stemberger (Stemberger & Bernhardt, 1999) has proposed a connectionist model of phonological acquisition in which segments are taken to be the basic unit. Children omit segments in words and clusters, or most often substitute segments as they engage in phonological learning. Superpositional memory is a key characteristic of Stemburger's model. Partially overlapping phonological and semantic units are activated in the representation of words. Superpositional memory provokes interactions between the representations of words and promotes a regular system. Representations of groupings for similar units are referred to as a "gang" and can cause other groupings of units to access the same units, even if they are not present in the input. This process is a "gang effect." Superpositional memory and gang effects influence the accessibility of units. Units will become highly accessible because there are overlapping representations (due to superpositional memory) and gang effects favoring retrieval of specific target representations. Learning (conceptualized as a change in the level of accessibility), occurs based on differences in difficulty of sound production and shifts in the type and token frequency of a unit. Systematization is observed in many of the common production patterns observed in child speech; children's *cup*, which requires three slots, may lose one, resulting in /key/. To correct such an error, an increase in the level of activation is needed. Once the activation level is increased, the new three-slot model will serve as the attractor state. Stemberger's analysis is of interest as an attempt to consider actual child data in considering predictions of neural net models for speech acquisition. More commonly, modeling paradigms center on simulations to demonstrate system learning and do not consider actual child data relative to how behavioral correlates of systems unfold.

Cognitively oriented proposals centered on later developmental periods for speech accuracy include "neighborhood density" approaches (Edwards et al., 2004). This work is based on Luce and colleagues' "Neighborhood Activation" model of spoken word recognition (Luce, 1986; Luce & Pisoni, 1998), whereby words are organized in memory into "similarity neighborhoods" based on their frequency of occurrence in the language and the density of words in their lexical neighborhood. In adults, words occurring often in the language from sparse lexical neighborhoods (i.e., few phonemically similar words, or lexical neighbors with which to compete for lexical selection) are recognized faster and more accurately than words occurring infrequently from dense lexical neighborhoods (i.e., many similar sounding words with which to compete for lexical selection). The construct of neighborhood density emphasizes linguistic processing and retrieval issues above the level of peripheral perceptual and production system operations, focusing on psycholinguistic processes rather than neural variables supporting cognition. Incorporation of linguistic processing variables emphasizes the "functional load" of cognitive processing as foundational to emergence of intelligible speech. This inclusion implies that speech acquisition is not modular in the sense portrayed in phonological treatments but based on a complex system necessarily including cognitive processing. The size of the child's lexicon is considered integral to accuracy of retrieval for phonological forms. Like many other models related to the process of speech acquisition presented in this chapter, this particular model is not inclusive of the whole process of acquisition, only of the phase of the process related to full emergence of word intelligibility. It strongly emphasizes the interface of the lexicon with phonological processing, presenting a challenge to hypotheses of modularity common in phonological theories.

4 Summary

Contemporary theoretical perspectives and research paradigms considering the nature of speech acquisition emerge from a widely varied milieu of philosophical traditions on the origins of complex knowledge. Synthesis will await some resolution of the philosophical foundations from which these diverse perspectives emerge. As well, considerations of the whole process of acquisition and how varied available proposals encompass that process forms a necessary part of an overarching synthesis. The challenge for contemporary study of phonological acquisition is to continue multidisciplinary interactions in spite of the lack of coherence of present proposals. With continued exposure, these paradigms can begin to achieve a true scholarly level of cross-fertilization with the comprehensiveness necessary to move toward a coherent theory of speech acquisition.

REFERENCES

Abler, W. L. (1989) On the particulate principle of self-diversifying systems, *Journal of Social and Biological Structure*, 12, 1–13.

Archangeli, D. & Pulleyblank, D. (1994) *Grounded Phonology*. Cambridge, MA: MIT Press.

Baker, E. (2006) Management of speech impairment in children: The journey so far and the road ahead. *Advances in Speech-Language Pathology*, 8, 156–64.

Bernhardt, B. H. & Stemberger, J. P. (1998) *Handbook of Phonological Development: From a Nonlinear Constraints-Based Perspective.* San Diego, CA: Academic Press.

Boë, L.-J. & Maeda, S. (1997) Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux-nés et des adultes: Conséquences pour l'ontogenèse et la phylogenèse. *Journées d'Etudes Linguistiques de Nantes 1997: La Voyelle dans Tous ces États* (pp. 98–105).

Boliek, C. A., Hixon, T. J., Watson, P. J., & Morgan, W. J. (1996) Vocalization and breathing during the first year of life. *Journal of Voice*, 10, 1–22.

Boliek, C. A., Hixon, T. J., Watson, P. J., & Morgan, W. J. (1997) Vocalization and breathing during the second and third years of life. *Journal of Voice*, 11, 373–90.

Browman, C. P. & Goldstein, L. (1992) Articulatory phonology: An overview. *Phonetica*, 49, 155–80.

Bybee, J. (2001) *Phonology and Language Use*. Cambridge: Cambridge University Press.

Callan, D. E., Kent, R. D., Gunther, F. H., & Vorperian, H. K. (2000) An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech and Hearing Research*, 43, 721–36.

Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York: Harper & Row.

Davis, B. L. & MacNeilage, P. F. (1990) Acquisition of correct vowel production: A quantitative case study. *Journal of Speech and Hearing Research*, 33, 16–27.

Davis, B. L. & MacNeilage, P. F. (1995) The articulatory basis of babbling. *Journal of Speech and Hearing Research*, 38, 1199–1211.

Davis, B. L., MacNeilage, P. F., & Matyear, C. (2002) Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica*, 59, 75–107.

de Boer, B. (2001) *The Origins of Vowel Systems*. Oxford: Oxford University Press.

de Boer, B. (2005) Evolution of speech and its acquisition, *Adaptive Behavior*, 13, 281–92.

Dekkers, J., Leeuw, F. van der, & Weijer, J. van de (2000) *Optimality Theory: Phonology, Syntax, and Acquisition*. New York: Oxford Press.

Descartes, R. (1824) La dioptrique. In V. Coursin (ed.), *Oeuvres de Descartes*, trans. M. D. Boring. Paris: n.p. (Original work published 1638.)

DePaolis, R. A. (2006) The influence of production on the perception of speech. In D. Bamman, T. Magnitskaia & C. Zaller (eds.), *Proceedings of the 30th Boston University Conference on Language Development* (pp. 142–53). Somerville, MA: Cascadilla Press.

Edwards, J., Beckman, M. E., & Munson, B. (2004) The interaction between vocabulary size and phonotactic probability effects on children's production, accuracy and fluency in nonword repetition. *Journal of Speech Language and Hearing Research*, 47, 421–36.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971) Speech perception in infants. *Science*, 171, 303–6.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fentress, J. C. (1984) The development of coordination. *Journal of Motor Behavior*, 16, 99–134.

Fodor, J. A. (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980) Implications for speech production of a general theory of action. In B. Butterworth (ed.), *Language Production* (pp. 373–420). New York: Academic Press.

Gathercole, S. E. & Baddeley, A. D. (1993) *Working Memory and Language.* Mahwah, NJ: Lawrence Erlbaum.

Gelman, S. A. (2003) *The Essential Child: Origins of Essentialism in Everyday Thought.* New York: Oxford University Press.

Green, J. R., Moore, C. A., & Reilly, K. J. (2002) The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*, 45, 66–79.

Green, J. R., Moore, C. A., Ruark, J. L., Rodda, P. R., Morvee, W., & VanWitzenburg, M. (1997) Development of chewing in children from 12 to 48 months: Longitudinal study of EMG patterns. *Journal of Neurophysiology*, 77, 2704–16.

Goad, H. & Rose, Y. (2003) Segmental-prosodic interaction in phonological development: A comparative investigation. Special issue of *Canadian Journal of Linguistics*, 48, 139–452.

Greenberg, J. (1966) Synchronic and diachronic universals in phonology. *Language*, 42, 508–17.

Guenther, F. H. (1995) A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, 72, 43–53.

Guenther, F. H. & Perkell, J. S. (2004) A neural model of speech production and its application or studies of the role of auditory feedback in speech, In B. Maassen, R. Kent, H. Peters, P. H. H. M. van Lieshout, & W. Hulstijn (eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 29–50). New York: Oxford University Press.

Hardcastle, W. J. & Hewlett, N. (1999) *Coarticulation, Theory, Data, and Techniques*. Cambridge: Cambridge University Press.

Hauser, M. D. (1996) *The Evolution of Communication*. Cambridge, MA: MIT Press.

Hauser, M. D., Chomsky, N., & Fitch, T. (2002) The facility of language: What is it, who has it, and how did it evolve? *Science*, 298, 1569–79.

Holmgren, K., Lindblom, B., Aurelius, G., Jalling, B., & Zetterstrom, R. (1986) On the phonetics of infant vocalization. In B. Lindblom & R. Zetterstrom (eds.), *Precursors of Early Speech* (pp. 51–63). New York: Stockton Press.

Jakobson, R. (1968) *Child Language, Aphasia, and Phonological Universals*. The Hague: Mouton.

James, W. (1890) *The Principles of Psychology*. New York: Henry Holt.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986) The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, 14, 29–59.

Kant, I. (1924) *Critique of Pure Reason* (trans F. M. Miller). New York: MacMillan. (Original work published in 1781.)

Kauffman, S. (1995) *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. New York: Oxford University Press.

Keating, P. A. (1990) The window model of coarticulation: Articulatory evidence. In J. Kingston & M. E. Beckman (eds.), *Papers in Laboratory Phonology, I: Between the Grammar and Physics of Speech* (pp. 451–70). Cambridge: Cambridge University Press.

Kent, R. D. & Bauer, H. R. (1985) Vocalizations of one year olds. *Journal of Child Language*, 12, 491–526.

Kent, R. D. & Vorperian, H. K. (1995) *Anatomic Development of the Craniofacial-Oral-Laryngeal Systems*. San Diego: Singular Publishing Group Inc.

Kern, S. & Davis, B. L. (2009) Emergent complexity in early vocal acquisition: Cross-linguistic comparisons of canonical babbling. In I. Chirotan, C. Coupé, E. Marsico, & F. Pellegrino (eds.), *Approaches to Phonological Complexity*. Berlin: Mouton de Gruyter.

Kirby, S. (1999) *Function, Selection and Innateness: The Emergence of Language Universals.* New York: Oxford University Press.

Kohler, K. (2000) Investigating unscripted speech: Implications for phonetics and phonology. *Phonetica*, 57, 85–94.

Koopmans-van Beinum, F. & Van der Steldt, J. (1986) Early stages in the development of speech movements. In B. Lindblom & R. Zetterstrom (eds.), *Precursors of Early Speech* (pp. 37–49). New York: Stockton Press.

Lee, S. (2003) Perceptual influences on speech production in Korean learning infant babbling. Unpublished dissertation, The University of Texas at Austin, Austin, Texas.

Lindblom, B. (1992) Phonological units as adaptive emergents of lexical development. In C. Ferguson, L. Menn, & C. Stoel-Gammon, *Phonological Development* (pp. 131–63). Parkton: York Press.

Lindblom, B. (2004) The organization of speech movements: Specification of units and modes of control. In J. Slifka, S. Manuel, & M. Matthies (eds.), *From Sound to Sense: 50+ Years of Discoveries in Speech Communication* (pp. 164–72). Boston, MA: MIT.

Locke, J. & St John, J. A. (1854) *The Works of John Locke*. London: Henry G. Bohn.

Locke, J. L. (1983) *Phonological Acquisition and Change*. New York: Academic Press.

Locke, J. L. (1993) *The Child's Path to Spoken Language*. Cambridge, MA: Harvard University Press.

Luce, P. A. (1986) *Neighborhoods of Words in the Mental Lexicon* (Research on Speech Perception Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A. & Pisoni, D. B. (1998) Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–36.

MacNeilage, P. F. & Davis, B. L. (1993) Motor explanations of babbling and early speech patterns. In B. Boysson-Bardies, S. de Schoen, P. Jusczyk, P. MacNeilage, & J. Morton (eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 341–52). Dordrecht: Kluwer.

MacNeilage, P. F. & Davis, B. L. (2000) Origin of the internal structure of words. *Science*, 288, 527–31.

MacNeilage, P. F., Davis, B. L., Kinney, A., & Matyear, C. M. (2000) The motor core of speech: A comparison of serial organization patterns in infants and languages, *Child Development*, 71, 153–63.

MacNeilage, P. F., Davis, B. L., Matyear, C. L., & Kinney, A. (1999) Origin of speech output complexity in infants and in languages. *Psychological Science*, 10, 459–60.

Maddieson, I. (1986) *Patterns of Sounds*, 2nd edn. Cambridge: Cambridge University Press.

McCarthy, J. J. & Prince, A. S. (1993) Prosodic morphology: Constraint interaction and satisfaction. Unpublished manuscript, University of Massachusetts, Amherst & Rutgers University, New Brunswick.

Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004) Role of vocal tract morphology in speech development: Perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood. *Journal of Speech, Language, and Hearing Research*, 47, 1059–80.

Moore, C. A., Caulfield, T. J., & Green, J. R. (2001) Relative kinematics of the rib cage and abdomen during speech and non-speech behaviors of 15-month-old children. *Journal of Speech, Language, and Hearing Research*, 44, 80–94.

Newmeyer, Frederick J. (1998) *Language Form and Language Function*. Cambridge, MA: MIT Press.

Nittrouer, S., Studdert-Kennedy, M., & McGown, R. S. (1989) The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel sequences spoken by children and adults. *Journal of Speech and Hearing Research*, 32, 120–32.

Oller, D. K. (1980) The emergence of the sounds of speech in infancy. In G. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (eds.), *Child Phonology, I: Production* (pp. 93–112). New York: Academic Press.

Oller, D. K. (2000) *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum.

Oller, D. K. & Griebel, U. (2008) The Origins of syllabification in human infancy and in human evolution. In Davis, B. L. and Zajdo, K. (Eds.) *The Syllable in Speech Production* (pp. 29–62). New York: Routledge/Taylor & Francis Publishers.

Pater, J., Stager, C., & Werker, J. (2004) The perceptual acquisition of phonological contrasts. *Language*, 80, 384–402.

Pesetsky, D. (1999) Linguistic universals and universal grammar. In R. A. Wilson & F. Keil (eds.), *MIT Encyclopedia of Cognitive Science* (pp. 476–8). Cambridge, MA: MIT Press.

Pierrehumbert, J. (2003) Probabilistic theories of phonology. In R. Bod, J. B. Hay, & S. Hannedy (eds.), *Probability Theory in Linguistics* (pp. 177–228). Cambridge, MA: MIT Press.

Prince, A. & Smolensky, P. (1993) *Optimality Theory: Constraint Interaction in Generative Grammar.* Technical Report, Rutgers University Center for Cognitive Science and Computer Science Department, University of Colorado at Boulder.

Prince, A. & Smolensky, P. (2004) *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell.

Rose, Y. (2009) Internal and external influences on child language productions. In I. Chirotan, C. Coupé, E. Marsico, & F. Pellegrino (eds.), *Approaches to Phonological Complexity*. Berlin: Mouton de Gruyter.

Roug, L., Landburg, I., & Lundburg, L. J. (1989) Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life. *Journal of Child Language*, 17, 19–40.

Rumelhart, D. E. & McClelland, J. L. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, I: Foundations*. Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996) Statistical learning by 8-month old infants. *Science*, 274, 1926–8.

Saltzman, E. L. & Munhall, K. G. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1, 333–82.

Smith, A. & Goffman, L. (1998) Stability and patterning of speech movement sequences in children and adults. *Journal of Speech, Language, and Hearing Research*, 41, 18–30.

Smith, A., Johnson, J., McGillem, C., & Goffman, L. (2000) On the assessment of stability and patterning of speech movement sequences. *Journal of Speech, Language, and Hearing Research*, 43, 277–86.

Stager, C. L. & Werker, J. F. (1997) Infants listen for more phonetic detail in speech perception than in word learning tasks. *Nature*, 388, 381–2.

Stark, R. (1980) Stages of speech development in the first year of life.

In G. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (eds.), *Child Phonology, I: Production* (pp. 113–42). New York: Academic Press.

Steels, L. (2006) Experiments on the emergence of human communication. *Trends in Cognitive Sciences*, 10, 347–9.

Stemberger, J. P. & Bernhardt, B. H. (1999) The emergence of faithfulness. In B. MacWhinney (ed.), *The Emergence of Language* (pp. 417–46). Mahwah, NJ: Lawrence Erlbaum.

Stoel-Gammon, C. (1985) Phonetic inventories 15–24 months: A longitudinal study. *Journal of Speech and Hearing Research*, 23, 506–12.

Studdert-Kennedy, M. (1998) The particulate origins of language generativity: From syllable to gesture. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases* (pp. 202–21). Cambridge: Cambridge University Press.

Teixeira, E. R. & Davis, B. L. (2002) Early sound patterns in the speech of two Brazilian Portuguese speakers, *Language and Speech*, 45, 179–204.

Tomasello, M. (1998) *Constructing a Language*. Cambridge, MA: Harvard University Press.

Trubetzkoy, Nikolai S. (1929) Zur allgemeinen Theorie des phonologischen Vokalsystems. *Travaux du Cercle Linguistique de Prague*, 1, 39–67.

Vihman, M. M. (1993) Variable paths to early word production. *Journal of Phonetics*, 21, 61–82.

Vihman, M. M., Ferguson, C. E., & Elbert, M. (1986) Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, 7, 3–40.

Vihman, M. M. & Kunnari, S. (2006) The sources of phonological knowledge. *Recherches Linguistiques de Vincennes*, 1, 133–64.

Vihman, M. M. & Velleman, S. (2000) The construction of a first phonology. *Phonetica*, 57, 255–66.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005) Development of vocal tract length during childhood: A Magnetic Resonance Imaging study. *Journal of the Acoustical Society of America*, 117, 338–50.

Werker, J. F. & Tees, R. C. (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.

FURTHER READING

Jackendoff, R. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford: Oxford University Press.

MacNeilage, P. F. & Davis, B. L. (2005) The evolution of language and speech. *Evolutionary Psychology Handbook* (pp. 698–723). Cambridge, MA: MIT Press.

Marcus, G. (2003) *The Birth of the Mind: How a Tiny Number of Genes Creates the Complexities of Human Thought*. New York: Basic Books.

Vihman, M. M. (1996) *Phonological Development: The Origins of Language in the Infant*. Oxford: Blackwell.

9 Coarticulation and Connected Speech Processes

EDDA FARNETANI AND
DANIEL RECASENS

1 Speech Contextual Variability

1.1 Coarticulation

A fundamental and extraordinary characteristic of spoken language, of which we speakers are not even conscious, is that the movements of different articulators for the production of successive phonetic segments overlap in time and interact with one another: as a consequence, the vocal tract configuration at any point in time is influenced by more than one segment. This is what the term "coarticulation" describes. The acoustic effects of coarticulation can be observed by means of spectrographic analysis: any acoustic interval, auditorily defined as a phonetic segment, will show the influence of neighboring phones in various forms and degrees. These effects are usually not audible, which is why their descriptive and theoretical study in various languages became possible only after physiological and acoustical methods of speech analysis became available and widespread during the last 40 years.

Table 9.1 shows how coarticulation can be described in terms of: (1) the main articulators involved; (2) some of the muscles considered to be primarily responsible for the articulatory movements; (3) the movements that usually overlap in contiguous segments; (4) the major acoustic consequences of such overlap. As for lingual coarticulation, the tongue tip/blade and the tongue body can act quasi-independently as two distinct articulators, so that their activity in the production of adjacent segments may overlay in time.

Jaw movements are not included in the table since the jaw contributes both to lip and to tongue positioning and, therefore, may be considered part of the labial and lingual subsystems. Mandibular movements are analyzed especially when the goal of the experiment is to establish the role of the jaw in shaping the vocal tract and thus distinguish between active and passive tongue or lip movements, or to investigate how the jaw contributes to or compensates for coarticulatory variations.

Table 9.1 Coarticulation: Levels of description

Articulators	*Muscles*	*Articulatory activity*	*Acoustic consequences*
LIPS	Orbicularis oris/Risorius	Lip rounding/ spreading	Formant changes
TONGUE	Genioglossus, and other extrinsic and intrinsic lingual muscles	Tongue front/back and high/low displacement	Formant changes
VELUM	Levator palatini	Velum lowering	Nasal formants and changes in oral formant structure
LARYNX	Posterior cricoarytenoid/ Lateral cricoarytenoid, Interarytenoid	Vocal fold abduction/ adduction	Signal (a)periodicity

Examples of coarticulation in terms of muscle activity and of articulatory movements are given in Figures 9.1 and 9.2. They show overlapping activity both at the level of commands and that of execution.

Data from Hirose and Gay (1972) illustrate coarticulation at the myomotoric level. Electromyographic activity of four muscles during the production of the sequences /əpɪb/ and /əpɪp/ indicates that the activity of orbicularis oris for the production of the first /p/ is overlapped by the activity of the genioglossus for the production of the following front vowel. Moreover the activity of the laryngeal muscles responsible for abducting and adducting the vocal folds also overlap: the onset of lateral cricoarytenoid activity (LCA, adducting) occurs when the posterior cricoarytenoid activity (PCA, abducting) is at its peak, that is, at the middle of the /p/ closure.

Figure 9.1 describes tongue-tip–tongue-body coarticulation in the /kl/ cluster of the English word *weakling*, analyzed with electropalatography (EPG) and synchronized with oral and nasal airflow and with the acoustic signal (from Hardcastle, 1985). The sequence of the EPG frames (sampled every 7.5 ms) describes the articulation of the /kl/ cluster: it can be seen that the tongue-body closure for the velar consonant is overlapped by the tongue-tip/blade gesture for the following /l/, detectable by a light front contact as early as frame 130. The following frames show complete overlap of /k/ and /l/ closures for about 20 ms.

Figure 9.2 is an example of velar and lingual coarticulation in the sequences /ˈana/ and /ˈini/ in Italian, analyzed with EPG and oral/nasal flow measurements (Farnetani, 1986). As for velar coarticulation, the nasal flow curves (continuous

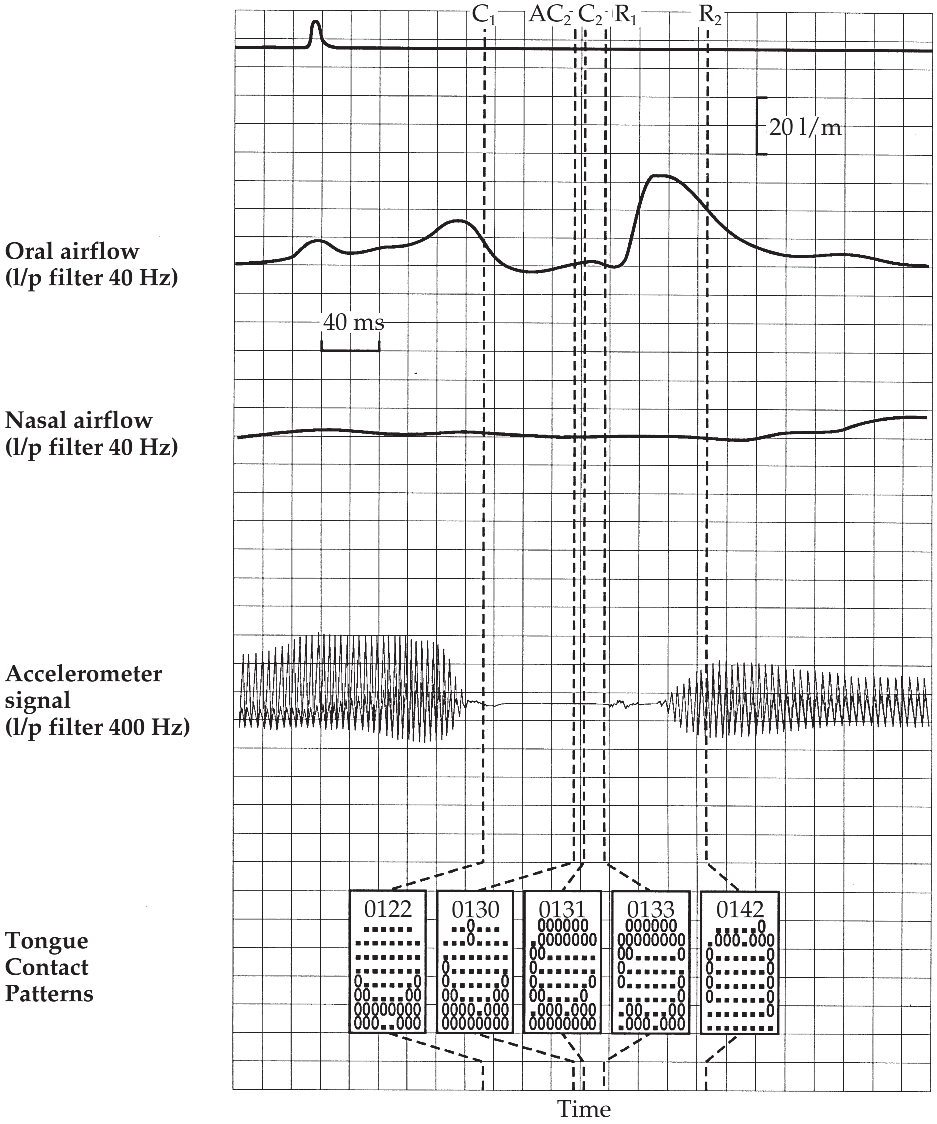


Figure 9.1 Oral and nasal flow curves, acoustic signal, and synchronized EPG activity in the production of the English word *weakling* within phrase. (From Hardcastle, 1985)

thick lines) indicate that in /ˈana/ the opening of the velopharyngeal port for the production of /n/ occurs just after the acoustic onset of the initial /a/ and lasts until the end of the final /a/; in /ˈini/ there is only a slight anticipation of velopharyngeal opening during the initial /i/, but after /n/ the port remains open until the end of the utterance. Thus, velar C-to-V coarticulation extends both in the

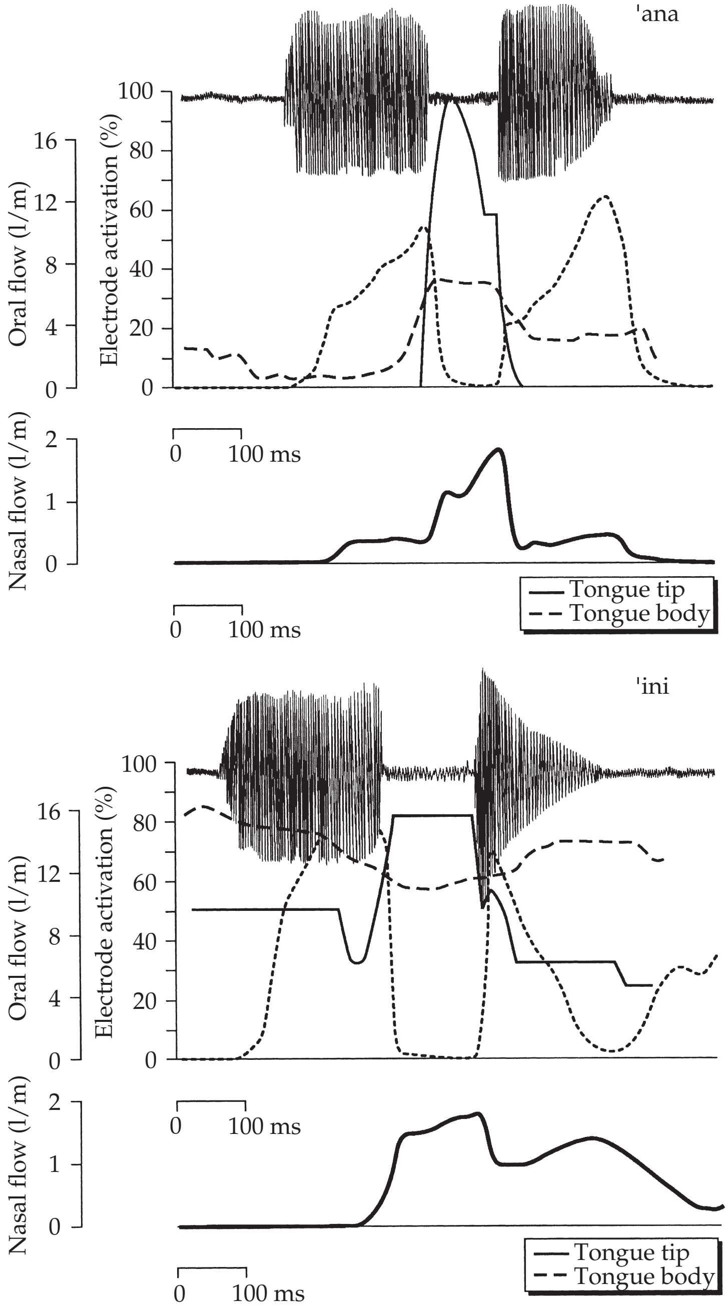


Figure 9.2 Acoustic signal, oral and nasal flow curves, and synchronized EPG curves during /'ana/ and /'ini/ produced in isolation by an Italian subject. (From Farnetani, 1986)

anticipatory and in the carryover direction in the sequence /ˈana/, while extending mostly in the carryover direction in the sequence /ˈini/. The two EPG curves represent the evolution of tongue-to-palate contact over time. It can be seen that during tongue-tip/blade closure for /n/, tongue-body contact is much larger in the context of /i/ than in the context of /a/, indicating that the tongue-body configuration during the consonant is strongly affected by the vowels. These patterns describe V-to-C lingual coarticulation, that is, the effect of the vowel on the articulation of the consonant.

These examples clearly show that speaking is coarticulating gestures. The central theoretical issues in the studies of coarticulation concern its origin, function, and control. Coarticulation has been observed in all languages so far analyzed, and can be considered a universal phenomenon, even if it appears to differ among languages. Before exploring these issues, the assimilatory and connected speech processes will be discussed at some length in the next sections.

1.2 Assimilation

Assimilation refers to contextual variability of speech sounds, by which one or more of their phonetic properties are modified and become like those of the adjacent segment. Are assimilation and coarticulation qualitatively different processes, the former reflecting an auditory approach to phonetic analysis, and the latter an instrumental articulatory/acoustic approach? The answers are various and controversial. Standard generative phonology (Chomsky & Halle, *The Sound Pattern of English*, 1968) makes a clear-cut distinction between assimilation and coarticulation. Assimilation pertains to the domain of linguistic competence, is accounted for by phonological rules, and refers to modifications of features defined as the minimal categorical-classificatory constituents of a phoneme; hence, assimilatory processes are part of the grammar and are language-specific. Coarticulation, by contrast, results from the physical properties of the speech mechanism and is governed by universal rules; hence, it pertains to the domain of performance and cannot be part of the grammar. Chomsky and Halle include as coarticulation effects "the transition between a vowel and an adjacent consonant, the adjustments in vocal tract shape made in anticipation of subsequent motions, etc." (1968, p. 295).

Quite often context-dependent changes involving the same articulatory structures have different acoustic and perceptual manifestations in different languages so that it is possible to distinguish what can be considered universal phonetic behavior from language-particular rules. A classic example is the difference between vowel harmony, an assimilatory process present in a limited number of languages such as Hungarian, and vowel-to-vowel coarticulation, a process attested in a number of languages and probably present in all (Fowler, 1983). In other cases, cross-language differences are not easily interpretable, and inferences on the nature of the underlying production mechanisms can be made by manipulating some of the speech parameters, for example segmental durations. In a study of vowel nasalization in Spanish and American English, Solé and Ohala (1991) were able to distinguish phonological (language-specific) nasalization from phonetic (universal) nasalization by manipulating speech rate. They found a quite different distribution of the temporal patterns of nasalization in the two languages as a function of speaking rate: the extent of nasalization on the vowel preceding the nasal consonant was proportional to the varying vowel duration in American English, while it remained constant in Spanish. A temporal increase of nasalization as vowel duration increases in American English must be intentional, i.e. phonological; on the other hand, the short, constant extent of nasalization in Spanish must be an automatic consequence of the speech mechanism, as it reflects the minimum time necessary for the lowering gesture of the velum.

But a strict dichotomy between universal and language-specific variations fails to account for the many cross-language data showing that coarticulation differs in degree across languages. A typical example is Clumeck's study on velar coarticulation, which was found to differ in temporal extent across all the six languages analyzed (Clumeck, 1976). In this case, what criterion can be used to decide in which language the patterns are unintentional and automatic, and in which they are to be ascribed to the grammar?

Likewise, within a given language, context-dependent variations may exhibit different articulatory patterns which can be interpreted either as the result of different underlying processes, or just as quantitative variations resulting from the same underlying mechanism. For example, Figure 9.3 shows the variations in the articulation of the phoneme /n/ as a function of the following phonetic segment, a palatal semivowel (in /anja/) and a postalveolar or alveolopalatal affricate consonant (in /antʃa/). The utterances are pseudowords produced in isolation by an Italian speaker, and analyzed with EPG.

In each graph the two curves represent the evolution of the tongue-to-palate contact over time. We can see that in /anja/, as the tongue tip/blade (continuous line) achieves maximum front contact for /n/ closure, the tongue body (dashed line) moves smoothly from /a/ to /j/ suggesting overlapping activity of two distinct articulators, and two distinct goals. In /antʃa/, instead, the typical /n/ configuration has disappeared; the front contact has decreased by some percentage points, and the back contact has appreciably increased; the cluster seems to be produced with one tongue movement. These differences may indicate that two distinct processes are at work in the two utterances: anticipatory coarticulation of /j/ on /n/ in the former, and place assimilation of /n/ to /tʃ/ in the latter.

Current theories of coarticulation have controversial views on whether there are qualitative or quantitative or even no differences between assimilatory and coarticulatory processes. It will be seen that at the core of the different positions are different answers to the fundamental issues addressed above, i.e., the domain, the function and the control of coarticulation.

1.3 Connected speech processes

In speech, the phonetic form of a word is not invariable but can vary as a function of a number of linguistic, communicative, and pragmatic factors (e.g., information

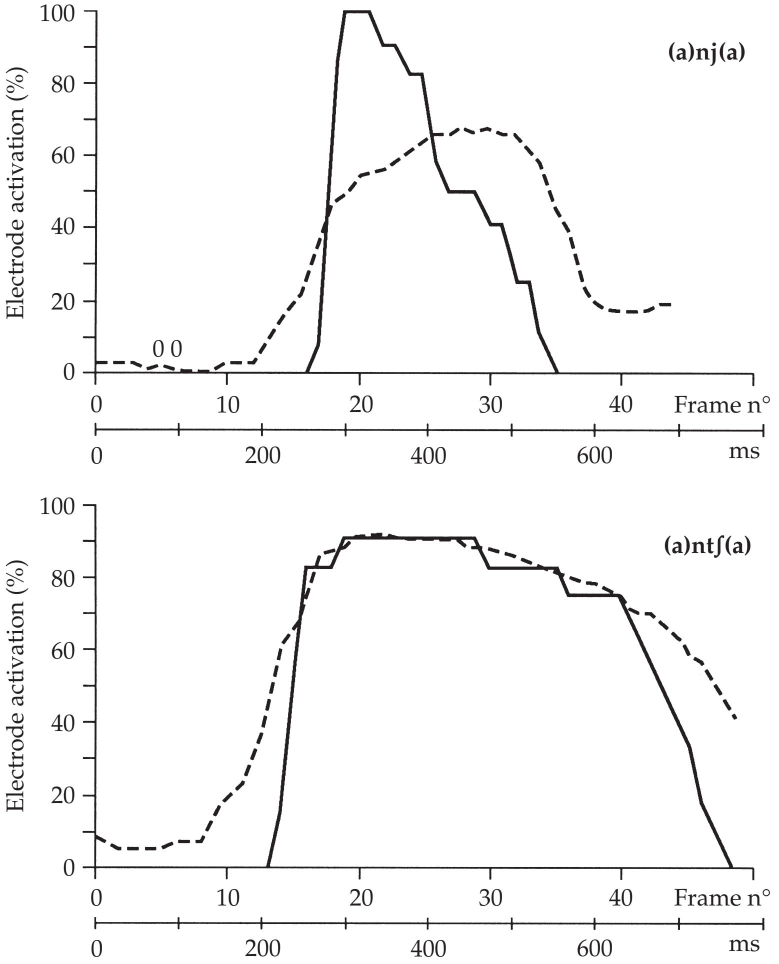


Figure 9.3 EPG curves during /anja/ and /antʃa/ produced by an Italian subject. Continuous lines: tongue-tip/blade contact; dashed lines: tongue-body contact. (From Farnetani, 1986)

structure, style, communicative situation). These variations are generally referred to as "alternations," and the phonetic processes accounting for them have been termed "connected speech processes" (Jones, 1969; Gimson, 1970). According to Gimson (p. 287), these processes describe the phonetic variations characterizing continuous speech when compared to a word spoken in isolation. The author makes a detailed list of connected speech processes in English: assimilation of place, manner, and voicing; reduction of vowels to schwa in unaccented words; deletion of consonants and vowels. According to the author, the factors that contribute to modify the phonetic properties of a word are "the pressures of its sound environment or of the accentual or rhythmic group of which it forms part," and the speed of the utterance.

Kohler (1990) proposes an explanatory account of connected speech processes in German. Basing his analysis of German on the differences between careful and casual pronunciation of the same items, the author arrives at the conclusion that the so-called connected speech processes are a global phenomenon of reduction and articulatory simplification. These processes include /r/ vocalization to [ɐ] (when not followed by a vowel), weak forms, elision, and assimilation. From an analysis of the sound categories undergoing such changes, he infers that connected processes result from articulatory constraints, such as minimization of energy, which induce a reorganization of the articulatory gestures. He also proposes a formalization of the processes in terms of sets of phonetic rules, which generate any reduced segmental pronunciation.

The two accounts, although substantially different in their perspectives, have in common the assumption that connected speech processes imply modifications of the basic units of speech, i.e., elimination and replacements of articulatory gestures, changes in articulation places, etc. Hence, the main difference between connected speech processes and the phonological assimilations described in section 1.2 is that the latter occur independently of how a word is pronounced, while the former occur in some cases (e.g., in rapid, casual speech), but are absent in others (Chomsky & Halle, 1968, p. 110).

Recent theories of coarticulation also consider connected speech processes and propose their own accounts, as will seen below in section 2.6.

2 Theoretical Accounts of Coarticulation

2.1 Pioneering studies: Joos' overlapping innervation theory

That speech is a continuum, rather than an orderly sequence of distinct sounds as listeners perceive it, was pointed out long ago (Sweet, 1877, cited by Wood, 1993). Sweet saw speech sounds as points "in a stream of incessant change" and this promoted the view that coarticulatory effects result from the transitional movements conjoining different articulatory targets, and reflected acoustically in the transitions to and from targets. Menzerath and de Lacerda (1933), to whom the term "coarticulation" is attributed, showed that segments can be articulated together, not merely conjoined to each other. The pioneering acoustic analysis of American English vowels conducted by Joos (1948) revealed that vowels vary as a function of neighboring consonants not only during the transitional periods but also during their steady state. Referring to temporal evolution of the second formant, Joos observed that "the effect of each consonant extends past the middle of the vowel, so that at the middle the two effects overlap" (p. 105). In his theoretical account of the phenomenon, he contests the "glide" hypothesis, which attributes coarticulation to the inertia of vocal organs and muscles: since no shift from one articulatory position to another can take place instantaneously, a transition intervenes between successive phones. Joos proposes instead the "overlapping innervation wave theory" (p. 109): each phonetic segment command is an invariant "wave" that "waxes and wanes smoothly"; "waves for successive phones overlap in time."

As will be seen below, these early hypotheses on the sequential ordering of speech segments have been highly influential in the development of coarticulation theories.

2.2 Coarticulation as a component of the Grammar

The evolution of featural phonology after Chomsky and Halle (1968) is marked by a gradual appropriation of coarticulation into the domain of linguistic competence.

2.2.1 The theory of feature spreading Daniloff and Hammarberg (1973) and Hammarberg (1976) were the promotors of the "feature-spreading" account of coarticulation. The view that coarticulation is a pure physiological process due to mechano-inertial constraints of the speech apparatus entails a sharp dichotomy between intent and execution, and implies that articulators are unable to carry out the commands as specified. The way to overcome this dichotomy is to assume that coarticulation itself is part of the phonological component. The arguments in support of this assumption are: (1) phonology is prior to phonetics, i.e., the phonology component underlies the phonetic implementation of speech sounds; (2) phonological segments are abstract entities, and cannot be altered by the physical speech mechanism; (3) the speech mechanism can only execute higher level commands. Hence, the variations associated to coarticulation must be the input to the speech mechanism. How? Segments have inherent and derived properties. These latter result from coarticulation, which alters the properties of a segment. Phonological rules stipulate which features get modified, and the phonetic representation, which is the input of the speech mechanism, specifies the details of articulation and coarticulation.

The departure from Chomsky and Halle's view of coarticulation was probably necessary in face of emerging data on anticipatory lip protrusion (Daniloff & Moll, 1968) and velar coarticulation (Moll & Daniloff, 1971) showing that coarticulatory movements can be initiated at least two segments before the influencing one. These findings revealed that coarticulation is not the product of inertia. Another reason (Hammarberg, 1976) was that coarticulation cannot be accounted for by universal rules, owing to interlanguage differences.

Why does coarticulation occur? The function of coarticulation is to smooth out the differences between adjacent sounds: coarticulatory modifications accommodate the segments so that the transitions between them are minimized. In Daniloff and Hammarberg's view anticipatory coarticulation is always a deliberate process, while carryover coarticulation is in part the effect of inertia and in part a feedback assisted strategy that accommodates speech segments to each other.

2.2.2 Henke's articulatory model The articulatory model of Henke (1966) best accounts for experimental data on the extent of coarticulation. It contrasts with another well-known account of coarticulation, the "articulatory syllable" model, proposed by Kozhevnikov and Chistovich (1965). This model is based on data on anticipatory labial coarticulation in Russian, where segments appeared to coarticulate within, but not across C_nV sequences. Unlike the C_nV model, Henke's model does not impose top-down boundaries on anticipatory coarticulation; instead, input segments are specified for articulatory targets in terms of binary phonological features (+ or −), unspecified features being given the value 0. Coarticulation rules assign a feature of a segment to all preceding unspecified segments by means of a look-ahead scanning mechanism. The spread of features is blocked by a specified feature: for example the feature [+nasal] will be anticipated to all preceding segments unspecified for nasality.

2.2.3 Feature specification and coarticulation: Coarticulatory resistance While a number of experimental results are compatible with the hypothesis of feature spreading and the look-ahead mechanism, many others contradict the spatial and/or the temporal predictions of the model. First, the model cannot explain the extensive carryover effects observed in a number of studies on V-to-V coarticulation (for example, Magen, 1989; Recasens, 1989). Other disputed aspects of the theory are:

1 the adequacy of the concept of specified versus unspecified features for blocking or allowing coarticulation;
2 the hypothesis that a look-ahead mechanism can account for the temporal extent of anticipatory coarticulation.

As for the first issue, it appears that segments specified for a contradictory feature in asymmetric VCV sequences can nonetheless be modified by coarticulation. Data on lip rounding in French (Benguerel & Cowan, 1974) and English (Sussman & Westbury, 1981) indicate that in an /iC_nu/ sequence type, lip rounding for /u/ can start during /i/. Also, data on lingual coarticulation show that tongue displacement towards a vowel can begin during a preceding cross-consonantal vowel even if the two vowels are specified for conflicting features, for example, F2 and tongue-dorsum lowering effects from /a/ on /i/ (Öhman, 1966, for Swedish and English; Butcher & Weiher, 1976, for German; Farnetani et al., 1985, for Italian; Magen, 1989 for American English). Most interestingly, these transconsonantal V-to-V effects appear to vary in degree across languages, indicating that the same vowel categories are subject to different constraints that favor or disfavor coarticulatory variations in different languages (see Manuel & Krakow, 1984, comparing Swahili, Shona and English; Manuel, 1987, comparing three Bantu languages; Choi & Keating, 1991, comparing English, Polish, Russian, and Bulgarian).

As for phonologically unspecified segments, some experimental data are compatible with the idea that they completely acquire a contextual feature. Figure 9.4

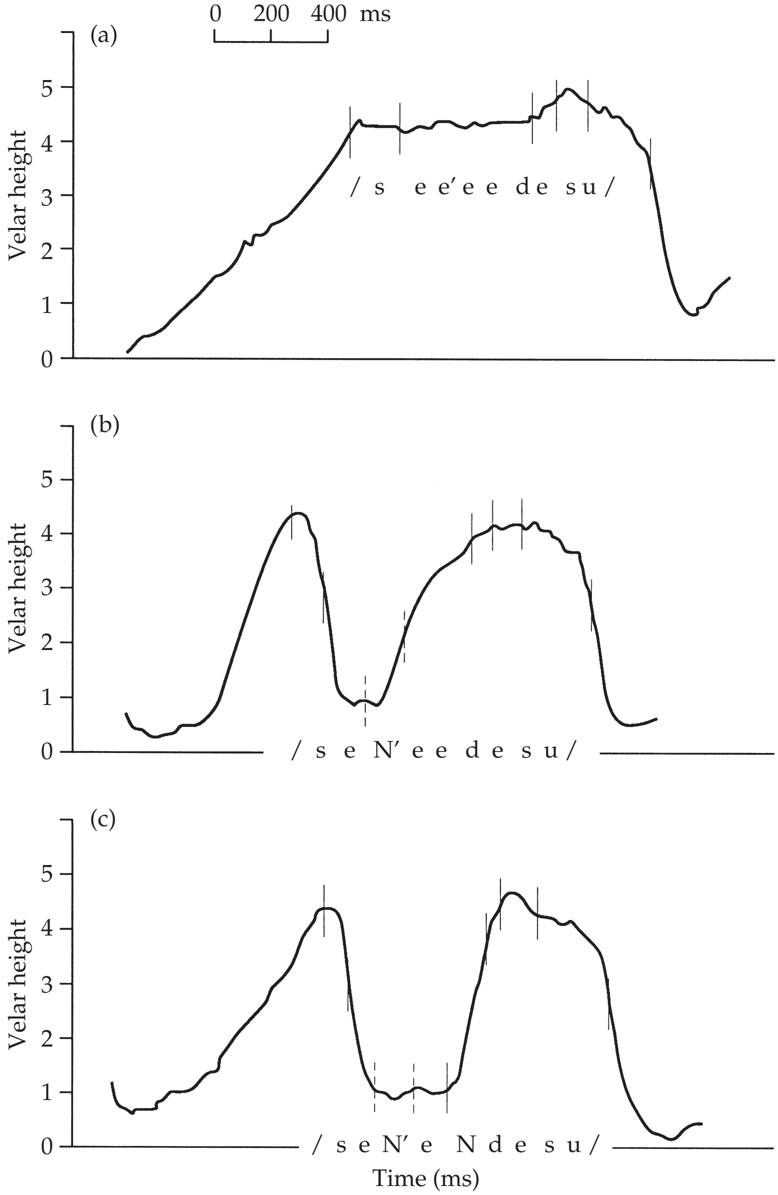


Figure 9.4 Velar movement observed with fiberscope during Japanese utterances containing the vowel /e/ in oral and nasal contexts. (From Ushijima & Sawashima, 1972)

(from Ushijima & Sawashima, 1972) illustrates how, as predicted by the feature-spreading model, the Japanese vowel /e/ unspecified for nasality acquires the nasality feature in a symmetric context. The figure shows the amount of velum height during the vowel /e/ in Japanese, surrounded by oral consonants (panel a), by nasal consonants (panel c), and in a mixed environment (panel b). It can be seen that during /e/ the velum is as high as for the oral consonants in (a), and as low as for the nasal consonants in (c): in both cases the velum height curve runs nearly flat across the vowel. Instead, in the asymmetric example (b) the curve traces a trajectory from a high to a low position during the /e/ preceded by /s/ and the reverse trajectory during the /e/ followed by /d/. The symmetric sequences show that /e/ is completely oral in an oral context and completely nasalized in a nasal context, indicating that this vowel has no velar target of its own but acquires that of the contextual phonetic segment(s). The trajectories in the asymmetric sequences do not contradict the hypothesis that this vowel has no target for velar position, but contradict the assumption that contextual features are spread in a categorical way. Accordingly, /e/ would have to be completely nasalized from its onset when followed by a nasal, and completely oral when followed by an oral consonants, and this does not seem to occur.

Many other data indicate that phonologically unspecified segments may nonetheless exhibit some resistance to coarticulation and, therefore, are specified for articulatory targets. English data on velar movements (Bell-Berti, 1980; Bell-Berti & Krakow, 1991) show that the oral vowels are not articulatorily neutral to velar height, and have their own specific velar positions even in a non-nasal environment. As for lip position, Engstrand's data (1981) show that in Swedish /u-u/ sequences with intervocalic lingual consonants, protrusion of the upper lip relaxes during the consonants and the curve may form a "trough" between the vowels, suggesting that such consonants are not neutral to lip position. Troughs in lingual muscle activity during /ipi/ were first observed by Bell-Berti and Harris in their 1974 study (see further below for a different account of troughs).

Subsequent research on lingual consonants in Catalan, Swedish, and Italian (Recasens, 1984a, 1984c, 1987; Engstrand, 1989; Farnetani, 1990, 1991) revealed that consonants unspecified for tongue-body features coarticulate to different degrees with the surrounding vowels. As for coronals, those studies show that the amount of tongue-body coarticulation tends to decrease from alveolars to postalveolars, and from liquids (provided that /l/ is clear and the rhotic is a tap or an approximant) to stops to fricatives.

Figure 9.5 is an example of how different consonants coarticulate to different degrees with the surrounding vowels /i/ and /a/ in Italian, as measured by EPG. The trajectories represent the amount of tongue-body contact over time during the coronals /t/, /d/, /z/, /ʃ/ and clear /l/ and the bilabial /p/ in symmetric VCV sequences. The /iCi/ trajectories exhibit troughs of moderate degree for most consonants; /z/ shows the largest deviation indicating that the production of this consonant requires a lowering of the tongue body from the /i/ position. In the context of /a/, the consonants /p/ and /l/ coarticulate strongly with this vowel, as they show little or no contact, while for /t/, /d/, and /z/ the tongue body

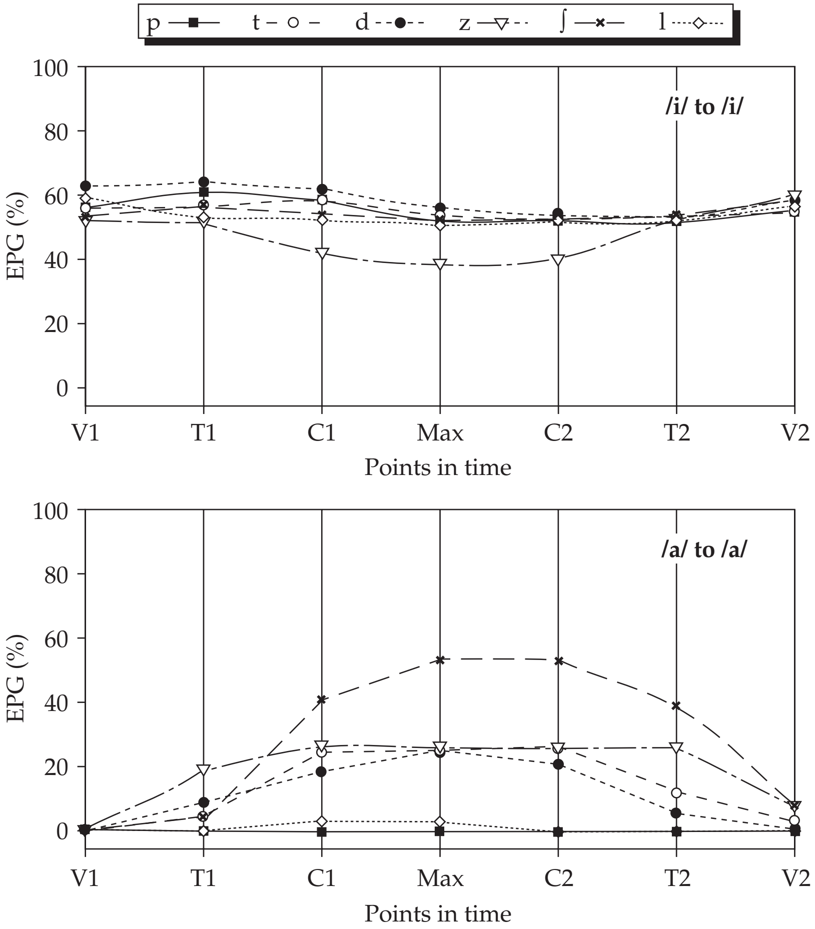


Figure 9.5 Tongue-body EPG contact at various points in time during symmetric (C)VCV isolated words in Italian. V1, V2 correspond to vowel mid-points; T1, T2 to offset of V1 and onset of V2 respectively; C1, C2 correspond to onset and release of consonant closures/constrictions respectively; Max refers to the point of maximum contact during the consonant. (From Farnetani, 1991)

needs to increase contact to about 20 percent. During /ʃ/, tongue-body contact reaches the same value, i.e., between 50 and 60 percent, in the two vocalic contexts, indicating that this consonant is maximally resistant to coarticulation.

The overall data on tongue-body V-to-C coarticulation indicate that no alveolar consonant fully coarticulates with the adjacent vowels, which suggests the presence of a functional and/or physical coupling between tip/blade and body. The differences in coarticulation across consonants can be accounted for by consonant-specific manner constraints: fricatives must constrain tongue dorsum position to ensure the appropriate front constriction and the intraoral pressure required for noise production; the production of stops and several laterals imposes lesser constraints, and allows for a wider range of coarticulatory variations.

The notion of coarticulatory resistance was introduced by Bladon and Al-Bamerni (1976) in an acoustic study of V-to-C coarticulation in /l/ allophones. Their data indicated that coarticulatory variations decrease from clear to dark to syllabic /l/. These graded differences could not be accounted for by binary feature analysis, which would predict complete blocking of coarticulation in the case of dark /l/ since this consonant variety is specified as [+back]. The authors propose a numerical index of coarticulatory resistance to be attached to the feature specification of each allophone. A subsequent study on tongue-tip/blade displacement in alveolar consonants in clusters (Bladon & Nolan, 1977) confirmed the idea that feature specification alone cannot account for the observed coarticulatory behavior. The coarticulation resistance scale has been developed more recently by the degree of articulatory constraint model of coarticulation or DAC (section 2.5.4 below).

All these studies show that the assignment of contextual binary features to unspecified segments through phonological rules fails to account for the presence versus absence of coarticulation, for its graded nature, and for the linguistically relevant aspects of coarticulation associated with this graded nature, i.e., the different degree of coarticulation exhibited by the same segments across languages. Explanation of these facts must be carried out in terms of articulatory, aerodynamic-acoustic, and perceptual constraints and, therefore, requires factors outside the world of phonological features.

2.3 *Coarticulation as speech economy*

2.3.1. Adaptive variability in speech The theoretical premise at the basis of Lindblom's theory of speech variability is that the primary scope of phonetics is not to describe how linguistic forms are realized in speech, but to explain and derive the linguistic forms from "substance-based principles pertaining to the use of spoken language and its biological, sociological, and communicative aspects" (Liljencrants & Lindblom, 1972, p. 859). Accordingly, in his theory of "Adaptive Variability" and "Hyper-/Hypo-speech" (Lindblom, 1983, 1989, 1990), phonetic variation is not viewed as mere consequence of inertia in the speech mechanism, but rather as a continuous adaptation of speech production to the demands of the communicative situation. Variation arises because production strategies change as a result of the interaction between system-oriented and output-oriented motor control. Some situations will require an output with a high degree of perceptual contrast, others will require less perceptual contrast and will allow more variability. Thus, the acoustic characteristics of the same item will exhibit a wide range of variation reflected along the continuum of over- to under-articulation, or hyper- to hypo-speech.

2.3.2 Low-cost and high-cost production behavior What is the function of coarticulation within the hyper–hypo framework? Coarticulation, manifested as a reduced displacement and a shift of articulatory movements towards the contextual phonetic segments, is a low-cost motor behavior, an economical way of speaking. Its pervasiveness indicates that the speech motor system, like other kinds of motor behavior, is governed by the principle of economy.

In his study on vowel reduction, Lindblom (1963) introduced the notion of acoustic target, an ideal context-free spectral configuration which, in the case of vowels, is represented by the asymptotic values towards which formant frequencies aim. Lindblom's study showed that targets are quite often not realized: his data on CVC syllables indicated that the formant frequency values at vowel midpoint change monotonically with changes in vowel duration. At long vowel durations, formants tend to reach the target values; as vowel duration decreases, the formant movements are reduced and tend to shift towards the values of the adjacent consonants. This target undershoot process is shown in Figure 9.6.

Its continuous nature reveals that vowel reduction is an articulatory process, largely dependent on duration, rather than a phonological process. Indeed, the direction of the change towards the segmental context, as well as different degrees of undershoot as a function of the extent of the CV transitions (i.e., vowel reduction is minimal when the consonant-to-vowel distance is small), indicate that reduction is ruled by a coarticulatory rather than by a centralization mechanism leading towards a schwa-like configuration.

Lindblom's account of the relation between duration, target undershoot, and coarticulation was that reduction is the automatic response of the motor system to an increase in the rate of the motor commands. When successive commands on one articulator are issued at very short temporal intervals, the articulator has insufficient time to complete the response before the next signal arrives, and has to respond to different commands simultaneously, thus inducing both vowel shortening and reduced formant displacement. Subsequent research showed that the system response to high rate commands does not automatically result in reduced movements (Kuehn & Moll, 1976; Gay, 1978), and that reduction can occur also at slow rates (Nord, 1986). These studies indicated that speakers can adapt to different speaking situations and choose different production strategies to avoid or to allow reduction/coarticulation.

In the revised model of vowel undershoot (Moon & Lindblom, 1994), vowel duration is still the main factor, but variables associated with speech style, such as the rate of formant frequency change, can substantially modify the amount of formant undershoot. The model is based on an acoustic study of American English stressed vowels produced in clear speech and in citation forms, i.e., in overarticulated versus normal speech. Data on vowel duration and F2 indicate that vowels tend to be longer and less reduced in clear speech than in citation forms. A second finding is that clear speech is in most cases characterized by larger formant velocity values than citation forms. This means that the degree of context-dependent undershoot depends on speech style and tends to decrease with an increase in velocity of the articulatory movements. In this model where the speech motor mechanism is seen as a second-order mechanical system, it is proposed that undershoot is controlled by three variables reflecting the articulation

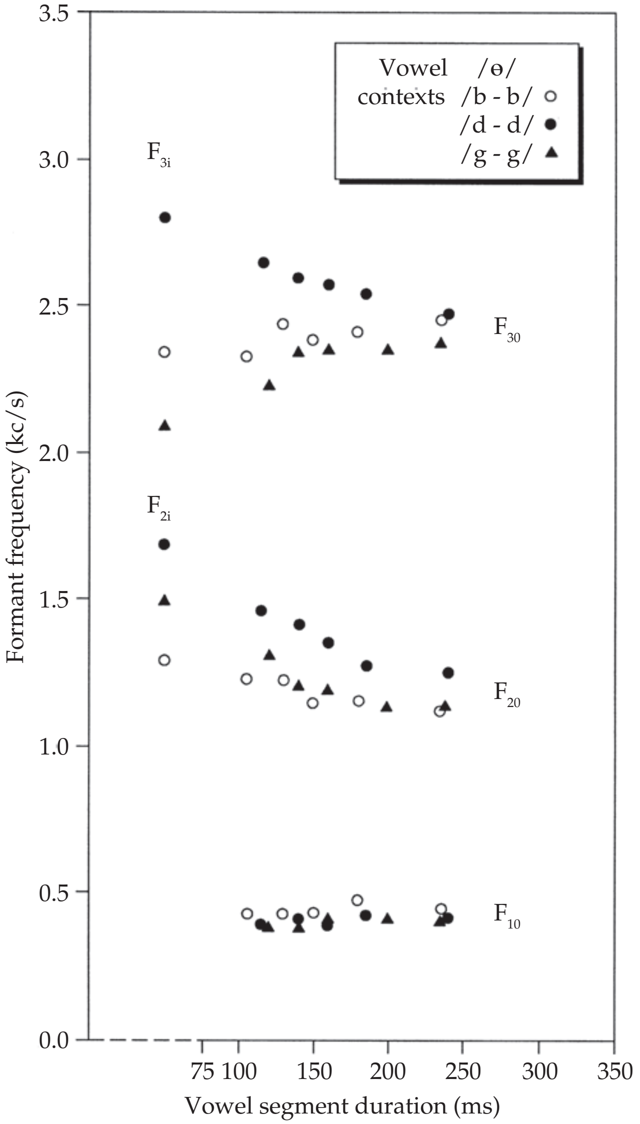


Figure 9.6 Mean F1, F2 and F3 frequencies during the steady state of the Swedish vowel /ɵ/ plotted against vowel duration. The vowel is in the contexts of /b/, /d/ and /g/. As the vowel shortens, the F2 and F3 frequencies shift towards the formant values of the /bV/, /dV/ and /gV/ initial boundaries, F2i and F3i. (Reprinted with permission from B. Lindblom (1963), "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, 35, 1773–81. Copyright 1963, Acoustical Society of America.)

strategies available to speakers: duration, input articulatory force, and time constant of the system. An increase in input force and/or an increase in speed of the system response (i.e., a decrease in stiffness) contribute to increase the movement amplitude/velocity, and hence to decrease the amount of context-dependent undershoot. Thus, there appears to be an undershoot-compensatory reorganization of articulatory gestures in clear speech.

2.3.3 Natural speech as a low-cost strategy The experiment carried out by Lindblom et al. (1975) shows that a low-cost strategy, characterized by coarticulatory variations, is the norm in natural speech. The authors analyzed apical consonants in VCV utterances. Using a numerical model of apical stop production, they showed that the best match between the output of the model and spectrographic data of natural VCV utterances produced in isolation is a tongue configuration always compatible with the apical closure but characterized by a minimal displacement from the preceding vowel. In other words, the experiment shows that among a number of tongue-body shapes that facilitate tongue-tip closure, the tongue body always tends to take those requiring the least amount of movement and an adaptative behavior to the articulatory configuration of the adjacent vowels.

Lindblom's hypothesis, that the more speech style shifts towards the hypospeech pole the larger will be the amount of coarticulation, is confirmed by a number of studies on connected speech: Krull (1987, 1989), for Swedish; Duez (1991) for French; Farnetani (1991) for Italian.

2.3.4 The Locus equation metrics as a measure of coarticulation Locus equations were conceived by Lindblom (1963), who defined them as linear regressions of the onset of the F2 transition on the F2 target, measured at the vowel nucleus. He formulated locus equations as F2 onset = K F2vowel + c, where the constants K and c are the slope and the intercept, respectively. Lindblom showed that the data points were aligned at about the regression line and that the slope and the intercept varied as a function of consonant place of articulation.

Krull (1989) pursued Lindblom's locus equations experiments. Most importantly, she found that the variations of slope and intercept as a function of consonant place in CV syllables were proportional to the extent of coarticulation between the vowel and the preceding consonant, such that flatter slopes indicate a more invariant locus and steeper slopes an increase in coarticulation. In other words, locus equation data are strongly related to the underlying coarticulatory behavior. Krull also found for CVC sequences that prevocalic stop consonants undergo stronger anticipatory vowel effects than postvocalic consonants undergo carryover vowel effects.

In a later study, Chennoukh et al. (1997) reported that locus equations depend on consonant place and on degree of coarticulation. Moreover, in a series of experiments conducted by Sussman and colleagues locus equations have been shown to account for degree of coarticulation in different conditions: coarticulation, obtained by locus equations, decreases in VC versus CV utterances with a stop consonant due to the greater articulatory precision in the production of prevocalic than of postvocalic consonants (Sussman et al., 1997); slopes are identical for single and geminate open syllables and lower for closed syllables (Sussmann & Modarresi, 2003); even though there is less coarticulation for a coda stop in VC and C##V sequences than for a syllable onset stop in CV sequences, the locus equation slopes for the across-syllable and word-boundary conditions still differ as a function of place of articulation (Modarresi et al., 2004).

Lindblom et al. (2002) have proposed a novel interpretation of troughs repeatedly observed in speech. A trough is described as an apparent discontinuity of anticipatory coarticulation that takes the form of a momentary deactivation of tongue or lips movement after the first vowel in a VCV sequence. This event, according to Lindblom et al. (2002), could suggest a segment-by-segment activation pattern, as opposed to a V-to-V trajectory with an independent and superimposed consonant gesture as proposed by Öhman (1967).

2.4 *The window model of coarticulation*

Keating (1985, 1988a, 1988b, 1990) formulated an articulatory model which, in the author's opinion, can account for the continuous changes in space and time observed in speech as well as for interlanguage differences in coarticulation. Keating agrees that phonological rules cannot account for the graded nature of coarticulation, but she contests the assumption that such graded variations are to be ascribed to the speech production mechanism (Keating, 1985). Her proposal is that all graded spatial and temporal contextual variations be accounted for by the phonetic rules of the grammar.

2.4.1 The windows Input to the window model is the phonological representation in terms of binary features. For a given articulatory or acoustic dimension, a feature value is associated with a range of values called a window. Specified features are associated with narrow windows and allow for little contextual variation; unspecified features are associated with wide windows and allow for large contextual variation. Windows are connected by interpolation functions called "paths" or contours. Paths should represent the articulatory or acoustic variations over time in a specific context (see Figure 9.7 showing some selected sequences of windows and contours).

Wide windows specify very little about a segment. On this crucial point, Boyce et al. (1991) argue that, if supposedly unspecified segments are associated in production with characteristic articulatory positions, it becomes hard to reconcile the demonstration of any kind of target with the notion of underspecification. The authors propose instead that phonologically unspecified features can influence speech production in another way: they may be associated with cross-speaker variability (as shown by their lip protrusion data during unspecified consonants) and with cross-dialectal variability.

2.4.2 Cross-language differences According to Keating, interlanguage differences in coarticulation may originate from phonology or from phonetics. Phonological

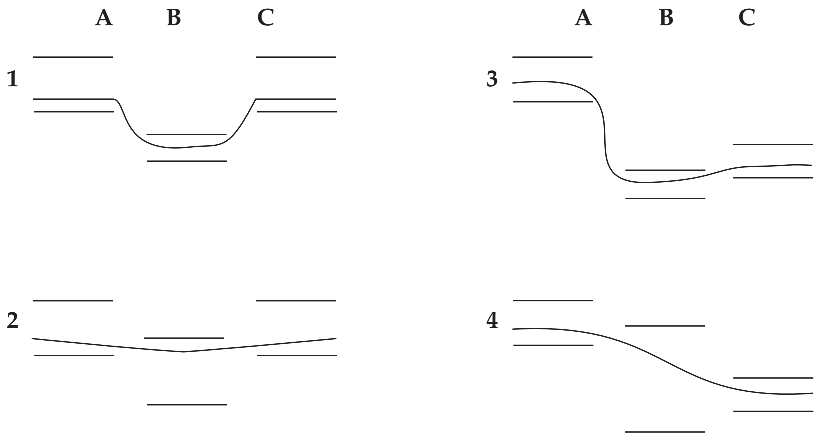


Figure 9.7 Windows and paths modeling articulator movements in three-segment sequences (selected from Keating, 1988a). The effects of narrow vs. wide windows on the interpolation contours can be observed in both the symmetric (1, 2) and the asymmetric (3, 4) sequences.

differences occur when, for a given feature, phonological assimilatory rules operate in one language and not in another. Phonetic differences are due to a different phonetic interpretation of a feature left unspecified. Speech analysis will help determine which differences are phonological and which are phonetic.

In a study of nasalization in English using airflow measurements, Cohn (1993) compared nasal flow contours in nasalized vowels in English with those of nasal vowels in French and of nasalized vowels in Sundanese. Vowel nasality is phonological in French and described as the output of a phonological spreading rule in Sundanese. Cohn found that in the nasalized vowels of Sundanese the flow patterns have plateau-like shapes very similar to the French patterns; in nasalized English vowels, instead, the shapes of the contours describe smooth trajectories from the [–nasal] to the [+nasal] adjacent segments. The categorical versus gradient quality of nasalization in Sundanese versus English indicates that nasalization is the output of phonological assimilatory rules in the former language and results from phonetic interpolation rules in the latter.

Manuel (1987) disagrees with Keating's tenet that all phonetic changes have to be accounted for by grammatical rules simply because they are not universal. Referring to interlanguage differences in V-to-V coarticulation, Manuel proposes that language-particular behavior, apparently arbitrary, can itself be deduced from the interaction between universal characteristics of the motor system and language-specific phonological facts such as the inventory and distribution of vowel phonemes. Her hypothesis is that V-to-V coarticulation is regulated in each language by the requirement that the perceptual contrast among vowels be preserved, i.e., by output constraints, which can be strict in some languages and loose in others. There ought to be more coarticulatory variations in languages with smaller vowel inventories, where there is less possibility of confusion, than in languages with a larger number of vowels, where coarticulation may lead to articulatory/acoustic overlap of adjacent vowel spaces. This hypothesis was tested by comparing languages with different vowel inventories (Manuel & Krakow, 1984; Manuel, 1987, 1990). Results of these studies support the output constraints hypothesis. Thus, if the output constraints of a given language are related to its inventory size and to the distribution of vowels in the articulatory/acoustic space, then no particular language-specific phonetic rules are needed since different degrees of coarticulation across languages can be predictable to some extent.

2.5 *Coarticulation as coproduction*

The coproduction theory has been elaborated through collaborative work of psychologists and linguists, starting from Fowler (1977, 1980, 1985), Fowler et al. (1980) and Bell-Berti and Harris (1981). In conjunction with the new theory, Kelso et al. (1986), Saltzman and Munhall (1989) and Saltzman (1991) have developed a computational model, the task-dynamic model, whose aim is to account for the kinematics of articulators in speech. Input to the model are the phonetic gestures, the dynamically defined units of gestural phonology, proposed as an alternative to features by Browman and Goldstein (1986, 1989, 1990a, 1990b, 1992).

The present account centers on four topics: the nature of phonological units, coarticulation resistance, anticipatory labial and velar coarticulation, and the DAC model of coarticulation.

2.5.1 The dynamic nature of phonological units The central point of Fowler's criticism of feature-based theories (Fowler, 1977, 1980) is the dichotomy between the abstract, discrete, and timeless units posited at the level of language knowledge, and the physical, continuous, and context-dependent articulatory movements at the level of performance. In other words, she contests the assumption that what speakers know about the phonological categories of their language is substantially different from the units they use when they speak. According to Fowler, all current accounts of speech production need a translation process between the abstract and the physical domain: the speech plan supplies the spatial targets to be reached and a central clock specifies when the articulators have to move to the targets ("The articulator movements are excluded from the domain of the plan except as it is implied by the different successive articulatory positions"; Fowler, 1977, p. 99). An alternative proposal that overcomes the dichotomy between linguistic and production units and gets rid of a time program separated from the plan, is to modify the phonological units of the plan. The plan must specify the act to be executed, not only describe "an abstract summary of its significance" (Fowler et al., 1980, p. 381). The production units, the articulatory gestures, must be planned actions serially ordered, specified dynamically and context-free. It is their specification in terms of dynamic parameters (such as force and stiffness) that automatically determines the kinematics of speech movements. Gestures have their own intrinsic temporal structure, which allows them to overlap in time when executed, and

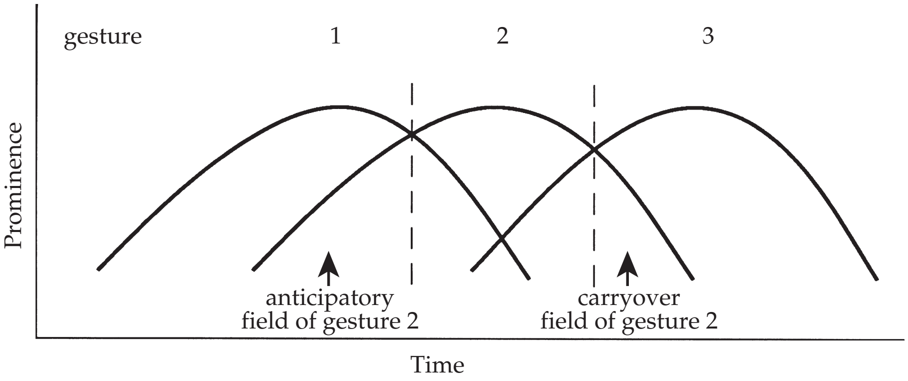


Figure 9.8 Representation of a sequence of three overlapping gestures. (From Fowler & Saltzman, 1993)

the degree of gestural overlap is controlled at the plan level. So gestures are not altered by adjacent gestures but coproduced with them. Figure 9.8 taken from Fowler and Saltzman (1993) illustrates the coproduction of articulatory gestures.

The activation of a gesture increases and decreases smoothly in time, and so does its influence on the vocal tract shape. In the figure, the vertical lines delimit a temporal interval (possibly corresponding to an acoustic segment) during which gesture 2 is maximally prominent, i.e., it has maximal influence on the vocal tract shape, while the overlapping gestures 1 and 3 have a weaker influence. The influence of gesture 2 is clearly less before and after this interval during its initiation and relaxation period, respectively.

The view of gestures as intervals of activation gradually waxing and waning in time, echoes the early insight by Joos (1948) who proposed the "innervation wave theory" to account for coarticulation (section 2.1 above).

2.5.2 Coarticulation resistance Articulatory gestures are implemented by coordinative structures, i.e., by transient functional dependencies among the articulators that contribute to a gesture. These constraints are established to ensure invariance of the phonetic goal. For instance, upper lip, lower lip, and jaw are functionally linked in the production of bilabial closures, so that one will automatically compensate for a decreased contribution of another due to perturbation or coarticulatory variations (see Löfqvist, this volume).

How are coarticulatory variations accounted for within the gestural framework? According to Fowler and Saltzman (1993), variations induced by coproduction depend on the degree to which the gestures share articulators, i.e., on the degree of spatial overlap. When subsequent gestures share only one articulator, such as the jaw in /VbV/ sequences, the effects of gestural interference will be irrelevant, and temporal overlap between vocalic and consonantal gestures will take place with minimal spatial perturbations. The highest degree of spatial perturbation occurs when two overlapping gestures share the articulators directly involved in the production of gestural goals, and impose competing demands on them. Browman and Goldstein (1989) and Saltzman and Munhall (1989) propose that the phasing of gestures may be context-free and that the output of a gestural conflict may be simply a blend of the influence of the overlapping gestures. According to Fowler and Saltzman (1993), the outcome of gestural blending depends on the degree of "blending strength" associated with the overlapping gestures: "stronger" gestures tend to suppress the influence of "weaker" gestures, while the blending of gestures of similar strength will result in an averaging of the two influences. In agreement with experimental findings (Bladon & Nolan, 1977; Recasens, 1984b; Farnetani & Recasens, 1993; Fowler & Brancazio, 2000), Fowler and Saltzman's account of coarticulatory resistance implies that gestures with a high degree of blending strength resist interference from other gestures, and at the same time themselves induce strong coarticulatory effects. On this account, the highest degree of blending strength appears to be associated with consonants requiring extreme constrictions and/or placing strong constraints on articulator movements, while a moderate degree of blending strength appears to be associated with vowels. A compatible proposal may be found in Lindblom (1983), according to which coarticulatory adaptability, maximal in vowels and minimal in lingual fricatives, varies as a function of the phonotactically based sonority categories.

The coproduction account of coordination and coarticulation also implies that speakers do not need to perform a continuous feedforward control of the acoustic output and consequent articulatory adjustments. Likewise, cross-language differences do not result from online control of the output. Languages may differ in degree of coarticulation in relation to their inventories, but these differences are consequences of the different gestural set-up, i.e., the parameters that specify the dynamics of gestures and their overlap, which are learned by speakers of different languages during speech development.

2.5.3 Anticipatory extent of labial and velar coarticulation According to the coproduction theory, articulatory gestures have their own intrinsic duration. Hence, the temporal extent of anticipatory coarticulation must be constant for a given gesture. Compatibly, Bell-Berti and Harris (1979, 1981, 1982), proposed the "frame" or time-locked model of anticipatory coarticulation on the basis of experimental data on lip rounding and velar lowering. The model states that the onset of an articulator movement is independent of the preceding phone string length and occurs at a fixed time before the acoustic onset of the segment with which it is associated.

Findings reported in other studies, however, are more consistent with the look-ahead model (section 2.2.2) and reveal that the onset of anticipatory lip rounding or anticipatory velar lowering is not fixed but extends as a function of the number of neutral segments preceding the influencing segment (see Daniloff & Moll, 1968 and Sussman & Westbury, 1981 for lip rounding coarticulation, and Moll & Daniloff, 1971 for velar coarticulation). Yet, other results on velar coarticulation in Japanese

(Ushijima & Sawashima, 1972; Ushijima & Hirose, 1974) and in French (Benguerel et al., 1977a, 1977b) indicate that velar lowering for a nasal consonant does not start earlier in sequences of three than of two oral vowels preceding the nasal.

An important finding of Benguerel et al. (1977a, 1977b), apparently disregarded in the literature, was the distinction between velar lowering associated with the oral segments preceding the nasal, and a subsequent more rapid velar lowering for the nasal which causes the opening of the velar port. Bladon and Al-Bamerni (1982) reported similar findings on velar coarticulation in CV_nN sequences in English: speakers seem to use two production strategies, either a single velar opening gesture, or a two-stage gesture whose onset is aligned with the first oral vowel and whose higher velocity stage is coordinated with the nasal consonant. Perkell and Cohen (1986) and Perkell (1990) were the first to observe two-stage patterns in lip rounding movements, which converge with Bladon and Al-Bamerni's observations: in /iC_nu/ utterances there was a gradual onset of lip protrusion linked to the offset of /i/, followed by an increase in velocity during the consonants and an additional protrusion motion closely linked with /u/ and quite invariant. The authors interpreted the composite movements of the two-stage patterns as a mixed coarticulation strategy, and proposed a third model of anticipatory coarticulation, the hybrid model. According to this model, the early onset of the protrusion movement would reflect a look-ahead strategy, while the rapid increase in protrusion at a fixed interval before the rounded vowel would reflect a time-locked strategy. Figure 9.9 compares the three models of anticipatory coarticulation. Perkell's data on three English subjects indicated that two of the three subjects used the two-stage pattern and, therefore, were consistent with the hybrid model (Perkell, 1990).

Boyce et al. (1990) argue that many of the conflicting results on the extent of anticipatory coarticulation stem from the assumption that phonologically unspecified segments are also articulatorily neutral (see section 2.2.3): a number of studies have attributed the onset of lip rounding or velar lowering to anticipatory coarticulation without testing first whether or not the phonologically neutral contextual phonetic segments had specific lip or velar target positions. Data on velar lowering for vowels in nasal and oral contexts reported by Bell-Berti and Krakow (1991) show that the early onset of velar lowering in two-stage patterns is associated with the characteristic velar positions of the oral vowels, while the second stage is associated with the production of the nasal consonant; therefore, the two-stage patterns of interest do not reflect a mixture of coarticulation strategies but simply a vocalic gesture followed by a consonantal gesture. Moreover, the study shows that the patterns of velar movement are not random but depend on speech rate and on the number of vowels in the string, e.g., the two-movement pattern prevails in longer versus shorter utterances.

In a subsequent study on four American speakers, Perkell and Matthies (1992) tested whether the onset of the initial phase of lip protrusion in /iC_nu/ utterances is related to consonant-specific protrusion targets as proposed by Boyce et al. (1990), and whether the second phase starting at the maximum acceleration event is indeed stable as predicted by the hybrid and coproduction models, or else is

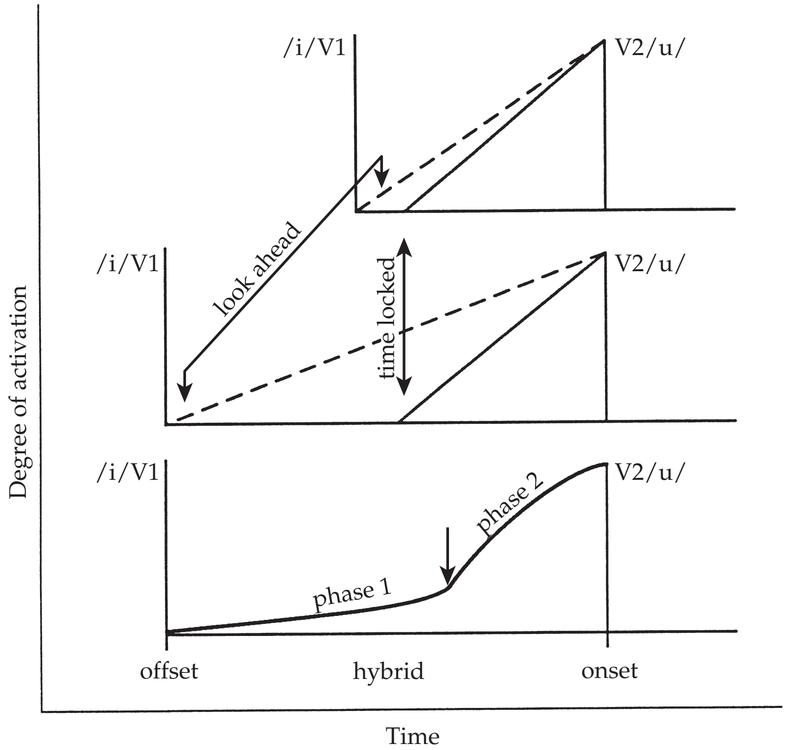


Figure 9.9 Schematic representation of the three models of anticipatory lip-rounding coarticulation proposed by Perkell (see text for description).

itself affected by the number of consonants. In agreement with Boyce et al. (1990), the movement patterns in the control /iC$_n$i/ utterances showed consonant-related protrusion gestures (especially for /s/), while those in the /iC$_n$u/ utterances exhibited earlier labial activity when the first consonant in the string is /s/. Both data confirm that the consonant contributes to the onset of lip movement. The analysis of the second-phase movement, i.e., of the /u/-related component of lip protrusion, revealed that the interval between the acceleration peak and the onset of /u/ tended to vary as a function of consonant duration for three subjects (although the correlations were very low, with R^2 ranging from 0.057 to 0.35). According to the authors, the timing and the kinematics of this gesture reflect the simultaneous expression of competing constraints, that of using the same kinematics (as predicted by the time-locked model), and that of starting the protrusion gesture for /u/ when it is permitted by the relaxation of the retraction gesture for /i/ (as predicted by the look-ahead model). The variability between and within subjects would reflect the degree to which such constraints are implemented. Also, according to data for French (Abry & Lallouache, 1995), the lip-protrusion movement measured from acceleration maximum to protrusion maximum varies in duration as a function of the consonant interval; however, in disagreement with the look-ahead model, its duration does not decrease from the /iC_1y/ utterance type to the utterance /iy/; in other words, lip protrusion can expand in time but cannot be compressed.

The possibility that the slow onset of the lip-protrusion movement occurring around the offset of /i/ may reflect a passive movement due to the relaxation of the /i/ retraction gesture rather than an active look-ahead mechanism has not yet been explored. Sussman and Westbury (1981) did observe for /iC_nu/ sequences that lip protrusion started before the onset of orbicularis oris activity, and suggested that this movement might be the passive result of the cessation of risorius activity and simply reflects a return of the lips to the neutral position.

As it might be expected, cross-language studies indicate that anticipatory coarticulation varies in timing and amplitude across languages. Clumeck (1976) observed that the timing and amplitude of velar lowering varies across the six languages analyzed; Lubker and Gay (1982) showed that anticipatory lip protrusion is much more extensive in Swedish than in English; an investigation on lip rounding in English and Turkish conducted by Boyce (1990) indicated that, while the English patterns were consistent with the coproduction model, the plateau-like protrusion curves of Turkish rendered lip-rounding coarticulation a phonological process.

2.5.4 The DAC model of coarticulation The degree of articulatory constraint or DAC model of coarticulation has been proposed by Recasens, mostly with data for the Catalan language, in order to deal with the complexity of lingual coarticulatory effects in speech (Recasens et al., 1997; Recasens, 2002). It claims that the size, temporal extent, and direction of lingual coarticulation are conditioned by the severity of the requirements imposed on the tongue for the production of vowels and consonants. Vowels and consonants are assigned specific DAC values. Thus, front vowels are more constrained than low and back rounded vowels in line with the biomechanical properties involved in displacing the tongue dorsum upwards and frontwards, and the least constrained vowel is schwa since it is has no clear articulatory target. Differences in degree of constraint are available for consonants as well: the DAC value is highest for consonants requiring much articulatory precision in the performance of frication for lingual fricatives, trilling for the alveolar trill, and tongue predorsum lowering and tongue postdorsum retraction for dark /l/; labials are least constrained since the tongue body does not intervene essentially in their production.

An increase in degree of articulatory constraint causes an increase in coarticulatory resistance and coarticulatory dominance, i.e., in the strength of the coarticulatory effects from and onto the adjacent segments, respectively. Thus, in comparison to the less constrained alveolar /n/, the alveolopalatal /ɲ/ is less coarticulation-sensitive to tongue-dorsum lowering effects from /a/ and exerts more prominent tongue-dorsum raising effects on this vowel. On the other hand, alveolopalatals are subject to depalatalization by more constrained consonants involving tongue-dorsum lowering and retraction (lingual fricatives, the alveolar trill, dark /l/), while palatalization of the latter consonants by alveolopalatals is less prone to occur (Recasens & Pallarès, 2001).

A relevant aspect of the DAC model concerns patterns of coarticulatory direction. It shows that vowels and consonants usually favor a particular coarticulatory direction over another, i.e., anticipation or carryover, based on the displacement and temporal characteristics of the lingual gestures involved. Thus, dark /l/ favors anticipation (the tongue lowers and backs in anticipation of the tongue-tip raising gesture for this consonant), while alveolopalatals such as /ɲ/ may favor the carryover component over the anticipatory component (which is in line with the articulatory configuration for this consonant being /n/-like at closure onset and /j/-like at closure release). When the complex coarticulatory interactions in VCV sequences are taken into consideration, vowel effects turn out to be affected by consonantal effects at the temporal site where the two coarticulation types conflict with each other. Therefore, the prominence of the vowel-dependent anticipatory effects decreases with an increase in the degree of consonant-dependent carryover coarticulation, while the strength of the vowel-dependent carryover effects varies inversely with the salience of the consonant-dependent anticipatory component. Consequently, dark /l/ allows more anticipatory than carryover tongue-dorsum raising effects from /i/ (as shown in the top graph of Figure 9.10), and the alveolopalatal /ɲ/ more carryover than anticipatory tongue-dorsum lowering effects from /a/ (as shown in the bottom graph of the figure). Trends in vowel-dependent coarticulatory direction may also be predicted for VCV sequences with consonants showing less clear patterns of C-to-V coarticulatory direction provided that sufficient attention is paid to their manner requirements and tongue-body configuration characteristics.

Data on the temporal extent of coarticulation have led to proposals by the DAC model about the role of planning and mechanical factors in speech production. In contrast with previous accounts (see section 2.5.3), they show that vowel anticipatory effects in tongue contact and displacement in VCV or longer VCVCV sequences are not planned to start invariably at the same moment in time. Instead, they begin earlier when the immediately preceding consonant and/or the transconsonantal vowel are relatively unconstrained than when they are highly constrained, e.g., when V1 is /a/ and the consonant is labial or alveolar than when V1 is /i/ and the consonant is alveolopalatal. Among highly constrained consonants, as pointed out above, those that exert less C-to-V carryover allow more vowel anticipation than those that exert more prominent C-to-V carryover effects, both during the consonant and the preceding vowel. This is not to say, however, that anticipatory effects do not involve any planning at all: anticipatory effects turn out to be more fixed than carryover effects, though influenced to some extent by the articulatory requirements for the contextual segments. Carryover effects, on the other hand, are more variable because they are conditioned by inertia and by the biomechanical requirements for the contextual segments, and may be particularly long if resulting from articulatory overshoot (e.g., in the sequence /iɲV/ where /i/ and /ɲ/ reinforce each other).

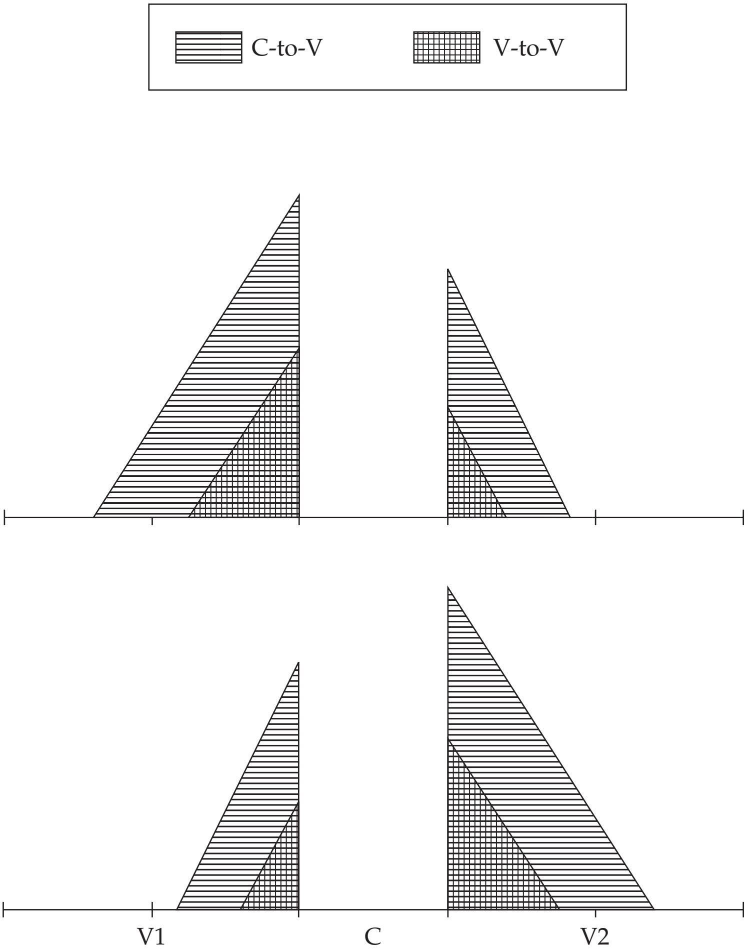


Figure 9.10 Schematic representation of C-to-V and V-to-V effects in VCV sequences with dark /l/ (top) and alveolopalatal /ɲ/ (bottom). (From D. Recasens and M. D. Pallarès (1998) "An electropalatographic and acoustic study of temporal coarticulation for Catalan dark /l/ and German clear /l/," *Phonetica*, 55, 53–79. Reproduced with permission from S. Karger AG, Basel.)

The DAC model has confirmed Fowler and Saltzman's notion that gestures specified for competing demands prevail over or are overriden by other gestures depending on gestural strength (see section 2.5.2). Indeed, differences in DAC value between the two consonants in a consonant cluster may account for changes in place of articulation in CC sequences composed of dentals, alveolars, and alveolopalatals produced with a tongue-front articulator (Recasens, 2006). Three scenarios may take place here: regressive assimilation in unconstrained + constrained consonant sequences, by which C1 adapts completely to the C2 place of articulation during its entire duration; blending in unconstrained + unconstrained consonant sequences through the addition of the closure extent for the two consonants; two separate place of articulation targets for C1 and C2 and possible C1-to-C2 carryover coarticulation effects in constrained + unconstrained sequences. Clear /l/ and dark /l/ do not undergo some or any of these three processes, which reveals that laterality may also contribute significantly to an increase in articulatory constraint. The model may also explain syllable-position-dependent differences in the same consonant cluster structures just referred to (Recasens, 2004): when not subject to the coarticulatory influence of other highly constrained consonants, more unconstrained consonants show less closure or constriction fronting and less dorsopalatal contact syllable-finally than syllable-initially; highly constrained consonants, on the other hand, do not exhibit such syllable-position-dependent articulatory effects.

2.6 Connected speech processes

2.6.1 Articulatory model According to Browman and Goldstein (1990b, 1992), gestural phonology provides an explanatory and unifying account of apparently unrelated speech processes (coarticulation, allophonic variations, alternations) requiring a number of separate phonological rules in featural phonology. Here the phonological structure of an utterance is modeled as a set of overlapping gestures specified on different tiers (see Figure 9.11 for the vocal tract variables). Gradient variations in overlap, or quantitative variations in gestural parameters, can account for a large number of allophonic variations as a function of stress and position, as well as for the alternations observed in connected speech. Connected speech processes such as assimilations, deletions, and reductions or weakenings can be accounted for by an increase in gestural overlap and a decrease in gestural amplitude. In casual rapid speech, subsequent consonantal gestures can so far overlap as to hide each other when they occur on different tiers, or to completely blend their characteristics when they occur on the same tier. Hiding gives rise to perceived deletions and/or assimilations, while blending gives rise to perceived assimilations. For example, the deletion of /t/ in a rapid execution of the utterance “perfect memory” is only apparent; X-ray trajectories reveal the presence of the /t/ gesture, overlapped by the following /m/ gesture. Figure 9.11 shows a schematic gestural representation of part of the utterance “perfect memory” spoken (a) in isolation and (b) within a fluent phrase.

In the figure the extent of each box represents the duration (or activation interval) of a gesture. It can be seen that within each word articulatory gestures always overlap, but in version (b) the labial closure for the initial /m/ of word 2 overlaps and hides the alveolar gesture for the final /t/ of word 1. According to the authors, hidden gestures may be extremely reduced in magnitude or completely deleted. As pointed out by Browman and Goldstein (1990b, p. 366): “Even deletion, however, can be seen as an extreme reduction, and thus as an endpoint in a continuum of gestural reduction, leaving the underlying representation unchanged.”

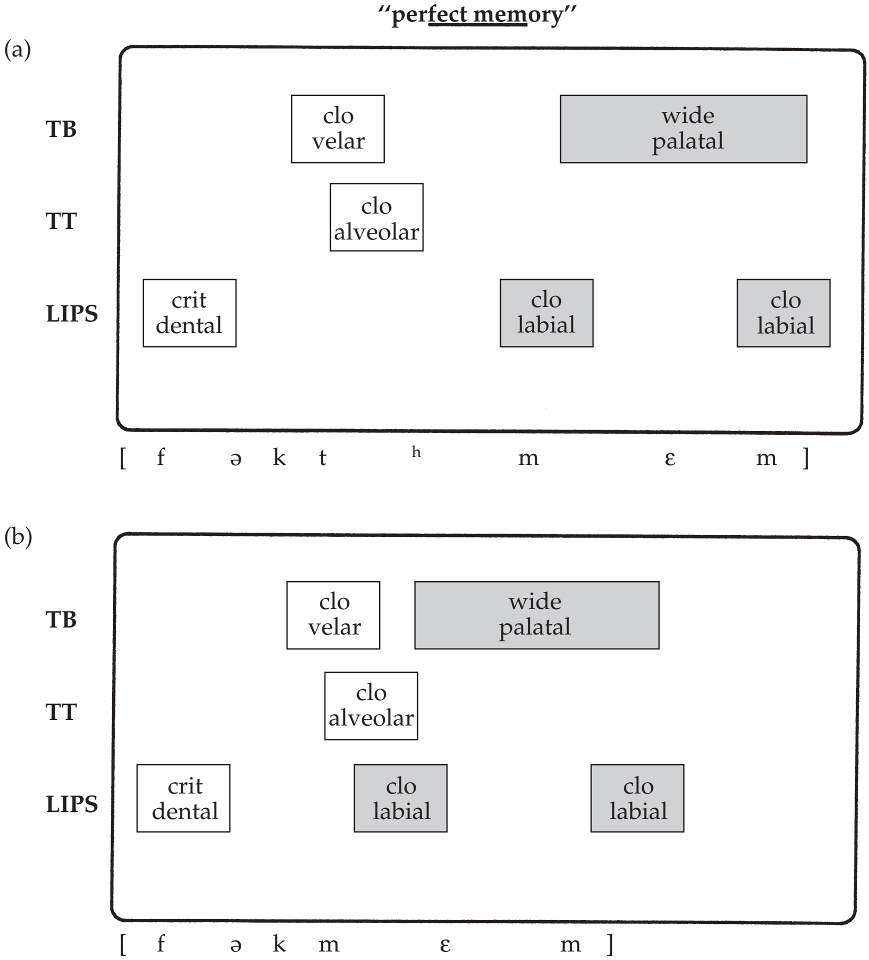


Figure 9.11 Representation of the phonological structure of the utterance "perfect memory" produced in isolation (a) and in continuous speech (b). Vocal tract variables are from top: tongue body, tongue tip, and lips. (From Browman & Goldstein, 1989)

An example of within-tiers blending is the palatalization of /s/ followed by a palatal in the utterance "this shop": the articulatory analysis should show a smooth transition between the first and the second consonant, not the substitution of the first by the second. Finally, in CVCV utterances with unstressed schwa as first vowel, the gestures on the consonantal tier can overlap the schwa gesture so far as to completely hide it, giving the impression of deletion of the unstressed syllable.

2.6.2 Experimental data Consistently with both gestural theory and Lindblom's hyper-/hypo-speech account, an increase in coarticulation and reduction in rapid, fluent speech has been found to hold in a number of acoustic and articulatory studies (see section 2.3.2).

The proposition that position-dependent allophonic variations proceed continuously rather than categorically is supported by experimental data on contrast neutralization. This approach differs from generative phonology which accounts for contrast neutralization through rules that delete the feature(s) responsible for the contrast. An acoustic-perceptual experiment on vowel contrast neutralization in devoiced syllables in Japanese shows that contrast is not completely neutralized (Beckman & Shoji, 1984): listeners are able to recover the underlying vowels /i/ and /u/, possibly from coarticulatory information present in the preceding consonant. In an acoustic-perceptual study on neutralization of the voicing contrast in word-final obstruents in German, Port and O'Dell (1985) found that voicing is not completely neutralized and that listeners are able to distinguish the voiced from the voiceless consonants with better-than-chance accuracy.

Also the majority of English data on alveolar-velar place assimilation in connected speech reported so far is consistent with the proposition that the nature of the segmental adaptive changes is gradient. EPG studies on VC-CV sequences, where C1 is an alveolar stop (Kerswill & Wright, 1989; Wright & Kerswill, 1989; Nolan, 1992), show an intermediate stage between absence of assimilation and complete assimilation, which the authors refer to as residual alveolar gesture. It is also shown that the occurrences of partial and complete assimilations increase from careful/slow speech to normal/fast speech. Most interestingly, the rate of correct identification of C1 decreases, as expected, from unassimilated to assimilated alveolars, but never falls to zero, suggesting that also in the cases of apparently complete assimilation where lingual alveolar contact is absent, listeners can make use of some residual cues to the place distinction. The data are in agreement with the hypothesis that in English the assimilation of alveolars to velars is a continuous process. This is confirmed by recent research on alveolar nasal + velar stop clusters (Hardcastle, 1994).

Other data (Barry, 1991; Nolan et al., 1993) challenge some of the assumptions of gestural phonology. The cross-language study by Barry (1991) on English and Russian alveolar–velar clusters confirms that assimilation in English is a graded process. In Russian, instead, assimilation never occurs when C1 is an oral stop; when C1 is a nasal, assimilation may be continuous or categorical depending on syllabic structure. Data on /s#ʃ/ sequences reported by Nolan et al. (1993) do show intermediate articulations between two-gesture and one-gesture patterns, as predicted by gestural phonology. Accordingly, the one-gesture or static patterns should reflect complete spatio-temporal overlap, i.e. they should show a blending of the /s/ and /ʃ/ influences and a duration comparable to that of a single consonant. Contrary to this hypothesis, preliminary results indicate that the static patterns have the spatial characteristic of a typical /ʃ/, and are 16 percent longer than an initial /ʃ/. Recent EPG and durational data on Italian clusters with C1 = /n/ followed by an oral C1 differing in place and manner of articulation suggest that both categorical and continuous processes may coexist in a language, the occurrence of the ones or the others depending on cluster type and individual speech style. Moreover, the finding that in Italian the alveolar–velar assimilation in /nk/ clusters is always categorical indicates, in agreement with Barry (1991), that the assimilatory process for the same cluster type may differ qualitatively across languages (Farnetani & Busà, 1994). Also, as pointed out by the DAC model (section 2.5.4), different predictions appear to be needed for sequences of consonants produced with close and distant primary tongue articulators; in particular, complete place adaptation appears to be an efficient strategy for the implementation of two consecutive consonants produced with the same or a close articulator and differing in manner of articulation requirements.

3 Summary

This excursus on the problem of contextual variability shows, on one hand, the incredible complexity of the speech production mechanism, which renders the task of understanding its underlying control principles so difficult. It shows, on the other hand, the enormous theoretical and experimental ongoing progress, as reflected in continuously evolving and improving models, and in increasingly rigorous and sophisticated research methodologies.

We started with the questions of the origin, function, and control of coarticulation. At the moment there is no single answer to these questions. For generative phonology, assimilations, connected speech processes, and coarticulation are different steps linking the domain of competence with that of performance, with no bottom-up influences from the physical to the cognitive structure of the language. For both the theory of "adaptive variability" and the theory of gestural phonology the origin of coarticulation lies in speech (in its plasticity and adaptability for the former, in its intrinsic organization in time for the latter). Both theories assume that the nature of speech production itself is at the root of linguistic morpho-phonological rules, which are viewed as adaptations of language to speech processes, sometimes eventuating in historical sound changes. However, there is a discrepancy between the two theories on the primacy of production vs. perception in the control of speech variability. Gestural phonology considers acoustics/perception as the effect of speech production, whilst Lindblom's theory of "adaptive variability" sees acoustics/perception as the final goal of production, hence perception itself shapes production.

Two general control principles for speech variability have been repeatedly advocated: economy (by Lindblom and by Keating) and output constraints (advocated by Lindblom for the preservation of perceptual contrast across styles within a language and extended by Manuel to account for interlanguage differences in coarticulation).

If we confront the various articulatory models with experimental data, it seems that the overall results on coarticulation resistance are more consistent with the gestural model than with other models, although certain patterns of coarticulation resistance could be better explained if aerodynamic/acoustic constraints, in addition to articulatory constraints, were taken into account (Sussman & Westbury, 1981;

Engstrand, 1983). The challenging hypothesis of gestural phonology that connected speech processes are not substantially different from coarticulation processes (i.e., are continuous and do not imply qualitative changes in the categorical underlying units) is supported by a large number of experimental results. However, recent data, based on both spatial and temporal parameters, indicate that assimilation can also be a rule-governed categorical process.

As for anticipatory coarticulation, no model in its present version can account for the diverse results within and across languages: the review shows that articulatory structures and languages differ both quantitatively and qualitatively in the way they implement this process. Lingual coarticulation appears to be subject to a more restricted set of mechanisms than labial and velar coarticulation which calls for models (such as the DAC model) relying on detailed information about the articulatory constraints involved in the production of specific vowel and consonant types. Languages differ considerably in the anticipatory coarticulation strategies for lips and velum. Thus, English and Swedish seem to differ quantitatively in lip-rounding anticipation (Lubker & Gay, 1982), while the plateau-patterns observed in some languages (Boyce, 1990; Cohn, 1993) suggest that the process is phonological in some languages and phonetic in others. Most intriguing in the data on anticipatory coarticulation are the discrepancies among the results for the same language, as those on vowel nasalization in American English (cf. Moll & Daniloff, 1971 vs. Bell-Berti, 1980, vs. Solé & Ohala, 1991). Such discrepancies might be due to different experimental techniques, or the different speech material may itself have conditioned the speaking style or rate and hence the coarticulatory patterns. The discrepancies might also reveal actual regional variants, suggesting ongoing phonetic changes, yet to be fully explored.

NOTE

We thank Michael Studdert-Kennedy for his extensive comments on the form and substance of the present work, and Bjorn Lindblom and Bill Hardcastle for their valuable suggestions. This research was supported in part by projects ESPRIT/ ACCOR, WG 7098 from the European Union, and HUM2006-03743 from the Spanish Ministry of Education.

REFERENCES

Abry, C. & Lallouache, M. T. (1995) Le MEM: un modèle d'anticipation paramétrable par locuteur. Données sur l'arrondissement en français. *Bulletin du Laboratoire de la Communication Parlée*, Centre National de la Recherche Scientifique, Grenoble, 3, 85–99.

Barry, M. (1991) Temporal modelling of gestures in articulatory assimilation. *Proceedings of the 12th International Congress of Phonetic Sciences* (Aix-en-Provence), 4, 14–17.

Beckman, M. & Shoji, A. (1984) Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. *Phonetica*, 41, 61–71.

Bell-Berti, F. (1980) Velopharyngeal function: A spatial-temporal model. In N. J. Lass (ed.), *Speech and Language: Advances in Basic Research and Practice* (pp. 291–316). New York: Academic Press.

Bell-Berti, F. & Harris, K. (1974) More on the motor organization of speech gestures. *Haskins Laboratories Status Report on Speech Research*, 37/38, 73–7.

Bell-Berti, F. & Harris, K. (1979) Anticipatory coarticulations: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America*, 65, 1268–70.

Bell-Berti, F. & Harris, K. (1981) A temporal model of speech production. *Phonetica*, 38, 9–20.

Bell-Berti, F. & Harris, K. (1982) Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America*, 71, 449–54.

Bell-Berti, F. & Krakow, R. (1991) Anticipatory velar lowering: A coproduction account. *Journal of the Acoustical Society of America*, 90, 112–23.

Benguerel, A. P. & Cowan, H. (1974) Coarticulation of upper lip protrusion in French. *Phonetica*, 30, 41–55.

Benguerel, A. P., Hirose, H., Sawashima, M., & Ushijima, T. (1977a) Velar coarticulation in French: A fiberscopic study. *Journal of Phonetics*, 5, 149–58.

Benguerel, A. P., Hirose, H., Sawashima, M., & Ushijima, T. (1977b) Velar coarticulation in French: An electromyographic study. *Journal of Phonetics*, 5, 159–67.

Bladon, A. & Al-Bamerni, A. (1976) Coarticulation resistance in English /l/. *Journal of Phonetics*, 4, 137–50.

Bladon, A. & Al-Bamerni, A. (1982) One-stage and two-stage temporal patterns of velar coarticulation. *Journal of the Acoustical Society of America*, 72, S104.

Bladon, A. & Nolan, F. (1977) A video-fluorographic investigation of tip and blade alveolars in English. *Journal of Phonetics*, 5, 185–93.

Boyce, S. (1990) Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America*, 88, 2584–95.

Boyce, S., Krakow, R., & Bell-Berti, F. (1991) Phonological underspecification and speech motor organization. *Phonology*, 8, 219–36.

Boyce, S., Krakow, R., Bell-Berti, F., & Gelfer, C. (1990) Converging sources of evidence for dissecting articulatory movements into core gestures. *Journal of Phonetics*, 18, 173–88.

Browman, C. P. & Goldstein, L. M. (1986) Towards an articulatory phonology. In C. Ewan & J. Anderson (eds.), *Phonology Yearbook 3* (pp. 219–52). Cambridge: Cambridge University Press.

Browman, C. P. & Goldstein, L. M. (1989) Articulatory gestures as phonological units. *Phonology*, 6, 201–51.

Browman, C. P. & Goldstein, L. M. (1990a) Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18, 299–320.

Browman, C. P. & Goldstein, L. M. (1990b) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 341–76). Cambridge: Cambridge University Press.

Browman, C. P. & Goldstein, L. M. (1992) Articulatory phonology: An overview. *Phonetica*, 49, 155–80.

Butcher, A. & Weiher, E. (1976) An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics*, 4, 59–74.

Chennoukh, S., Carré, R., & Lindblom, B. (1997) Locus equation in the light of articulatory modelling. *Journal of the Acoustical Society of America*, 102, 2380–8.

Choi, J. D. & Keating, P. (1991) Vowel-to-vowel coarticulation in three Slavic languages. *UCLA Working Papers in Phonetics*, 78, 78–86.

Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York: Harper and Row.

Clumeck, H. (1976) Patterns of soft palate movements in six languages. *Journal of Phonetics*, 4, 337–51.

Cohn, A. C. (1993) Nasalisation in English: Phonology or phonetics. *Phonology*, 10, 43–81.

Daniloff, R. & Hammarberg, R. (1973) On defining coarticulation. *Journal of Phonetics*, 1, 239–48.

Daniloff, R. & Moll, K. (1968) Coarticulation of lip rounding. *Journal of Speech and Hearing Research*, 11, 707–21.

Duez, D. (1991) Some evidence of second formant locus-nucleus patterns in spontaneous speech in French. *PERILUS XII*, University of Stockholm, 109–26.

Engstrand, O. (1981) Acoustic constraints or invariant input representation? An experimental study of selected articulatory movements and targets. *Reports from Uppsala University Department of Linguistics*, 7, 67–95.

Engstrand, O. (1983) Articulatory coordination in selected VCV utterances: A means-end view. *Reports from Uppsala University Department of Linguistics*, 10, 1–145.

Engstrand, O. (1989) Towards an electropalatographic specification of consonant articulation in Swedish. *PERILUS X*, University of Stockholm, 115–56.

Farnetani, E. (1986) A pilot study of the articulation of /n/ in Italian using electro-palatography and airflow measurements. *15e Journées d'Études sur la Parole*. GALF 23–6.

Farnetani, E. (1990) V-C-V lingual coarticulation and its spatiotemporal domain. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 93–130). Dordrecht: Kluwer.

Farnetani, E. (1991) Coarticulation and reduction in consonants: Comparing isolated words and continuous speech. *PERILUS XIV*, University of Stockholm, 11–15.

Farnetani, E. & Busà, M. G. (1994) Italian clusters in continuous speech. *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, 1, 359–62.

Farnetani, E. & Recasens, D. (1993) Anticipatory consonant-to-vowel coarticulation in the production of VCV sequences in Italian. *Language and Speech*, 36, 279–302.

Farnetani, E., Vagges, K., & Magno-Caldognetto, E. (1985) Coarticulation in Italian /VtV/ sequences: A palatographic study. *Phonetica*, 42, 78–99.

Fowler, C. A. (1977) *Timing Control in Speech Production.* Bloomington: Indiana University Linguistics Club.

Fowler, C. A. (1980) Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8, 113–33.

Fowler, C. A. (1983) Realism and unrealism: A reply. *Journal of Phonetics*, 11, 303–22.

Fowler, C. A. (1985) Current perspectives on language and speech production: A critical overview. In R. G. Daniloff (ed.), *Speech Science* (pp. 193–278). London: Taylor and Francis.

Fowler, C. A. & Brancazio, L. (2000) Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation. *Language and Speech*, 43, 1–41.

Fowler, C. A., Rubin, P., Remez, R., & Turvey, M. (1980) Implications for speech production of a general theory of action. In B. Butterworth (ed.), *Language Production I: Speech and Talk* (pp. 373–420). London: Academic Press.

Fowler, C. A. & Saltzman, E. (1993) Coordination and coarticulation in speech production. *Language and Speech*, 36, 171–95.

Gay, T. (1978) Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America,* 63, 223–30.

Gimson, A. C. (1970) *An Introduction to the Pronunciation of English*, 2nd edn. London: Edward Arnold.

Hammarberg, R. (1976) The metaphysics of coarticulation. *Journal of Phonetics*, 4, 353–63.

Hardcastle, W. (1985) Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication*, 4, 247–63.

Hardcastle, W. (1994) EPG and acoustic study of some connected speech processes. *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, 2, 515–18.

Henke, W. L. (1966) Dynamic articulatory model of speech production using computer simulation. Doctoral Dissertation, MIT.

Hirose, H. & Gay, T. (1972) The activity of the intrinsic laryngeal muscles in voicing control: An electro-myographic study. *Phonetica*, 25, 140–64.

Jones, D. (1969) *An Outline of English Phonetics*, 9th edn. Cambridge: Cambridge University Press.

Joos, M. (1948) Acoustic phonetics. *Language Monographs*, 23 (Supplement to *Language*, 24).

Keating, P. A. (1985) Universal phonetics and the organization of grammars. In V. Fromkin (ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 115–32). Orlando: Academic Press.

Keating, P. A. (1988a) Underspecification in phonetics. *Phonology*, 5, 275–92.

Keating, P. A. (1988b) The window model of coarticulation: Articulatory evidence. *UCLA Working Papers in Phonetics*, 69, 3–29.

Keating, P. A. (1990) Phonetic representations in a generative grammar. *Journal of Phonetics*, 18, 321–34.

Kelso, J. A. S., Saltzman, E., & Tuller, B. (1986) The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, 14, 29–59.

Kerswill, P. & Wright, S. (1989) On the limits of auditory transcription: A sociophonetic approach. *York Papers in Linguistics*, 14, 35–59.

Kohler, K. J. (1990) Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 69–92). Dordrecht: Kluwer.

Kozhevnikov, V. A. & Chistovich, L. A. (1965) *Speech: Articulation and Perception* (trans. US Department of Commerce, Clearing House for Federal Scientific and Technical Information), No. 30, 543. Washington, DC: Joint Publications Research Service.

Krull, D. (1987) Second formant locus patterns as a measure of consonant–vowel coarticulation. *PERILUS V*, University of Stockholm, 43–61.

Krull, D. (1989) Second formant locus pattern and consonant-vowel coarticulation in spontaneous speech. Phonetic experimental research at the Institute of Linguistics, *PERILUS X*, University of Stockholm, 87–108.

Kuehn, D. P. & Moll, K. L. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4, 303–20.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839–62.

Lindblom, B., (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773–81.

Lindblom, B. (1983) Economy of speech gestures. In P. F. MacNeilage (ed.), *The Production of Speech* (pp. 217–45). New York: Springer Verlag.

Lindblom, B. (1989) Phonetic invariance and the adaptive nature of speech. In B. A. G. Elsendoorn & H. Bouma (eds.), *Working Models of Human Perception* (pp. 139–73). London: Academic Press.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H & H theory.

In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 403–40). Dordrecht: Kluwer.

Lindblom, B., Pauli, S., & Sundberg, J. (1975) Modeling coarticulation in apical stops. In G. Fant (ed.), *Speech Communication*, vol. 2 (pp. 87–94). Uppsala: Almqvist and Wiksell.

Lindblom, B., Sussman, H., Modarresi, G., & Burlingame, E. (2002) The trough effect: Implications for speech motor programming. *Phonetica*, 59, 245–62.

Lubker, J. F. & Gay, T. (1982) Anticipatory labial coarticulation: Experimental, biological and linguistic variables. *Journal of the Acoustical Society of America*, 71, 437–48.

Magen, H. S. (1989) An acoustic study of vowel-to-vowel coarticulation in English. Doctoral dissertation, Yale University, New Haven, CT.

Manuel, S. (1987) Acoustic and perceptual consequences of vowel-to-vowel coarticulation in three Bantu languages. Doctoral dissertation, Yale University, New Haven, CT.

Manuel, S. (1990) The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America*, 88, 1286–98.

Manuel, S. & Krakow, R. (1984) Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Report on Speech Research*, 77/78, 69–78.

Menzerath, P. & Lacerda, A. de (1933) *Koartikulation Steuerung und Lautabgrenzung*. Bonn: Ferdinand Dummlers Verlag.

Moll, K. & Daniloff, R. (1971) Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America*, 50, 678–84.

Modarresi, G., Sussman, H., Lindblom, B., & Burlingame, E. (2004) Stop place coding: An acoustic study of CV, VC#, and C#V sequences. *Phonetica*, 61, 2–21.

Moon, S. J. & Lindblom, B. (1994) Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96, 40–55.

Nolan, F. (1992) The descriptive role of segments: Evidence from assimilation. In G. J. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 261–80). Cambridge: Cambridge University Press.

Nolan, F., Holst, T., & Kühnert, B. (1993) Modelling [s] to [ʃ] assimilation. *Journal of Phonetics*, 24, 113–38.

Nord, L. (1986) Acoustic studies of vowel reduction in Swedish. *Quarterly Progress and Status Report*, Department of Speech Communication, KTH, Stockholm, 4, 19–36.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–68.

Öhman, S. E. G. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41, 310–20.

Perkell, J. S. (1990) Testing theories of speech production: Implications of some detailed analyses of variable articulatory data. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 263–88). Dordrecht: Kluwer.

Perkell, J. S. & Cohen, C. (1986) Preliminary support for a "hybrid model" of anticipatory coarticulation. *Proceedings of the 12th International Congress of Acoustics*, Toronto, A3-6.

Perkell, J. S. & Matthies, M. L. (1992) Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability. *Journal of the Acoustical Society of America*, 91, 2911–25.

Port, R. & O'Dell, M. (1985) Neutralization of syllable-final voicing in German. *Journal of Phonetics*, 13, 455–71.

Recasens, D. (1984a) V-to-C coarticulation in Catalan VCV sequences: an articulatory and acoustical study. *Journal of Phonetics*, 12, 61–73.

Recasens, D. (1984b) Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America*, 76, 1624–35.

Recasens, D. (1984c) Timing and coarticulation: Alveolo-palatals and sequences of alveolar + [j] in Catalan. *Phonetica*, 41, 125–39.

Recasens, D. (1987) An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish V-C-V sequences. *Journal of Phonetics*, 15, 299–312.

Recasens, D. (1989) Long range coarticulation effects for tongue dorsum contact in VCVCV sequences. *Speech Communication*, 8, 293–307.

Recasens, D. (2002) An EMA study of VCV coarticulatory direction. *Journal of the Acoustical Society of America*, 111, 2828–41.

Recasens, D. (2004) The effect of syllable position in consonant reduction (evidence from Catalan consonant clusters). *Journal of Phonetics*, 32, 435–53.

Recasens, D. (2006) Integrating coarticulation, blending and assimilation into a model of articulatory constraints. In L. Goldstein, D. Whalen, & C. Best (eds.), *Laboratory Phonology 8* (pp. 611–34). Berlin/New York: Mouton de Gruyter.

Recasens, D. & Pallarès, M. D. (1998) An electropalatographic and acoustic study of temporal coarticulation for Catalan dark /l/ and German clear /l/. *Phonetica*, 55, 53–79.

Recasens, D. & Pallarès, M. D. (2001) Coarticulation, blending and assimilation in Catalan consonant clusters. *Journal of Phonetics*, 29, 273–301.

Recasens, D., Pallarès, M. D., & Fontdevila, J. (1997) A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America*, 102, 544–61.

Saltzman, E. (1991) The task dynamic model in speech production. In H. F. M. Peters, W. Hulstijin, & C. W. Starkweather (eds.), *Speech Motor Control and Stuttering* (pp. 37–52). Amsterdam: Excerpta Medica.

Saltzman, E. & Munhall, K. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1, 333–82.

Solé, M. J. & Ohala, J. (1991) Differentiating between phonetic and phonological processes: The case of nasalization. *Proceedings of the 12th International Congress of Phonetic Sciences* (Aix-en-Provence), 2, 110–13.

Sussman, H., Bessel, N., Dalston, E., & Majors, T. (1997) An investigation of stop place articulation as a function of syllable position: A locus equation perspective. *Journal of the Acoustical Society of America*, 101, 2826–38.

Sussman, H. & Modarresi, G. (2003) The stability of locus equation encoding of stop place. *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcelona), 2, 1931–34.

Sussman, H. M. & Westbury, J. R. (1981) The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research*, 24, 16–24.

Sweet, H. (1877) *Handbook of Phonetics*. Oxford: Clarendon.

Ushijima, T. & Hirose, H. (1974) Electromyographic study of the velum during speech. *Journal of Phonetics*, 2, 315–26.

Ushijima, T. & Sawashima, M. (1972) Fiberscopic observation of velar movements during speech. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, University of Tokyo, 6, 25–38.

Wood, S. (1993) Crosslinguistic cineradiographic studies of the temporal coordination of speech gestures. *Working Papers*, Lund University, Department of Linguistics, 40, 251–63.

Wright, S. & Kerswill, P. (1989) Electropalatography in the analysis of connected speech processes. *Clinical Linguistics and Phonetics*, 3, 49–57.

10 Theories and Models of Speech Production

ANDERS LÖFQVIST

"The purpose of models is not to fit the data but to sharpen the questions."
Samuel Karlin (11th R. A. Fisher Memorial Lecture,
Royal Society, 20 April 1983)

1 The Speech Signal and Its Description

For the purpose of the following presentation, it is convenient to view speech as audible gestures. A speaker creates variations in air pressure and air flow in the vocal tract by making valving actions with different parts of the vocal tract: the glottis, the velum, the tongue, the lips, and the jaw. The changes in pressure and flow give rise to the acoustic signal that we hear when perceiving speech. Most of the variations in the acoustic signal are made intentionally by the speaker to convey linguistic information. Other properties convey what is called paralinguistic information, such as attitudes and emotions, social and geographical dialect characteristics. In addition, there are properties reflecting biological characteristics of the speaker such as sex and age. The resulting acoustic signal is thus shaped by contributions from many different sources that are all overlaid on each other. The fact that listeners can usually identify these different sources suggests that they are recoverable from the acoustic signal.

In describing speech and language, it is common to use one of two modes that can be referred to as the linguistic and the dynamic mode (see Pattee, 1977, for a further elaboration of this distinction). In the linguistic mode, the units of language are described without a temporal domain. For example, most phonological descriptions use a set of symbols that can be arranged in different ways to produce different messages. Although the primitives used for this type of analysis vary depending on the theoretical framework being adopted, the units are commonly described as being discrete and serially ordered. The dynamic mode is used for describing articulatory and acoustic properties of speech. Here, the focus is on the time-varying properties of articulatory movements and/or the spectral characteristics of the speech signal. This necessarily implies a temporal domain. The linguistic units of speech can no longer be described as discrete, since a salient feature of speech production is that the units show considerable articulatory influence and overlap. This is commonly referred to as coarticulation, coproduction, blending, or aggregation (cf. Farnetani & Recasens, this volume). Thus, the movements associated with different production units blend seamlessly with each other and in the articulatory record there are no boundaries between units. Consequently, the movements necessary for the production of a given unit differ according to its context, and likewise its acoustic properties vary according to context. A further result of this overlap is that at any one point in time, the vocal tract is an aggregate of different production units (cf. Fowler & Smith, 1986; Saltzman & Munhall, 1989; Löfqvist, 1990). The obvious acoustic consequence is that a single temporal slice of the signal contains influences from several production units (see Fant, 1962, for an early discussion).

Throughout the history of the study of speech, much effort has been devoted to arguments about these two modes of description (cf. Ohala, this volume). One famous depiction of their different natures is provided by Hockett (1955), who makes an analogy between speech production and a row of raw Easter eggs on a conveyor belt, being smashed between the two rollers of a wringer. The implication is that the units of speech are distinct and serially ordered (perhaps also invariant and displaying their essential properties) before they are all smeared together in the process of articulation: "The flow of eggs before the wringer represents the impulses from the phoneme source; the *mess* that emerges from the wringer represents the output from the speech transmitter" (Hockett, 1955, p. 210; italics added.) We should note that Hockett does not imply that it is impossible to recover the original eggs that went into the mess. He duly comments that an inspector examining the passing mess could "decide, on the basis of the broken and unbroken yolks, the variously spread-out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer" (p. 210) and further notes that the inspector represents the hearer.

While Hockett's Easter egg analogy would seem to represent an extreme case, it is not unique. Rather, it represents a class of theories which have been called translation theories, because they view speech production as translating a mental representation into something completely different during the process of articulation (cf. Fowler et al., 1980; Fowler, 1993). Hockett's view is also understandable from its epistemological context. The discovery of coarticulation was made around the beginning of last century, and Menzerath and de Lacerda (1933) published the first systematic treatise on the subject. Not only did they show large contextual variability for productions of the same sound, but they also showed, as had others before them (see Hardcastle, 1981; Kühnert & Nolan, 1999; Farnetani & Recasens, this volume, for historical reviews), that it was impossible to draw boundaries between sounds in the articulatory record. These findings caused some consternation among speech scientists, since it had often been assumed that the same sound would be articulated in the same way irrespective of its context – an assumption that seems to reappear at certain intervals over time. Hence, the search was on to find the invariant, or essential, properties of the phoneme in production (or acoustics). In a review of the production efforts, Peter MacNeilage neatly sums up the shift in emphasis that has come to dominate work on speech motor control:

> it becomes clear that the more basic problem in speech production theory is not the one considered central to most theorists; namely, why articulators do not always reach the same position for a given phoneme. It is, How do articulators always come as close to reaching the same position as they do? One of the main conclusions of this paper is that the essence of the speech production process is not an inefficient response to invariant central signals, but an elegantly controlled variability of response to the demand for a relatively constant end. (MacNeilage, 1970, p. 184)

Before continuing, we should also note another shift of emphasis in the study of speech motor control. Much work in speech physiology was carried out within a paradigm in which two general issues dominated: chain versus comb models for the serial ordering of articulatory movements, and the role of peripheral feedback in speech production (Lashley, 1951; Kozhevnikov & Chistovich, 1965; Keele, 1968; see Kent, 1976, for a review of these issues). Briefly, in a chain model, the central motor commands to the articulators to produce a segment were supposed to be triggered by feedback from the periphery upon the completion of the articulatory movements for the previous one. In a comb model, the commands to the articulators for successive segments were assumed to be sent according to a plan or temporal scheme. In practice, one limitation in this approach was a tendency to subsume the question of feedback under the question of serial order, and phrase the alternatives as either a chain model incorporating feedback or a comb model without feedback. Of the two remaining alternatives, one was perhaps automatically ruled out, i.e., a chain model without feedback, but the possibility of a comb model incorporating feedback was not generally explored, in spite of the wealth of physiological studies of sensorimotor mechanisms (e.g. Granit, 1970; Matthews, 1972). In such a model, the role of feedback would not necessarily be limited to the sequencing of movements but rather would be important in the shaping of movements as well (as will be discussed in section 2.7 in terms of internal models). A further limitation was an insistence that signals from peripheral receptors go to higher centers with the resulting problem of apparently inadequate loop time. Another possibility could be that information from the periphery goes to lower levels of the nervous system such as the spinal cord or the brainstem; we will later explore the idea that these levels may play a crucial role in integrating signals from the periphery with signals from higher centers.

While coarticulation has been taken as a fundamental characteristic of speech and the basis for the rapidity with which information can be conveyed (Liberman et al., 1967), it is also likely a fundamental characteristic of most motor activities. Presumably, the reason why it has received so much attention in speech science is that speech can, at one level, be described as a succession of discrete segments, making it possible to study how the "canonical" forms of these segments are altered in the process of articulation. Very similar patterns of contextual variability can be found in typing. In typing, the goal is to produce a sequence of keystrokes, and it may thus be easier to define the targets in typing than in speech production. The movements of the fingers towards the keys show large contextual variability in both space and time (Salthouse, 1986). For example, successive keystrokes are made faster by fingers on alternate hands than by fingers on the same hand. The likely reason is that when alternate hands are used, there is no conflict between the fingers used for the strokes, since the two hands can operate independently. The time needed for a keystroke depends on the context in which a character occurs. The range of these contextual influences in typing appears to be limited to two or three characters. In contrast, coarticulatory influences in speech have been claimed to span up to six segments, but the size of the temporal window for coarticulatory influences remains under debate (cf. Farnetani & Recasens, this volume). The differences between strokes made by the same or alternating hands in typing are similar to coarticulation in speech, where the different parts of the vocal tract can operate relatively independently of each other.

2 Concepts and Issues in Movement Control

During speech, parts of the vocal tract are briefly coupled in a functional manner to produce the acoustic characteristics of speech sounds. For example, the production of the bilabial voiceless stop /p/ requires the following set of actions. The lips are closed by joint activity of the jaw and the lips. The velum is elevated to seal off the entrance into the nasal cavity. The glottis is widened and the longitudinal tension of the vocal folds is often increased to prevent glottal vibrations. These articulatory actions all contribute to a period of silence in the acoustic signal and an increase in oral air pressure associated with the stop consonant. Speech production thus involves control and coordination of different parts of the vocal tract. How this is achieved is not well understood. Speech motor control should properly be seen as an instance of the control of coordinated movements in general. As a preliminary to this discussion, we shall briefly review a line of experiments on speech production that will provide a suitable empirical and experimental background. After this review, the remaining parts of this section discuss a number of issues in the control of movement and their implications for speech motor control.

2.1 What happens when speech movements are perturbed?

Daily activities such as walking and picking up and moving objects often require rapid actions to cope with unexpected events such as stumbling or hitting an object with the hand. One valuable experimental paradigm for understanding movement coordination and control is to introduce unexpected perturbations to motor acts in a systematic manner. In a standard experiment, a subject is attached to a small motor that can be activated during some trials to generate a brief load. The rationale for this research is that the nature and time course of the response to the load may reveal the motor organization and reflex structure of the motor act. This paradigm has been applied to different types of motor behavior in humans such as posture control (e.g., Nashner & McCollum, 1985), hand and finger movements (e.g., Traub et al., 1980; Rothwell et al., 1982; Cole et al., 1984), and respiratory control (Newsom Davis & Sears, 1970). A number of studies have also used this method to study speech motor control (Folkins & Abbs, 1975; Folkins & Zimmermann, 1982; Abbs & Gracco, 1984; Kelso et al., 1984; Gracco & Abbs, 1985, 1988, 1989; Flege et al., 1988; Shaiman & Abbs, 1987; Shaiman, 1989; Munhall et al., 1994; Savariaux et al., 1995; Saltzman et al., 1998; Gomi et al., 2002; Honda et al., 2002; Shaiman & Gracco, 2002).

From these speech perturbation studies, some general conclusions can be drawn. First, compensations are rapid. Electromyographic responses can occur 20–30 ms after load onset. The latency is not fixed, however, but depends on when the load was applied with respect to onset of activity in the muscles responsible for the movement in question (Abbs et al., 1984). The short latencies suggest that the responses are not due to reaction time processes. Second, compensations are mostly task-specific. That is, they are neither stereotypic nor evident throughout the system, but rather tailored to the needs of the ongoing motor act. For example, when the jaw is loaded during the transition from a vowel to a bilabial stop, compensatory responses are made in the upper and lower lips to achieve the labial closure. On the other hand, when the jaw is loaded during the transition from a vowel to a dental fricative or a dental stop, a response is seen in the tongue (Kelso et al., 1984; Shaiman, 1989). Similarly, a load applied to the upper lip only elicited lower-lip responses for a bilabial stop /p/, but not for a labiodental fricative /f/, where the upper lip does not contribute to the constriction (Shaiman & Gracco, 2002). We should add a word of caution here, however, since task specificity is not always consistent across speakers. In particular, one of the subjects in the study by Shaiman (1989) showed increased lower-lip movement in addition to jaw and tongue compensatory movements when the jaw was perturbed during the utterance /ædæ/, which does not require lip activity. Similarly, the study by Kelso et al. (1984) found increased upper-lip EMG activity in perturbed productions of /bæz/. Third, compensations are flexible and distributed among articulators involved in a specific task. Thus, when the jaw is loaded in the production of a bilabial stop, responses can occur in the jaw itself and/or in the upper and lower lips (Shaiman, 1989). Fourth, compensations are functional and effective in the sense that the intended goal is normally achieved. For example, Munhall et al. (1994) perturbed the lower lip at the transition from the first vowel to the medial bilabial voiceless stop in the utterance /iˈpip/. The system was able to overcome the load, making the intended closure of the vocal tract and increasing the air pressure in the oral cavity: recordings of oral pressure revealed no differences in pressure between load and control productions.

While the results of these studies clearly indicate that the articulatory system is capable of rapid and functional responses to external loads, such loads may, nevertheless, affect the timing between different articulatory systems (Löfqvist & Gracco, 1991; Saltzman et al., 1998). For example, Munhall et al. (1994) also examined laryngeal responses to lower-lip perturbations during the production of a voiceless bilabial stop. In addition to lip and jaw actions to achieve the labial closure, a laryngeal response was evident by a delay of the onset of glottal abduction, measured relative to the onset of the preceding vowel. This delay was presumably made to maintain lip–larynx coordination at the onset of labial closure, and resulted in an increased acoustic duration of the preceding vowel. However, the period of bilabial closure for the stop was shortened by the perturbation while the laryngeal abduction–adduction movement increased in duration. The normal phasing between the oral and laryngeal movements was consequently disrupted at the release of the oral closure. As a result, Voice Onset Time increased in the perturbed trials since it depends in part on the timing between the oral and laryngeal events in stop production (e.g., Löfqvist & Yoshioka, 1984; Löfqvist, 1992).

2.2 Planning and execution of movements

While it is convenient to discuss movement control in terms of a plan and its execution, there is reason to believe that a clear separation between plan and execution is often not possible. One problem here concerns the representations used in speech planning. Current phonological representations would seem to require a great deal of detail to be filled in during the conversion into a phonetic representation, in particular temporal information. Another issue is how much motoric detail a plan can contain, an issue that will be taken up in more detail in section 2.3. Theories of speech planning have often used cases of speech errors, slips of the tongue or spoonerisms, as evidence. An example of such an error is when someone says "queer old dean" instead of the intended "dear old queen." Based on analysis of such speech errors, several models of the speech planning process have been proposed (e.g., Garrett, 1980; Levelt, 1989; Dell et al., 1993). These findings obviously suggest that utterances are planned, since it would otherwise be difficult to explain how an upcoming word could be exchanged with one that is preceding it. Both words would have to be activated at the same time for such an exchange to occur. Still, the nature of this plan is not clear. Using Hockett's Easter egg analogy, the plan in these models of speech production would seem to correspond to the organization of the eggs before they are smashed between the rollers. That is, the smashing process does not appear to be part of the plan.

2.3 Distributed control

The nervous system is made up of a complex network of interacting neurons and centers at different levels of the system. In motor control, one important function must involve integrating signals from higher centers with signals from the periphery which indicate the current situation. This involves selecting the appropriate muscles, activating them to a suitable degree, and establishing a proper sequence of activation. The integrative function for limb control is located in the spinal cord (cf. Humphrey & Freund, 1991; McCrea, 1992). Only at this level is all the relevant information present. The activity of the neural pool in the spinal cord is constantly changing as a function of central and peripheral inputs. Hence, a given central command will have different results depending on the current state of the pool. From the perspective of movement planning and execution, the executive and integrative function is thus played by lower levels of the nervous system. Indeed, it is not entirely clear that a general division between central and peripheral processes is possible. A metaphor would be that an intended movement is realized successively in more motoric detail as it is passed down through the system until it reaches the final common path from the motor neurons to the muscles. Speech production would seem to share the same form of control, where the brainstem plays the integrative role. The rapid and functional compensations following perturbations to articulators are in agreement with such a distributed system.

2.4 Coordinate spaces

One persistent problem in movement control concerns the coordinate space in which movements are planned and represented (see Hollerbach, 1990, for a general discussion, and Munhall et al., 1991, for discussion of speech movements). In unrestrained reaching movements, the hand usually traverses a relatively straight path in an extrinsic cartesian coordinate system. If the same movement is described in an intrinsic coordinate system represented by the joint angles of the shoulder and elbow, a plot of elbow angle versus shoulder angle typically shows a curved path. One can similarly compare articulator path shapes observed during speech production in extrinsic versus intrinsic coordinates. For example, jaw movements can be represented in extrinsic or intrinsic coordinate space, where the latter involves at least rotation and translation of the jaw, possibly also yaw. For tongue movements, the situation is even more complex. Due to its mechanical linkage to the jaw, movements of the tongue are partly due to jaw rotation and translation, and partly to the activities of intrinsic and extrinsic tongue muscles. Moreover, the tongue has a hydrostatic skeleton, like an elephant's trunk, unlike the joints of the legs, the arms, and the jaw (cf. Smith & Kier, 1989; Stone, this volume).

Straight-path trajectories in extrinsic space have often been cited as evidence that movements are planned in extrinsic space. For speech, this argument can possibly be bolstered by the fact that the result of the speech production process is a time-varying acoustic signal. It has been argued that speech movements are controlled with respect to such acoustic effects. The acoustic effects depend on the transfer function of the vocal tract. Planning and control of speech movements in an acoustic coordinate system seem plausible. We should perhaps add, however, that tongue movements usually do not follow straight lines in extrinsic coordinate space but rather show curved paths (Houde, 1968; Perkell, 1969; Kent and Moll, 1972; Schönle, 1988; Munhall et al., 1991; Löfqvist & Gracco, 1999, 2002).

However, one traditional cause for concern is that control in extrinsic space requires the motor control system to solve the so-called inverse problem. A solution to the inverse problem entails going backwards from the desired movement trajectory to the muscle forces required to produce the movements. In arm movement control, the inverse problem involves mapping backwards from the desired movement goal in extrinsic space to the required muscular forces. For speech, the same mappings would be involved, perhaps with the added step of going from acoustic coordinates to vocal tract coordinates. The inverse problem is mathematically ill-posed in the sense that it is unclear whether a solution exists, is unique, and depends continuously on initial conditions (Tikhonov & Arsenin, 1977). One component of the inverse problem for arm movements is that the arm has excess degrees of freedom – seven (e.g., Alexander, 1992). Excess degrees of freedom in this context imply that the number of controlled spatial variables for the arm is less than the number of controlled joint angular variables. In such a case, the mapping from spatial variables to joint variables is indeterminate, since the same final position of the hand can be achieved by very many possible combinations of joint movements and, consequently, of very many different combinations of muscle activity patterns (there are 22 distinct muscles in the arm). In speech, the problem is the one-to-many mapping from acoustic signal to vocal tract area function as well as the excess degrees of freedom of the articulatory system. Models of speech production arguing for acoustically based targets would assume implicitly that speech movements are planned in acoustic space. Interestingly, when the acoustic properties of the vocal tract are changed experimentally, e.g., by having subjects wear a dental prosthesis, speakers do not compensate immediately for the induced changes (Hamlet & Stone, 1976, 1978; Baum &, McFarland, 1997; Munhall & Jones, 2003). Rather, such a manipulation requires some time for adjustment, possibly indicating that an inverse mapping has to be solved anew. These results might superficially seem to contradict the finding of immediate compensations when jaw movements are constrained by a bite-block held between a subject's teeth (Lindblom et al., 1979; Lubker, 1979; Fowler & Turvey, 1980). Note that the bite-block does not necessarily change the transfer function of the vocal tract in the same way as a dental prosthesis.

Studies using parallel processing (Jordan, 1990; Jordan & Rumelhart, 1992) suggest that the traditional computational concerns about the inverse problem may be exaggerated. We should also remember that speech and most other skilled movements are highly learned motor activities. Depending on what definition we use for mastering speech, it takes human infants two or three years to acquire it. Thus, in many instances of movement control, the motor system may not have to perform an exhaustive inverse computation, since learning can reduce the number of possible actions. In addition, speech is continuous and the changes in the vocal tract transfer function are slow, so that there is a continuity constraint in the sense that abrupt changes do not occur. Furthermore, much of the discussion about the inverse problem has received input from the field of robotics, but natural systems may take a loan on physics and evolution to solve this problem. Brains and nervous systems are not general-purpose devices, but rather special-purpose devices that have evolved to solve ecologically significant problems in a world governed by relatively stable and predictable physical forces.

In an attempt to alleviate the inverse problem, some investigators have argued that motor control is formulated in terms of muscular coordinates. One such model is the equilibrium-point model (Asatryan & Feldman, 1965; Feldman, 1966; see Bizzi et al., 1992, and commentaries for a review; Ostry et al., 1996; Perrier et al., 1996). According to the equilibrium-point model, the target of the movement is specified by the length and stiffness of agonist–antagonist muscle pairs working across a joint. This specification is made via central commands. We noted above that the tongue is a muscular hydrostat lacking joints. Equilibrium-point control of the tongue would nevertheless seem possible. By specifying relationships between the three major extrinsic tongue muscles the tongue can be moved up and down, forward and backward. Control of tongue shape can similarly be made by changing the relation between on the one hand the transverse and vertical muscles, and on the other hand the longitudinal muscles. There is some experimental evidence for such a control model. For example, Bizzi and colleagues (Polit & Bizzi, 1979) studied arm movements in monkeys who had been deprived of sensory information from the arm. The animals could still reach a visually presented target using the arm without kinesthetic or visual feedback about arm position. They could even do so when the arm was momentarily perturbed in the opposite direction, slowing the movement. Initially, it was thought that the target was set once and for all before the initiation of movement. In a later study (Bizzi et al., 1984), the perturbation was applied in the opposite direction during the reaching task. That is, the perturbation moved the arm towards the target position and thus assisted the movement. Contrary to expectations, this did not result in the arm reaching the target faster. Instead, after the perturbation had been released, the arm moved away from the target and returned to the position on its trajectory before the perturbation was applied. Hence, the whole trajectory is apparently not specified at the onset of movement but rather continuously updated.

Using an equilibrium-point approach, the muscles controlling movement in a joint can be modeled as a mass-spring system. This has certain attractive features (cf. Cooke, 1980). One of them is that movement will proceed in the face of transient perturbations (cf. section 2.1). Another one is equifinality, i.e., the intended goal will be reached from different initial conditions. An influential model of speech motor control is built on similar ideas (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989).

Before concluding this section, it is worth noting that people working with vowels tend to propose an acoustic/perceptual coordinate system, while those working with consonants prefer a system based in articulation. The reason is that vowels can easily be described in acoustic terms, while consonants are more readily described by their articulation, since they are produced with contacts between articulators. For example, results presented by Löfqvist and Gracco (1997) showed that the lips were moving at close to their peak velocities at the instant of labial closure for a bilabial stop consonant. The high velocity at the impact resulted in tissue compression making the airtight seal for the stop consonant. In addition, mechanical interactions between the lips were observed, with the lower lip pushing the upper lip upward due to its higher velocity. These results were compatible with the idea of a virtual target for the lips that would have them move beyond each other. Such a control strategy would ensure that the lips will make a closure irrespective of variations in their onset positions. The idea of a virtual target can also be applied to other consonants, since whenever two articulators meet, one of them is of soft tissue, e.g., the tongue contacting the hard palate during a lingual consonant. The results for tongue movements presented by Löfqvist and Gracco (2002) are thus compatible with the idea of a virtual target for the tongue in making a stop closure.

2.5 Coordinative structures

The speech perturbation studies suggest another property of movement control. In coordinated action, the level of control is not the individual muscles but rather task-dependent groupings of muscles. For example, perturbations to the jaw during the formation of a labial closure are compensated for by any combination of lip and jaw activity. It thus appears that individual articulators can be flexibly marshaled during speech to perform the intended closure in the vocal tract (cf. Gracco & Abbs, 1986).

Such task-dependent groupings of muscles have been called coordinative structures or synergies. This particular view of movement control owes much of its initial formulation to the Russian physiologist N. Bernstein (Bernstein, 1967). Further discussion and elaboration of these concepts are found in Gelfand, Gurfinkel, Fomin, et al. (1971), Turvey (1977, 1991), Kelso et al. (1980), Kugler et al. (1980), and Lee (1984). A synergy is defined as "those classes of movements which have similar kinematic characteristics, coinciding active muscle groups and conducting types of afferentation" (Gelfand, Gurfinkel, Tsetlin, et al., 1971, p. 331). According to Lee (1984), synergies can be defined by coherent patterns of muscle activity and/or movement, and in terms of spatial, temporal, and scaling properties. Spatially, the same set of muscles should be activated. In the temporal domain, synchronicity, stable order, or stable phase relationships should hold between events. Relations among events should demonstrate a scaling relationship. Such a definition requires appropriate measurements for a synergy to be recognized. For speech, the arguments for synergies have mostly been based on temporal and spatial relationships between muscle and/or movement patterns. As will be discussed in section 3, there are some intriguing experimental problems in defining synergies using timing and scaling properties. The most convincing evidence for coordinative structures would appear to come from the perturbation studies reviewed in section 2.1. In particular, a theory of coordinative structures predicts task-specific responses.

Coordinative structures should be seen as linkages between muscles that are set up for the execution of specific tasks. For example, Kelso et al. (1983) had subjects make flexion–extension movements of the index finger in synchrony with stressed and unstressed syllables. They noted a coupling between speech and finger movements. When producing a stressed syllable, the subjects also increased the amplitude of the finger movements. Similarly, when the finger was mechanically perturbed, a change in the acoustic speech signal was also observed. The authors argue that these findings can be accounted for by a coordinative structure comprising the vocal tract and the hand, set up for the execution of a specific task. Thus, when one member of the synergy was perturbed, other members also showed a change. We should note that the functionality of this particular coupled change is not entirely clear. Perhaps we should entertain another interpretation of these results.

Movements such as walking and swimming are rhythmic, and such movements can be effectively modeled by coupled oscillators producing many different patterns of organization (cf. Cohen et al., 1988; see also Stewart & Golubitsky, 1992, ch. 8, for a discussion of locomotion in terms of coupled oscillators). According to the oscillator model, a perturbation to a coupled system would manifest itself throughout the system. This class of models is very powerful for simulating coordinated rhythmic movements such as those found in locomotion, swimming, and chewing. The question arises, however, whether such rhythmic patterns are a property of normal speech movements.

One attractive feature of coordinative structures is that they can provide a principled solution to the problem of controlling many degrees of freedom. We noted above that the arm has several degrees of freedom, and this provides flexibility in the control of arm movements but also introduces the problem of indeterminacy in managing all the degrees. A coordinative structure can be described as a set of constraints between muscles that are set up to make the set of muscles behave as a unit. Thus, control is simplified in the sense that the individual muscles need not be controlled independently of each other but rather as a functional unit. It is obvious, however, that while control may be simplified at one level, complexities arise on other levels. If coordinative structures are task-specific and set up for brief periods of time to execute a given movement, there must be a way for the system to keep track of these different coordinative structures, to put them together and break them apart at the appropriate time.

2.6 A gestural approach to speech production

Records of speech movements generally show a succession of opening and closing movements at different locations in the vocal tract. One approach to understanding speech motor control is to posit underlying gestures as the building blocks of speech. A gesture can be defined briefly as a class of functionally equivalent movement patterns (cf. Saltzman & Munhall, 1989). Again, a word of caution is in order, since introducing underlying representations always carries a certain risk – such representations have a tendency to show an unprincipled rate of multiplication. Parsimony and a judicious use of Occam's razor are often desirable in science, although we are also well advised to keep in mind that the famous razor has been described as an instrument used by scientists to cut their own throats. Still, using gestures as underlying representations has certain advantages. It can possibly bypass the translation problem by providing the underlying linguistic units with more motoric detail. In this view, a segment should be viewed as a set of gestures (see Löfqvist, 1990, for a defense of the segment).

Munhall and Löfqvist (1992) examined how the two successive laryngeal movements in the utterance "Kiss Ted," for the /s/ and the /t/, were affected by variations in speaking rate. At a slow rate, two independent movements were found. At fast rates, a single movement was observed. Interestingly, at intermediate rates, a blend of the two gestures was seen. These blends could be reasonably well modeled by adding together two underlying gestures at different degrees of overlap. By varying speaking rate, it was thus possible to view the gestures both in isolation and as aggregates. Hence, the assumed underlying gestures could be readily observed. Similar effects of speaking rate on velar movements have been presented by Boyce et al. (1990).

Using underlying gestures to account for movement control is not a new idea. Aiming movements have often been shown to be composed of a number of submovements (e.g., Woodworth, 1899). Here, a large initial movement is followed by smaller corrective movements. It has been suggested by Milner and Ijaz (1990) that irregularities in the tangential velocity of aiming movements can be accounted for by linearly superimposing submovements to create a single composite movement. Similarly, when a subject is suddenly required to switch to a new target after a reaching movement has started, the initial movement is not aborted. Rather, a second movement is blended with the first one (Flash, 1990), and the resulting tangential velocity of the movement can be modeled by adding two underlying movements. For speech, Öhman (1966, 1967) showed evidence of gestural blending in VCV sequences when the vowels and medial consonant shared the same articulator: In the sequences /aga/ and /igi/, the tongue shape during the closure for the /g/ is a blend of the gestures for the vowels and the consonant (see Saltzman & Munhall, 1989, for simulation of such patterns using gestural blending). Thus, blending of gestures may be a general strategy that the motor system applies in implementing successive elements of movements.

2.7 Internal models

If an articulatory pattern is to be maintained and transmitted across generations of speakers, the pattern would have to either be recoverable by auditory or audiovisual means, or follow from general principles of biomechanics and motor control. For example, the larynx is not visible and it would thus appear that the only way to master linguistic contrasts involving the larynx is by listening to the sound produced and relating it to the sounds of other members of the linguistic community. Thus, a hearing impairment affects both babbling (Oller & Eilers, 1988) and speech development (Smith, 1975). Also adults who lose part or all of their hearing change their speech (e.g., Waldstein, 1990; Lane & Webster, 1991; Leder & Spitzer, 1993; Schenk et al., 2003). The role of auditory feedback in speech used to be examined in such "natural experiments," i.e., persons with congenital or adventitious hearing impairment. However, technological advances have made it possible to manipulate the auditory feedback that a speaker hears from his own voice without any significant time delay.

There are basically two types of experimental approaches to altering f_0 feedback. One of them is to introduce a sudden upward or downward shift in the perceived signal. In this kind of experiment, the majority of the responses are opposite, i.e., the subject responds by shifting his/her fundamental frequency in the opposite direction to the altered f_0. However, a varying number of following responses also occur, i.e., where the produced f_0 is altered in the same direction as the modification. An alternative procedure is to gradually shift the fundamental frequency over a number of trials, and examine the change point for the shift. In addition to the expected opposite change in f_0, studies using this paradigm also report an after-effect, i.e., after normal auditory feedback has been restored, the subject continues to produce an f_0 similar to the one used in the experimental condition

Many studies have examined manipulations of fundamental frequency in both running speech and sustained phonations under different experimental conditions (e.g., Kawahara, 1995; Burnett et al., 1998; Donath et al., 2002; Natke & Kalveram, 2001; Natke et al., 2003; Jones & Munhall, 2000). This technique has also been applied to tone languages such as Mandarin, where a tight control of fundamental frequency is necessary to produce tonal contrasts (e.g., Jones & Munhall, 2002; Xu et al., 2004). The results of these studies suggest that responses to altered auditory feedback can occur 120–200 ms after the onset of the modification. Both response magnitude and latency are affected by the nature of the speech task. Similar online changes in voice output have also been observed in experiments where the voice amplitude has been shifted (Heinks-Maldonado & Houde, 2005; Bauer et al., 2006).

This technique has also been used to alter the spectral contents of the speech signal by changing formant frequencies during whispered (Houde & Jordan, 1998, 2002) and voiced (Purcell & Munhall, 2006a, 2006b; Villacorta et al., 2007) speech. In these experiments, the change in formant frequency has been made gradually over a number of stimuli. In response to the altered spectral properties of the signal, subjects change formants in the opposite direction to the manipulation, and an after-effect is also observed. In response to manipulation of f_0, amplitude, or formants, the compensation made by the subject is never complete, but only a fraction of the change. The interpretation of these studies is usually framed in terms of an internal model that the speaker has of the relationship between vocal tract changes and acoustics (e.g., Guenther, 1995; Perkell et al., 1997; Kawato, 1999; Tremblay et al., 2003; Tin & Poon, 2005). For speech, such a model maps the relationship between articulatory movements and the acoustic signal. For the model to be maintained, it has to be updated. Thus, a loss of hearing will eventually affect speech output, although the details of the time course of the decay of the internal model is unknown. The studies gradually modifying auditory feedback show that the model is being changed during the course of the experiment, and also that an after-effect exists. That is, the model remains in the changed state for a period of time until it returns to the initial, unperturbed, state. A study by Jones and Keough (2008) compared the responses of singers and nonsingers to altered auditory feedback during singing. Only the singers showed an after-effect, thus suggesting that singers rely more on an internal model than nonsingers for the control of fundamental frequency. In contrast, the rapid shift of the perceived fundamental frequency in the pitch shift experiments indicates that this effect is most likely in part an automatic reflex, since the temporal windows involved are too short to affect an internal model.

In addition to auditory feedback, there is also somatosensory feedback from different types of receptors in muscles and joints. By applying a force in the forward direction during jaw movements in speech, Tremblay et al. (2003) showed that subjects initially made more jaw protrusion than normally but that they eventually gradually adapted to the changed force field and made normal jaw protrusion. When the force field was back to normal, subjects showed an after-effect by making less jaw protrusion than normally. These kinematic adaptations were not due to auditory feedback, since acoustic analysis showed no effects of the altered force field. Interestingly, the adaptation only occurred during normal or silent speech, but not for nonspeech movements.

3 Serial Control of Speech Movements

During normal speech production, movements of the articulators have to be made in the proper sequence to produce an acoustic signal that transmits the intended message. Figure 10.1 shows aerodynamic and articulatory records of three productions of the utterance “It’s a papaya,” spoken at a conversational rate. The top trace shows the air pressure in the oral cavity; there are three local increases in pressure associated with the voiceless consonants of the utterance. The middle trace shows the vertical movements of the lower lip; the lip moves upwards for the labial closure of the two bilabial stop consonants. The bottom trace shows the opening in the glottis. The glottis opens three times for the production of the voiceless consonants; the glottal movements for the stop and the fricative in /ts/ blend together. The signals for the three productions have been temporally aligned at the first peak glottal opening, associated with the cluster /ts/. The duration of the three productions differ. A change in utterance duration is evident in all three signals for that utterance. That is, movements of the lower lip and the glottis as well as the increase in oral air pressure all shift together. This is, in a sense, self-evident, since if it did not occur, speech would break down. The temporal coordination between articulatory movements has to be maintained within certain limits for speech to be intelligible, across changes in speaking rate. How this temporal cohesion is achieved is not well understood, however. It has been suggested that variations in speaking rate result in a scaling between the different articulatory movements that are involved in the production process. This suggestion is based on the following theoretical view. If someone is writing a word on a paper with a pencil or on a blackboard with a piece of chalk, different parts of the body are used. When the word is written on paper, writing involves movements of the hand and around the wrist; when it is written on the blackboard, the arm

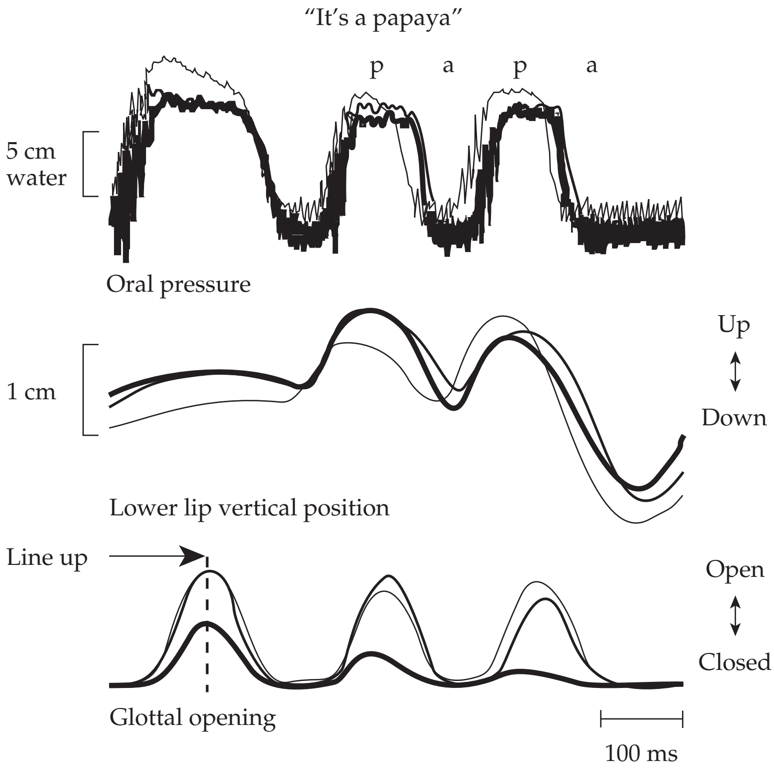


Figure 10.1 Records of three repetitions of the utterance "It's a papaya." The curves represent, from top to bottom, oral air pressure, lower-lip displacement, and glottal opening.

moves around the shoulder joint. Since the written pattern on the blackboard can be seen as a scaled version of the one on paper, it has generally been argued that there is a single underlying representation of the movement pattern that is instantiated by different parts of the body using a scaling relation. The alternative view, that each pattern is stored as a separate entity, is at least intuitively implausible and inefficient. Thus, the claim is that the pattern is stored as a "generalized motor program" that can be re-parameterized (see Schmidt, 1975). A generalized motor program predicts that when variations in speed and amplitude of a movement complex occur, the relationship between the individual movements should remain virtually unchanged. The reason is that a submovement interval should maintain a constant proportion of the whole movement interval. Hence, the model is usually referred to as a proportional duration model (see Heuer, 1991, for a general discussion of such models). Initially, several studies claimed that proportional timing was indeed found for motor activities like locomotion (Shapiro et al., 1981), handwriting (Viviani & Terzuolo, 1980), typing (Terzuolo & Viviani, 1979), and speech (Tuller & Kelso, 1984).

Gentner (1987) proposed a stronger test of proportional duration by examining whether the ratio between one movement interval and the duration of the whole movement sequence is unrelated to the duration of the whole movement sequence. The proportional duration model predicts that this should be the case, since the duration of all the components of a movement sequence should maintain a constant proportion of the overall duration. Studies applying this statistical analysis suggest that proportional timing does not occur in speech or any other motor activity that has been examined (cf. Sock et al., 1988; Wann & Nimmo-Smith, 1990; Löfqvist, 1991). The slope of the regression usually deviates from zero. One methodological uncertainty facing students of speech timing should be mentioned in this context. Studies of temporal phenomena by necessity have to break up the flow of articulatory movements into discrete intervals for measurement. To delimit these intervals, movement onset and offset, and peak velocity of movement are commonly used. It is, of course, possible that these events are not the ones that the nervous system uses for controlling movements. Kelso et al. (1986) suggested that the proper metric for constant relative timing is phase as measured on a phase plane, rather than ratio of articulatory intervals, and presented some evidence in support of this notion. In a phase plane representation, position is plotted against velocity. In a sequence of a vowel + labial consonant + vowel, a phase plane plot of the jaw or the lower lip shows an elliptical orbit. Using this kind of representation, movement onsets for different articulators can be defined in terms of phase relationships. Further studies have, however, failed to replicate their findings (Lubker, 1986; Nittrouer et al., 1988; Nittrouer, 1991). These results have implications for theories of speech motor control based on coordinative structures. When discussing coordinative structures in section 2.5, we noted that a definition based on temporal relations requires fixed intervals or scaling among components. One interpretation of scaling is proportional timing and this, as we have seen, does not appear to occur, or at least the scaling is not linear. An important task for speech motor control is to define the metric that governs temporal relations among speech movements.

Speech movements thus show temporal cohesion even though they do not appear to follow a proportional duration model. What are the rules governing this cohesion? Admittedly, not very much is known about this problem, although a reasonable assumption is that intervals that are important for the integrity of the speech signal will show relatively less variability than others. An experiment by Saltzman et al. (1992) tried to shed some light on this issue using the perturbation paradigm discussed above (see also Gracco & Abbs, 1989). As a subject was producing the pseudo-word /pæsæpæpple/, a mechanical load was applied to the lower lip, pulling the lip downwards; the load was applied at different points in time during the production of the utterance. They found that the temporal intervals between the successive bilabial closing movements for the stop consonants were systematically affected by the perturbation. Most of the timing changes occurred during the lip-opening phases of these intervals; these phases are associated with the production of the vowels. The closing phases were relatively resistant to temporal distortion, suggesting that their durations were, in some sense, more actively controlled.

In discussions about timing control of speech movements, a confusing issue has been whether timing is intrinsic or extrinsic. According to an extrinsic timing model, time is metered out by a central clock or time keeper that is, in a sense, outside the movement itself. Proponents of intrinsic timing argue that time may not be represented outside the movement but is rather inside it (cf. Fowler, 1980; Kelso & Tuller, 1987). It seems safe to conclude that the solution depends on the level of description being adopted. According to the equilibrium-point model of movement control, the endpoint of the movement is specified in terms of the relationship between the stiffness or activation thresholds of agonist and antagonist muscles. The duration of the movement trajectory thus depends on the dynamical system defined among muscles, and there may be no timing device keeping track of the progression of the movement. In this sense, time is not represented outside the movement by a time keeper but is rather intrinsic to it. However, for movements to be properly executed and sequenced, the equilibrium points have to be reset continuously. These changes have to be made at the appropriate points in time and the system must have some time-keeping mechanism to make them. At this level, the time keeping should thus more properly be considered extrinsic.

What are the properties of the clock or time keeper? Again, applying mechanical perturbations to movements may provide some clues. For rhythmic movements, phase resetting analysis can be used (see Winfree, 1980, and Glass & Mackey, 1988, for a general discussion). In this type of experiment, one measures the temporal shift that is introduced by a perturbation relative to the timing pattern of the pre-perturbation rhythm. If a phase shift is found, the implication is that a central clock does not drive the periphery in a unidirectional manner. Rather, the central–peripheral coupling is bi-directional, since feedback from the periphery affects the clock. Studies of rhythmic finger movements (Kay et al., 1991) suggest that mechanical perturbations do introduce shifts in the phasing of such movements. Results reported by Saltzman (1992) indicate that this is also the case for speech, at least when the speech task consists of the repetition of a single consonant-vowel syllable. Saltzman et al. (1998) reported results from a series of phase-resetting studies of speech production that investigate whether intergestural temporal cohesion is greater within segments than between segments. In this experiment, the coordinated gestures were bilabial closing and laryngeal devoicing gestures for /p/s in successive syllable onsets. The phase-resetting techniques are used to determine whether perturbations delivered during an ongoing rhythm have a permanent effect (i.e., phase shift) on the underlying temporal organization of the rhythm. In such studies, what is measured is the amount of temporal shift introduced by the perturbation, relative to the timing pattern that existed prior to the perturbation. The finding of a post-perturbation, steady-state phase shift using this method would support the hypothesis that there exists a central timing network that both drives the articulatory periphery and whose state is altered (phase-shifted) by feedback specific to events at the periphery.

The perturbations used in this study consisted of a downward pull on the lower lip that was delivered via a paddle resting on the lower lip and that was connected to a torque motor. The results led to several conclusions. First, the steady-state analyses of the speakers' repetitive utterances indicated that "permanent" phase shifts existed for both the lips and the larynx after the system returned to its pre-perturbation steady-state rhythm. These results support the hypothesis that central intergestural dynamics can be reset by peripheral articulatory events. Such resetting was strongest when the downwardly directed lower-lip perturbation was delivered near the initiation or acceleration portion of the actively controlled bilabial closing gesture for /p/. Second, analyses of the transient portions of the perturbed cycles of the repetitive utterances indicated that perturbation-induced steady-state phase shifts were almost totally attributable to changes occurring during the first two perturbed cycles. Third, in addition to the steady-state shifts in the timing between successive bilabial closing and laryngeal devoicing gestures for /p/, steady-state shifts in the *relative* phasing of gestures were also demonstrated. However, the individual temporal shifts of the bilabial and laryngeal gestures were an order of magnitude larger than the relative temporal shift between these gestures, and the lips and larynx appeared to be phase-advanced as a relatively coherent unit. Thus, these results not only demonstrate a resetting of a central "clock" for these utterances, but also imply that intergestural cohesion is greater within segments (i.e., between labial and laryngeal gestures during each /p/) than between segments (i.e., between labial or laryngeal gestures in the /p/s of successive syllables), as has been hypothesized by Byrd (1996), Löfqvist (1991), Nittrouer et al. (1988), Saltzman and Munhall (1989) (see also Gracco & Löfqvist, 1994). The notion of segment used here entails that the closing and opening of the lips together with the opening and closing of the glottis is made within a segment consisting of a stop consonant. In contrast, the intervals between successive lip openings and glottal openings are associated with different vowel and consonant segments.

4 Summary

The theoretical and empirical approaches to speech production that we have discussed in this chapter converge in their focus on understanding how the different parts of the vocal tract are flexibly marshaled and coordinated to produce the acoustic signal that the speaker uses to convey a message. A variety of experimental paradigms are currently being applied to the problem of coordination and control in motor systems with excess degrees of freedom. Progress in speech motor control is likely to benefit from input from other areas of movement control and in using a combined strategy of empirical studies and mathematical modeling.

ACKNOWLEDGMENT

This work was supported by Grant DC-00865 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

REFERENCES

Abbs, J. & Gracco, V. L. (1984) Control of complex motor gestures: Orofacial muscle responses to load perturbations during speech. *Journal of Neurophysiology*, 51, 705–23.

Abbs, J., Gracco, V. L., & Cole, K. (1984) Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming. *Journal of Motor Behavior*, 16, 195–232.

Alexander, R. McN. (1992) *The Human Machine*. New York: Columbia University Press.

Asatryan, D. & Feldman, A. (1965) Functional tuning of the nervous system with control of movement or maintenance of a steady posture, I: Mechanographic analysis of the work of the joint or execution of a postural task. *Biofizika*, 10, 837–46 [English translation 925–35].

Bauer, J., Mittal, K., Larson, C., & Hain, T. (2006) Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. *Journal of the Acoustical Society of America*, 119, 2363–71.

Baum, S. & McFarland, D. (1997) The development of speech adaptation to an artificial palate. *Journal of the Acoustical Society of America*, 102, 2353–9.

Bernstein, N. (1967) *The Coordination and Regulation of Movements*. London: Pergamon Press.

Bizzi, E., Accornero, N., Chapple, W., & Hogan, N. (1984) Posture control and trajectory formation during arm movement. *Journal of Neuroscience*, 4, 2738–44.

Bizzi, E., Hogan, N., Mussa-Ivaldi, F., & Giszter, S. (1992) Does the nervous system use equilibrium-point control to guide single and multiple joint movements? *Behavioral and Brain Sciences*, 15, 603–13.

Boyce, S., Krakow, R. A., Bell-Berti, F., & Gelfer, C. (1990) Converging sources of evidence for dissecting articulatory movements into core gestures. *Journal of Phonetics*, 18, 173–88.

Burnett, T., Freedland, M., Larson, C., & Hain, T. (1998) Voice frequency responses to manipulation in pitch feedback. *Journal of the Acoustical Society of America*, 103, 3153–61.

Byrd, D. (1996) A phase window framework for articulatory timing. *Phonology*, 13, 139–69.

Cohen, A., Rossignol, S., & Grillner, S. (eds.) (1988) *Neural Control of Rhythmic Movements in Vertebrates*. New York: Wiley.

Cole, K., Gracco, V. L., & Abbs, J. (1984) Autogenic and nonautogenic sensorimotor actions in the control of multiarticulate hand movements. *Experimental Brain Research*, 56, 582–5.

Cooke, J. (1980) The organization of simple, skilled movements. In G. Stelmach & J. Requin (eds.), *Tutorials in Motor Behavior* (pp. 199–212). Amsterdam: North-Holland.

Dell, G., Juliano, C., & Govindjee, A. (1993) Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, 17, 149–95.

Donath, T., Natke, U., & Kalveram, T. (2002) Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *Journal of the Acoustical Society of America*, 111, 357–66.

Fant, G. (1962) Descriptive analysis of the acoustic aspects of speech. *Logos*, 5, 3–17. (Reprinted in Fant, G. (1973). *Speech Sounds and Features* (pp. 17–23). Cambridge, MA: MIT Press.)

Feldman, A. (1966) Functional tuning of the nervous system during control of movement or maintenance of a steady posture, III: Mechanographic analysis of the execution by man of the simplest motor task. *Biophysics*, 11, 766–75.

Flash, T. (1990) Organization of human arm trajectory control. In J. Winters & S. Woo (eds.), *Multiple Muscle Systems: Biomechanics and Movement Organization* (pp. 282–301). New York: Springer.

Flege, J., Fletcher, S., & Homiedan, A. (1988) Compensating for a bite block in /s/ and /t/ production: Palatographic, acoustic and perceptual data. *Journal of the Acoustical Society of America*, 83, 212–28.

Folkins, J. & Abbs, J. (1975) Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research*, 18, 207–20.

Folkins, J. & Zimmermann, G. (1982) Lip and jaw interaction during speech: Responses to perturbation of lower-lip movement prior to bilabial closure. *Journal of the Acoustical Society of America*, 71, 1225–33.

Fowler, C. A. (1980) Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8, 113–33.

Fowler, C. A. (1993) Phonological and articulatory characteristics of spoken language. In G. Blasken, J. Dittman, H. Grimm, J. Marshall, & C.-W. Wallesch (eds.), *Linguistic Disorders and Pathologies: An International Handbook* (pp. 34–46). Berlin: Walter de Gruyter.

Fowler, C., Rubin, P., Remez, R., & Turvey, M. (1980) Implications for speech production of a general theory of action. In B. Butterworth (ed.), *Language Production I: Speech and Talk* (pp. 373–420). London: Academic Press.

Fowler, C. & Smith, M. (1986) Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (eds.), *Invariance and Variability of Speech Processes* (pp. 123–36). Hillsdale, NJ: Lawrence Erlbaum.

Fowler, C. & Turvey, M. (1980) Immediate compensation in bite-block speech. *Phonetica*, 37, 306–26.

Garrett, M. (1980) Levels of processing in sentence production. In B. Butterworth (ed.), *Language Production I: Speech and Talk* (pp. 177–220). London: Academic Press.

Gelfand, I., Gurfinkel, V., Fomin, S., & Tsetlin, M. (eds.) (1971) *Models of the Structural-Functional Organization of Certain Biological Systems*. Cambridge, MA: MIT Press.

Gelfand, I., Gurfinkel, V., Tsetlin, M., & Shik, M. (1971) Some problems in the analysis of movements. In I. Gelfand, V. Gurfinkel, S. Fomin, & M. Tsetlin (eds.), *Models of the Structural-Functional Organization of Certain Biological Systems* (pp. 329–45). Cambridge, MA: MIT Press.

Gentner, D. (1987) Timing of skilled movements: Test of the proportional duration model. *Psychological Review*, 94, 255–76.

Glass, L. & Mackey, M. (1988) *From Clocks to Chaos: The Rhythms of Life*. Princeton: Princeton University Press.

Gomi, H., Honda, M., Ito, T., & Murano, E. (2002) Compensatory articulation during bilabial fricative production by regulating muscle stiffness. *Journal of Phonetics*, 30, 261–79.

Gracco, V. L. & Abbs, J. (1985) Dynamic control of the perioral system during speech: Kinematic analyses of autogenic and nonautogenic sensorimotor processes. *Journal of Neurophysiology*, 54, 418–32.

Gracco, V. L. & Abbs, J. (1986) Variant and invariant characteristics of speech movements. *Experimental Brain Research*, 65, 156–66.

Gracco, V. L. & Abbs, J. (1988) Central patterning of speech movements. *Experimental Brain Research*, 71, 515–26.

Gracco, V. L. & Abbs, J. (1989) Sensorimotor characteristics of speech motor sequences. *Experimental Brain Research*, 75, 586–98.

Gracco, V. L. & Löfqvist, A. (1994) Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. *Journal of Neuroscience*, 14, 6585–97.

Granit, R. (1970) *The Basis of Motor Control*. London: Academic Press.

Guenther, F. (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102, 594–621.

Hamlet, S. & Stone, M. (1976) Compensatory vowel characteristics resulting from the presence of different types of experimental prostheses. *Journal of Phonetics*, 4, 199–218.

Hamlet, S. & Stone, M. (1978) Compensatory alveolar consonant production induced by wearing a dental prosthesis. *Journal of Phonetics*, 6, 227–48.

Hardcastle, W. (1981) Experimental studies in lingual coarticulation. In R. Asher & E. Henderson (eds.), *Towards a History of Phonetics* (pp. 50–66). Edinburgh: Edinburgh University Press.

Heinks-Maldonado, T. & Houde, J. (2005) Compensatory responses to brief perturbations of speech amplitude. *Acoustics Research Letters Online*, 6, 131–7.

Heuer, H. (1991) Invariant timing in motor-program theory. In J. Fagard & P. Wolfe (eds.), *The Development of Timing Control and Temporal Organization in Coordinated Action* (pp. 37–68). Amsterdam: Elsevier.

Hockett, C. (1955) *A Manual of Phonology* (*International Journal of American Linguistics*, Memoir 11). Baltimore: Waverly Press.

Hollerbach, J. (1990) Planning of arm movements. In D. Osherson, S. Kosslyn, & J. Hollerbach (eds.), *Visual Cognition and Action. An Invitation to Cognitive Science*, vol. 2 (pp. 183–211). Cambridge, MA: MIT Press.

Honda, M., Fujino, A., & Kaburagi, T. (2002) Compensatory responses of articulators to unexpected perturbation of the palate shape. *Journal of Phonetics*, 30, 281–302.

Houde, J. & Jordan, M. (1998) Sensorimotor adaptation in speech production. *Nature*, 279, 1213–16.

Houde, J. & Jordan, M. (2002) Sensorimotor adaptation of speech I: Compensation and adptation. *Journal of Speech Language and Hearing Research*, 45, 295–310.

Houde, R. (1968) *A Study of Tongue Body Motion During Selected Speech Sounds* (SCRL Monograph 2). Santa Barbara: Speech Communications Research Laboratory.

Humphrey, D. R. & Freund, H.-J. (eds.) (1991) *Motor Control: Concepts and Issues*. Chichester, UK: Wiley.

Jones, J. & Keough, D. (2008) Auditory-motor mapping for pitch control in singers and nonsingers. *Experimental Brain Research*, 190, 279–87.

Jones, J. & Munhall, K. (2000) Perceptual calibration of F0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America*, 108, 1246–51.

Jones, J. & Munhall, K. (2002) The role of auditory feedback during phonation: Studies of Mandarin tone production. *Journal of Phonetics*, 30, 303–20.

Jordan, M. (1990) Motor learning and the degrees of freedom problem. In M. Jeannerod (ed.), *Attention and Performance XII: Motor Representation and Control* (pp. 796–836). Hillsdale, NJ: Lawrence Erlbaum.

Jordan, M. & Rumelhart, D. (1992) Forward models: Supervised learning with a distal teacher. *Cognitive Psychology*, 16, 307–54.

Kawahara, H. (1995) Transformed auditory feedback: The collection of data from 1993.1 to 1994.12 by a new set of analysis procedures. *TR-H-120* (ATR Human Information Processing Laboratories, Kyoto), 1–52.

Kawato, M. (1999) Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9, 718–27.

Kay, B., Saltzman, E., & Kelso, J. A. S. (1991) Steady-state and perturbed rhythmical movements: A dynamical analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 183–97.

Keele, S. (1968) Movement control in skilled motor performance. *Psychological Bulletin*, 70, 387–403.

Kelso, J. A. S., Holt, K., Kugler, P., & Turvey, M. (1980) On the concept of coordinative structures as dissipative structures II: Empirical lines of convergence. In G. Stelmach & J. Requin (eds.), *Tutorials in Motor Behavior* (pp. 49–70). Amsterdam: North-Holland.

Kelso, J. A. S., Saltzman, E., & Tuller, B. (1986) The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, 14, 29–59.

Kelso, J. A. S. & Tuller, B. (1987) Intrinsic time in speech production: Theory, methodology, and preliminary observations. In E. Keller & M. Gopnik (eds.), *Motor and Sensory Processes of Language* (pp. 203–19). Hillsdale, NJ: Lawrence Erlbaum.

Kelso, J. A. S., Tuller, B., & Harris, K. S. (1983) A "dynamic pattern" perspective on the control and coordination of movement. In P. MacNeilage (ed.), *The Production of Speech* (pp. 137–73). New York: Springer.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. (1984) Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 812–32.

Kent, R. D. (1976) Models of speech production. In N. Lass (ed.), *Contemporary Issues in Experimental Phonetics* (pp. 79–104). New York: Academic Press.

Kent, R. D. & Moll, K. (1972) Cinefluorographic analyses of selected lingual consonants. *Journal of Speech and Hearing Research*, 15, 453–73.

Kozhevnikov, V. A. & Chistovich, L. A. (1965) *Speech: Articulation and Perception*. Washington, DC: Joint Publication Research Service.

Kugler, P., Kelso, J. A. S., & Turvey, M. (1980) On the concept of coordinative structures as dissipative structures I: Theoretical lines of convergence. In G. Stelmach & J. Requin (eds.), *Tutorials in Motor Behavior* (pp. 3–47). Amsterdam: North-Holland.

Kühnert, B. & Nolan, F. (1999) The origin of coartculation. In W. Hardcastle & N. Hewlett (eds.), *Coarticulation: Theory, Data and Techniques* (pp. 7–30). Cambridge: Cambridge University Press.

Lane, H. & Webster, J. (1991) Speech deterioration in postlingually deafened adults. *Journal of the Acoustical Society of America*, 89, 859–66.

Lashley, K. S. (1951) The problem of serial order in behavior. In L. A. Jeffress (ed.), *Cerebral Mechanisms in Behavior* (pp. 112–36). New York: Wiley.

Leder, S. & Spitzer, J. (1993) Speaking fundamental frequency, intensity, and rate of adventitiously profoundly hearing-impaired women. *Journal of the Acoustical Society of America*, 93, 2146–51.

Lee, W. (1984) Neuromotor synergies as a basis for coordinated intentional action. *Journal of Motor Behavior*, 16, 135–70.

Levelt, W. (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Liberman, A., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review*, 74, 431–61.

Lindblom, B., Lubker, J., & Gay, T. (1979) Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation. *Journal of Phonetics*, 7, 147–62.

Löfqvist, A. (1990) Speech as audible gestures. In W. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling* (pp. 289–322). Dordrecht: Kluwer.

Löfqvist, A. (1991) Proportional timing in speech motor control. *Journal of Phonetics*, 19, 343–50.

Löfqvist, A. (1992) Acoustic and aerodynamic effects of interarticulator timing in voiceless consonants. *Language and Speech*, 35, 15–28.

Löfqvist, A. & Gracco, V. L. (1991) Discrete and continuous modes in speech motor control. *PERILUS XIV*, University of Stockholm, Institute of Linguistics, 27–34.

Löfqvist, A. & Gracco, V. L. (1997) Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language, and Hearing Research*, 40, 877–93.

Löfqvist, A. & Gracco, V. (1999) Interarticulator programming in VCV sequences: Lip and tongue movements. *Journal of the Acoustical Society of America*, 105, 1864–76.

Löfqvist, A. & Gracco, V. (2002) Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America,* 111, 2811–27.

Löfqvist, A. & Yoshioka, H. (1984) Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication*, 3, 279–89.

Lubker, J. (1979) The reorganization time of bite-block vowels. *Phonetica*, 36, 273–93.

Lubker, J. (1986) Articulatory timing and the concept of phase. *Journal of Phonetics*, 14, 133–7.

MacNeilage, P. (1970) The motor control of serial ordering in speech. *Psychological Review*, 77, 182–96.

Matthews, P. (1972) *Mammalian Muscle Receptors and Their Central Actions*. London: Edward Arnold.

McCrea, D. (1992) Can sense be made of spinal interneuron circuits? *Behavioral and Brain Sciences*, 15, 633–43.

Menzerath, P. & de Lacerda, A. (1933) *Koartikulation, Steuerung und Lautabgrenzung* [Coarticulation, control, and segementation of speech]. Bonn: Ferdinand Dümmlers Verlag.

Milner, T. & Ijaz, M. (1990) The effect of accuracy constraints on three dimensional movement kinematics. *Neuroscience*, 35, 365–74.

Munhall, K. & Jones, J. (2003) Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Journal of the Acoustical Society of America*, 113, 532–43.

Munhall, K. & Löfqvist, A. (1992) Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics*, 20, 111–26.

Munhall, K., Löfqvist, A., & Kelso, J. A. S. (1994) Lip–larynx coordination in speech: Effects of mechanical perturbations to the lower lip. *Journal of the Acoustical Society of America*, 65, 3605–16.

Munhall, K., Ostry, D., & Flanagan, J. (1991) Coordinate spaces in speech planning. *Journal of Phonetics*, 19, 293–307.

Nashner, L. & McCollum, G. (1985) The organization of human postural movements: A formal basis and experimental synthesis. *Behavioral and Brain Sciences*, 8, 135–72.

Natke, U., Donath, T., & Kalveram, K. (2003) Control of voice fundamental frequency in speaking versus singing. *Journal of the Acoustical Society of America*, 113, 1587–93.

Natke, U. & Kalveram, K. (2001) Effects of frequency-shifted auditory feedback on long stressed and unstressed syllables. *Journal of Speeech Language and Hearing Research*, 44, 577–84.

Newsom Davis, J. & Sears, T. (1970) The proprioceptive reflex control of the intercostal muscles during their voluntary activation. *Journal of Physiology*, 209, 711–38.

Nittrouer, S. (1991) Phase relations of jaw and tongue tip movements in the production of VCV utterances. *Journal of the Acoustical Society of America*, 90, 1806–15.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988) Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America*, 84, 1653–61.

Oller, K. & Eilers, R. (1988) The role of audition in infant babbling. *Child Development*, 59, 441–9.

Öhman, S. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–68.

Öhman, S. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41, 310–20.

Ostry, D., Gribble, P., & Gracco, V. (1996) Coarticulation of of jaw movements in speech production: Is context sensitivity in speech kinematics centrally planned? *Journal of Neuroscience*, 16, 1570–9.

Pattee, H. (1977) Dynamic and linguistic modes of complex systems. *International Journal of General Systems*, 3, 259–66.

Perkell, J. (1969) *Physiology of Speech Production*. Cambridge, MA: MIT Press.

Perkell, J., Matthies, M., Lane, H. et al. (1997) Speeech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication*, 22, 227–50.

Perrier, P., Ostry, D., & Laboissière, R. (1996) The equilibrium point hypothesis and its application to speech motor control. *Journal of Speech and Hearing Research*, 39, 365–78. (Reprinted with commentaries in *Bulletin de la Communication Parlée*, 4, 1998, Institut de la Communication Parlée, Grenoble.)

Polit, A. & Bizzi, E. (1979) Characteristics of motor programs underlying arm movements in monkeys. *Journal of Neurophysiology*, 42, 183–94.

Purcell, D. & Munhall, K. (2006a) Compensation following real-time manipulation of formants in isolated vowels. *Journal of the Acoustical Society of America*, 119, 2288–97.

Purcell, D. & Munhall, K. (2006b) Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, 120, 966–77.

Rothwell, J., Traub, M., & Marsden, C. (1982) Automatic and "voluntary" responses compensating for disturbances of human thumb movements. *Brain Research*, 248, 33–41.

Salthouse, T. (1986) Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin*, 99, 303–19.

Saltzman, E. (1992) Biomechanic and haptic factors in the temporal patterning of limb and speech activity. *Human Movement Science*, 11, 239–51.

Saltzman, E. & Kelso, J. A. S. (1987) Skilled actions: A task dynamic approach. *Psychological Review*, 94, 84–106.

Saltzman, E., Löfqvist, A., Kay, B., Kinsella-Shaw, J., & Rubin, P. (1998) Dynamics of intergestural timing: A perturbation study of lip–larynx coordination. *Experimental Brain Research*, 123, 412–24.

Saltzman, E., Löfqvist, A., Kinsella-Shaw, J., Rubin, P., & Kay, B. (1992) A perturbation study of lip–larynx coordination. In J. Ohala, T. Neary, B. Derwing, M. Hodge, & G. Wiebe (eds.), *ICSLIP 92 Proceedings*, addendum (pp. 19–22). Edmonton: The University of Alberta.

Saltzman, E. & Munhall, K. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1, 333–82.

Savariaux, C., Perrier, P., & Orliaguet, J. (1995) Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *Journal of the Acoustical Society of America*, 98, 78–88.

Schenk, B., Baumgartner, W., & Hamzavi, J. (2003) Effect of the loss of auditory feedback on segmental parameters of vowels of postlingually deafened speakers. *Auris Nasus Larynx*, 30, 333–9.

Schmidt, R. (1975) A schema theory for discrete motor skill learning. *Psychological Review*, 82, 225–60.

Schönle, P. W. (1988) *Elektromagnetische Artikulographie* [Electromagnetic articulography]. Berlin: Springer.

Shaiman, S. (1989) Kinematic and electromyographic responses to perturbation of the jaw. *Journal of the Acoustical Society of America*, 86, 78–88.

Shaiman, S. & Abbs, J. (1987) Sensorimotor contributions to the temporal coordination of oral and laryngeal movements. *SMLC Preprints* (Speech Motor Control Laboratories, University of Madison) Spring–Summer 1987, 185–202.

Shaiman, S. & Gracco, V. (2002) Task-specific sensorimotor interactions in speech production. *Experimental Brain Research*, 146, 411–18.

Shapiro, D., Zernicke, R., Gregor, R., & Diestel, J. (1981) Evidence for generalized motor programs using gait pattern analysis. *Journal of Motor Behavior*, 13, 33–47.

Smith, B. (1975) Residual hearing and speech production in deaf children. *Journal of Speech and Hearing Research*, 18, 795–811.

Smith, K. & Kier, W. (1989) Trunks, tongues, and tentacles: Moving with skeletons of muscle. *American Scientist*, 77, 29–35.

Sock, R., Ollila, L., Delattre, C., Zilliox, C., & Zohair, L. (1988) Patrons de phases dans le cycle acoustique de détente en français. *Journal Acoustique*, 1, 339–45.

Stewart, I. & Golubitsky, M. (1992) *Fearful Symmetry*. Oxford: Blackwell.

Terzuolo, C. & Viviani, P. (1979) The central representation of learned motor patterns. In R. Talbot & D. Humphrey (eds.), *Posture and Movement* (pp. 113–21). New York: Raven Press.

Tikhonov, A. and Arsenin, V. (1977) *Solutions of ill-posed problems*. Washington, DC: W.H. Winstron.

Tin, C. & Poon, C. (2005) Internal models in sensorimotor integration: Perspectives from adaptive control theory. *Journal of Neural Engineering*, 2, S147–163.

Traub, M., Rothwell, J., & Marsden, C. (1980) A grab reflex in the human hand. *Brain*, 103, 869–84.

Tremblay, S., Shiller, D., & Ostry, D. (2003) Somatosensory basis of speech production. *Nature*, 423, 866–9.

Tuller, B. & Kelso, J. A. S. (1984) The timing of articulatory gestures: Evidence for relational invariants. *Journal of the Acoustical Society of America*, 76, 1030–6.

Turvey, M. (1977) Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (eds.), *Perceiving, Acting and Knowing: Toward an Ecological Psychology* (pp. 211–65). Hillsdale, NJ: Lawrence Erlbaum.

Turvey, M. (1991) Coordination. *American Psychologist*, 45, 938–53.

Villacorta, V., Perkell, J., & Guenther, F. (2007) Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America*, 122, 2306–19.

Viviani, P. & Terzuolo, C. (1980) Space-time invariance in learned motor skills. In G. Stelmach & J. Requin (eds.), *Tutorials in Motor Behavior* (pp. 525–33). Amsterdam: North-Holland.

Waldstein, R. (1990) Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. *Journal of the Acoustical Society of America*, 88, 2099–114.

Wann, J. & Nimmo-Smith, I. (1990) Evidence against the relative invariance of timing in handwriting. *Quarterly Journal of Experimental Psychology*, 42A, 105–19.

Winfree, A. (1980) *The Geometry of Biological Time*. New York: Springer.

Woodworth, R. (1899) The accuracy of voluntary movement. *Psychological Review*, 3, 1–114.

Xu, Y., Larson, C., Bauer, J., & Hain, T. (2004) Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *Journal of the Acoustical Society of America*, 116, 1168–78.

11 Voice Source Variation and Its Communicative Functions

CHRISTER GOBL AND
AILBHE NÍ CHASAIDE

1 Introduction

This chapter deals with acoustic aspects of phonation and explores how phonation is exploited in speech communication. The early sections focus on the *voice source* signal, which is typically defined as the airflow, or volume velocity, through the glottis. It varies in a periodic way which reflects the rapid opening and closing of the vibrating vocal folds. The source for voiceless sounds is not dealt with here.

In running speech the voice source signal is constantly modulated, and these modulations carry different kinds of information to the listener. Some of these variations are substantial and are clearly heard by the listener as shifts in voice quality, say as between modal and breathy voice. Some of the modulations are not perceived in this way, but contribute in other ways, either to the meaning of the message, or to our ability to identify the social or individual characteristics of the speaker. In the later sections of this chapter, we discuss the various determinants of voice source variation, and their role in linguistic, paralinguistic, sociolinguistic, and extralinguistic strands of spoken communication.

The glottal airflow constitutes the input signal to the vocal tract which acts as an acoustic filter. The configuration of the supraglottal vocal organs determines the specific resonance characteristics of this filter. For a given source signal, a large number of segments may be differentiated from each other on the basis of the particular patterning of resonances and anti-resonances that different supraglottal filters impart.

Figure 11.1 shows schematically the speech production process for two vowels [u] and [i]. In this illustration, the (amplitude) spectrum of the source is identical in both cases: it contains all harmonic components and has a constant slope of –12 dB per octave. This means that the amplitude of the harmonics decreases monotonically with increasing frequency, so that for every doubling of frequency, the amplitude has dropped by 12 dB. We should note that this is the ideal case. The true source spectrum does not have a constant slope, and may present local dips depending on the precise shape of the glottal pulse.

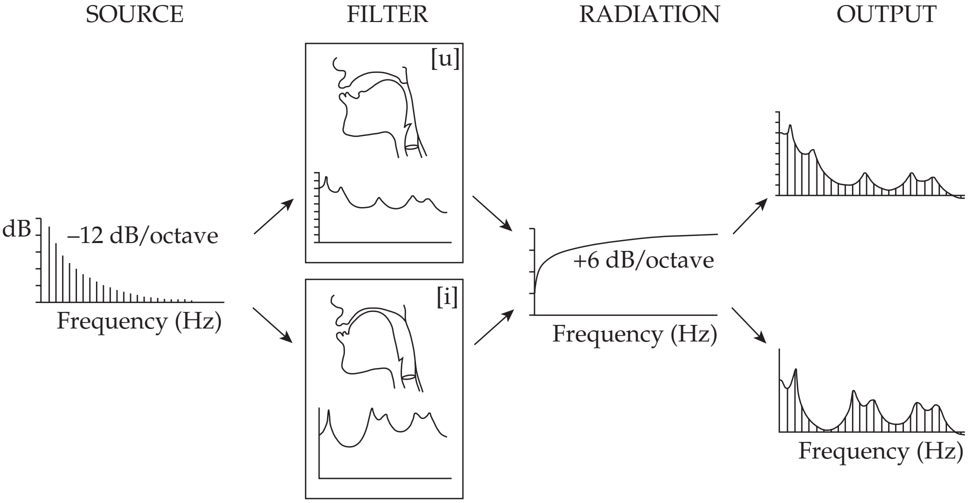


Figure 11.1 A schematic illustration of the speech production process for the vowels [u] and [i].

The filtering effect of the vocal tract, referred to as the transfer function, is rather different for the two vowels, due to the different positions adopted by the tongue and the lips. Source harmonics which fall at or near the peaks of the transfer function will be amplified by the filter. Harmonics which do not come near to the peaks will not be amplified and may be attenuated. Consequently, the output of the filter, i.e., the oral airflow, has a spectrum exhibiting peaks and valleys rather than the relatively evenly falling source spectrum, and these determine the different segmental qualities of the sounds we hear (in this instance the difference between [u] and [i]). Finally, the radiated sound pressure has a spectrum that is tilted by approximately +6 dB per octave in comparison to the spectral slope of the oral airflow.

The illustration in Figure 11.1 is of an idealized source which is assumed to be constant for the two vowels. In real speech, the source varies dynamically in a way that reflects the configuration of the glottis, the degree and type of any laryngeal tension that may be present, the respiratory effort being used, and even the aerodynamic consequences of any supraglottal stricture. Gobl (1988) illustrates how the source may vary in the course of a single utterance spoken with a neutral, modal mode of phonation. The variation is even greater if the speaker chooses to switch between different modes of phonation (e.g., breathy voice, creaky voice, etc.) as is often done for paralinguistic signaling of emotion and attitude. Different speakers may also vary considerably in terms of the habitual type of phonation they use.

Over the years, much work has been carried out on the acoustics of the filter, which corresponds to much of the segmental differentiation of place and manner of articulation. Concerning the voice source, a good deal is known about f_0 variation, and how it varies as a function of intonation, tone, and stress. Relatively little is known about other aspects of the voice source and how it varies in speech. There are of course many studies on the intensity variation of the speech signal. Although the amplitude of the speech output to some extent reflects the amplitude of the source, one should bear in mind that the total amplitude of the speech output is a function of both source and filter.

In the next section, ways of analyzing and measuring the voice source are discussed. This is followed in section 3 by brief illustrations of how the source varies for a number of different voice qualities. In section 4 we give an overview of the factors that determine voice source variation in speech and language.

2 Analyzing the Voice Source

2.1 Obtaining glottal flow: Inverse filtering

Most experimental studies of the voice source have been based on inverse filtering. This technique is effectively a reversal of the speech production process. The speech signal is passed through a filter whose transfer function is the inverse of the supraglottal transfer function. In principle this yields the voice source in its prefiltered form, as the filtering effect of the vocal tract is canceled. Figure 11.2 illustrates this process in the frequency domain – in terms of the signal's frequency components – and in the time domain – in terms of the glottal airflow, or its derivative. Cancelation of lip radiation is not shown here for reasons that are explained below.

The inverse filter should contain a specification of the frequencies and bandwidths of the anti-resonators (complex-conjugate zeros) required to cancel the formants (complex-conjugate poles) of the vocal tract transfer function at any given instant in time. It is important to use the right number of anti-resonators for the inverse filter, appropriate to the signal bandwidth as determined by the sampling frequency. The average spacing between the formants (poles) is determined by the length of the vocal tract: for a typical male with a vocal tract of 17.5 cm we can expect one formant on average per 1,000 Hz. So, for example, given a sampling frequency of 10 kHz (i.e., signal bandwidth of 5 kHz) and a vocal tract length of 17.5 cm, the number of anti-resonators should be 5. The specification of the precise frequency and bandwidth is very critical for the lower formants, especially F1. Any error here will result in some distortion of the glottal pulse. Minor errors in the higher formants have little effect on the main pulse shape and its corresponding frequency spectrum (Gobl, 1988).

An all-pole function adequately describes the transfer function for many sounds such as vowels. For certain sounds such as nasals and laterals the vocal tract transfer function contains zeros as well as poles and in principle these zeros should be canceled by the inclusion of corresponding poles in the inverse filter. As it is often difficult to estimate the zeros of the transfer function, most researchers tend, in practice, to use an all-pole model for all sounds. Although this simplifies the

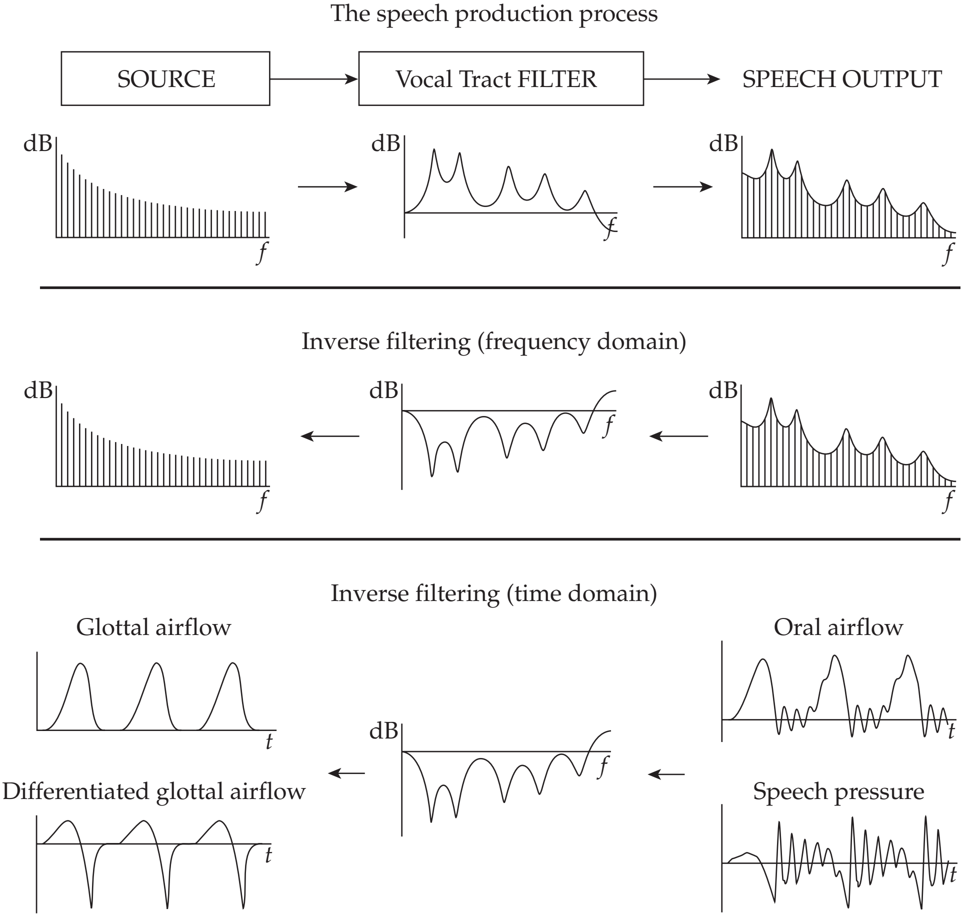

The speech production process
SOURCE
Vocal Tract FILTER
SPEECH OUTPUT
dB
f
dB
f
dB
f
Inverse filtering (frequency domain)
dB
f
dB
f
dB
f
Inverse filtering (time domain)
Glottal airflow
t
Oral airflow
t
dB
f
Differentiated glottal airflow
t
Speech pressure
t

Figure 11.2 Schematic representation of inverse filtering in the frequency and time domains.

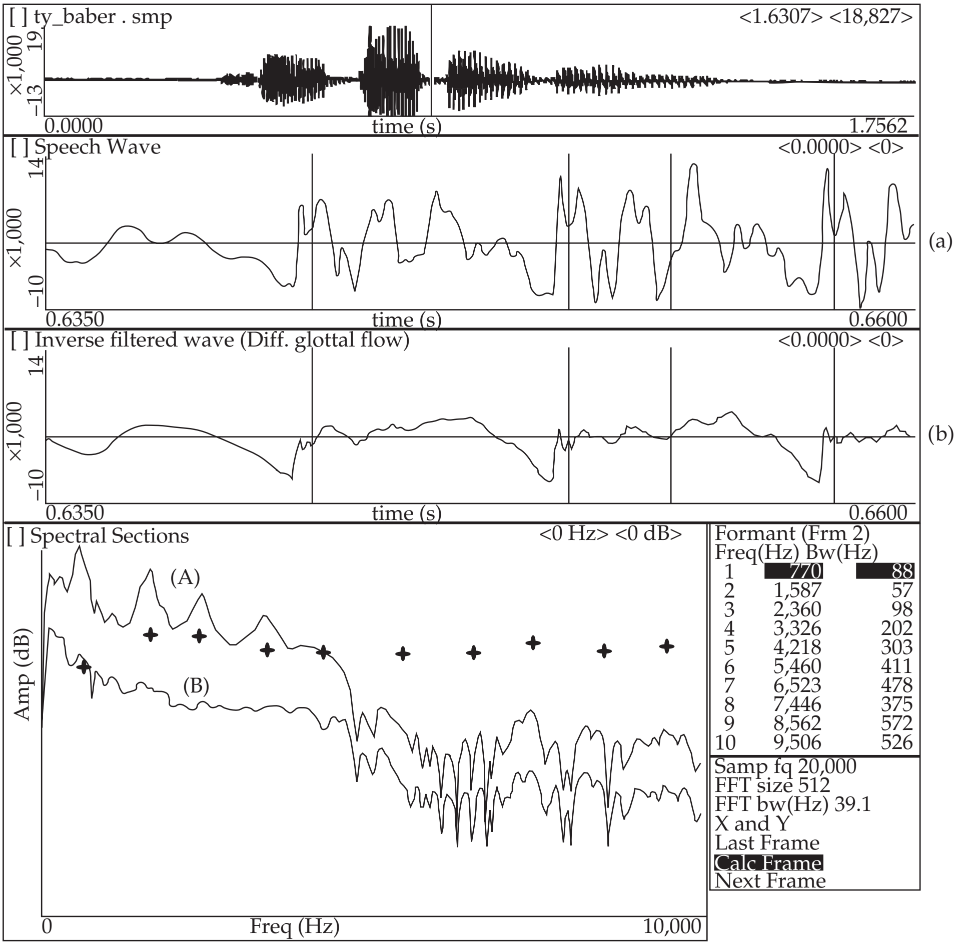
inverse filter specification, it does mean that sounds whose spectrum contains zeros are less accurately filtered. How this simplification may affect voice source estimates for nasalized speech is explored in Mahshie and Gobl (2003).

To obtain the true glottal flow from the speech pressure wave, the filtering effect of the sound radiation at the lips needs to be canceled as well. The radiation characteristics can be relatively accurately approximated by a first order differentiation (see, however, Fant, 1960, pp. 44–5, for a more detailed description). The spectral consequence of the differentiation is a relative boosting of higher frequencies by 6 dB per octave. This effect can easily be canceled by a simple integration of the signal (a real pole at zero frequency), as this is the inverse of differentiation. If the effect of the lip radiation is not canceled, the output of the inverse filter will correspond to the differentiated glottal flow, also referred to as the glottal flow derivative. Many researchers opt to work with this signal rather than the true glottal flow. The emphasis of higher frequencies by 6 dB per octave permits a more precise modeling of the spectral slope of the source signal. The strength of the glottal excitation as determined by the maximum flow discontinuity at glottal closure is also more easily obtained (see further section 2.4). It is also convenient for resynthesis purposes to combine the lip radiation with the source: one does not need first to remove it and then reintroduce it.

Inverse filtering based on the speech pressure waveform can yield detailed temporal and spectral information. However, the recording equipment and room are critical, and shortcomings in either condition can lead to artifacts in the measurements (see, for example, discussion and comments in Ladefoged et al., 1987). Ideally, an anechoic chamber should be used, although a well-damped "semi-anechoic" studio is generally also adequate. The recording equipment must preserve the phase characteristics of the signal even at very low frequencies. To achieve this, the choice of microphone and amplifiers is critical. High-quality digital or FM recordings are recommended, avoiding any built-in filters, data compression techniques, etc. which may introduce phase distortion. Analogue tape recorders distort the phase: however, suggestions have been made as to how this might be compensated for (Holmes, 1975; Hunt, 1978; Ljungqvist and Fujisaki, 1985; Hedelin, 1986). Recordings generally require some high-pass filtering to remove the inevitable intrusion of some inaudible low-frequency pressure fluctuations in the recording room. It is essential to ensure that this filtering is done using a filter with linear phase response.

Inverse filtering can also be carried out on recordings of oral airflow. In this case, a special airflow mask with a built-in differential pressure transducer is used: many studies have employed the circumferentially vented pneumotachograph mask designed by Martin Rothenberg (Rothenberg, 1973). When oral airflow is inverse filtered, the output is an estimate of the true glottal flow. If the differentiated glottal flow is required, a first order differentiator (a real zero at zero frequency) is added to the inverse filter. The main advantage of using oral airflow recordings is that absolute values of the airflow rate can be measured, which is not possible from recordings of the speech pressure wave. This is particularly useful for measuring the "DC-leakage" during phonation where the glottal cycle lacks complete closure during the so-called closed phase. The main disadvantage with this approach arises out of the limited frequency response of the mask. Even with the specially designed Rothenberg mask, the frequency response is limited to slightly over 1 kHz (see Badin et al., 1990; Hertegård & Gauffin, 1992). As a consequence, it does not provide for detailed spectral analysis of the source.

For a successful source analysis, it is of course essential that the estimate of the vocal tract transfer function be accurate. Many of the systems proposed for estimating the inverse filter involve fully automatic procedures, typically based on linear predictive coding (LPC) in one form or another. Unfortunately, they often do not yield satisfactory results for detailed source analysis, particularly where the vocal tract filter is undergoing rapid change or where the source involves a nonmodal mode of phonation. At present, the most accurate source



Formant (Frm 2)

	Freq(Hz)	Bw(Hz)
1	770	88
2	1,587	57
3	2,360	98
4	3,326	202
5	4,218	303
6	5,460	411
7	6,523	478
8	7,446	375
9	8,562	572
10	9,506	526

Samp fq 20,000
FFT size 512
FFT bw(Hz) 39.1

Figure 11.3 (Modified) Screen display illustrating an interactive inverse filtering method.

signal is obtained by using a method where the user interactively fine-tunes the formant frequencies and bandwidths of the inverse filter. Figure 11.3 is a slightly modified screen display illustrating the time and frequency domain information which guides the user in canceling the formant peaks (in the frequency domain) and corresponding formant oscillations (in the time domain). The upper window shows the speech waveform, with a cursor marking the pulse under analysis. The second and third windows show (a) the speech waveform and (b) the inverse filter output (the differentiated glottal flow) for this pulse. The lowest window shows the corresponding spectra for (A) the speech waveform and (B) the differentiated glottal flow. The points marked as crosses in the lowest window indicate the formants, determining the complex zeros of the inverse filter. Each of these points can be moved in a horizontal or vertical direction to manipulate the frequency and bandwidth respectively. With each manipulation, the screen is instantaneously updated to show the new inverse filtered waveform (b) and its spectrum (B). For further details on this particular implementation, see Gobl and Ní Chasaide (1999a).

At present no automatic procedure can achieve the level of accuracy that the trained researcher can. Using the combined time and frequency information, many aspects of the source can be measured more accurately than would otherwise be possible. Yet there are also disadvantages with this method. In fine tuning the filter, it is sometimes necessary to compromise between the time and frequency information, and here it is vital that a consistent approach be adopted. This of course demands considerable skill and experience, and entails a risk that different experimenters will adopt different strategies leading to inconsistent results. Even with a trained user, there is some risk of circularity with this procedure. As the experimenter has certain expectations of what the glottal flow should look like, it could lead to an avoidance of unlikely-looking but valid pulse shapes. But probably the greatest problem of all is that the manual interactive method is not suited to the analysis of large amounts of data. As the analysis typically proceeds on a pulse-by-pulse basis, it is very time-consuming. This, and the high degree of vigilance needed, has resulted in these types of studies being limited to small amounts of carefully analyzed data.

2.2 *Voice source models*

As stated above, the output signal of the inverse filter is an estimate of the glottal airflow or its derivative. Visual inspection may yield a first gross impression of some characteristics, whether the voicing is efficient, breathy, etc., but for fine comparisons precise measurements are required. In several studies these measurements were carried out directly from the estimated source signal (e.g., Huffman 1987; Hertegård & Gauffin, 1991; Holmberg et al., 1988; Löfqvist & McGowan, 1991; Laukkanen et al., 1997; Alku et al., 2002). An alternative method involves matching a parametric source model to the pulses obtained from the inverse filtering, and deriving the measurements from the modeled waveform.

For this approach to be successful, it is important that the model be a good representation of the true source and that it be flexible enough to capture the important variations that may occur. Traditionally, the voice source in parametric speech synthesis was implemented as a low-pass filtered impulse train (e.g., Liljencrants, 1969; Klatt, 1980). The only control parameters of this simple type of voice source are f_0 and the amplitude of the impulse. Its main drawback is that the spectral slope cannot be controlled: it is always perfectly regular, falling off monotonically at 12 dB per octave. Another drawback is that the phase characteristics of the filtered impulse are very different from that of the typical glottal pulse. The impulse response of the low-pass filter is time-reversed in comparison to the typical glottal waveform. This means that the main discontinuity of the waveform (corresponding to the main excitation) occurs at the rising branch rather than at the falling branch of the glottal pulse.

These drawbacks resulted in an inflexible and often unsatisfactory voice quality in this type of speech synthesizer, and prompted the development of more elaborate voice source models. These models all have a larger number of control parameters and a more accurate representation of the glottal waveform. They are therefore more capable of capturing the frequency characteristics (e.g., the spectral slope and the phase characteristics) of the natural glottal waveform.

2.2.1 The Liljencrants-Fant (LF) model A model which has gained popularity, and which is used below for a number of illustrations, is the LF voice source model presented in Fant et al. (1985). In addition to f_0, this model has four parameters to control the shape of the glottal pulse.

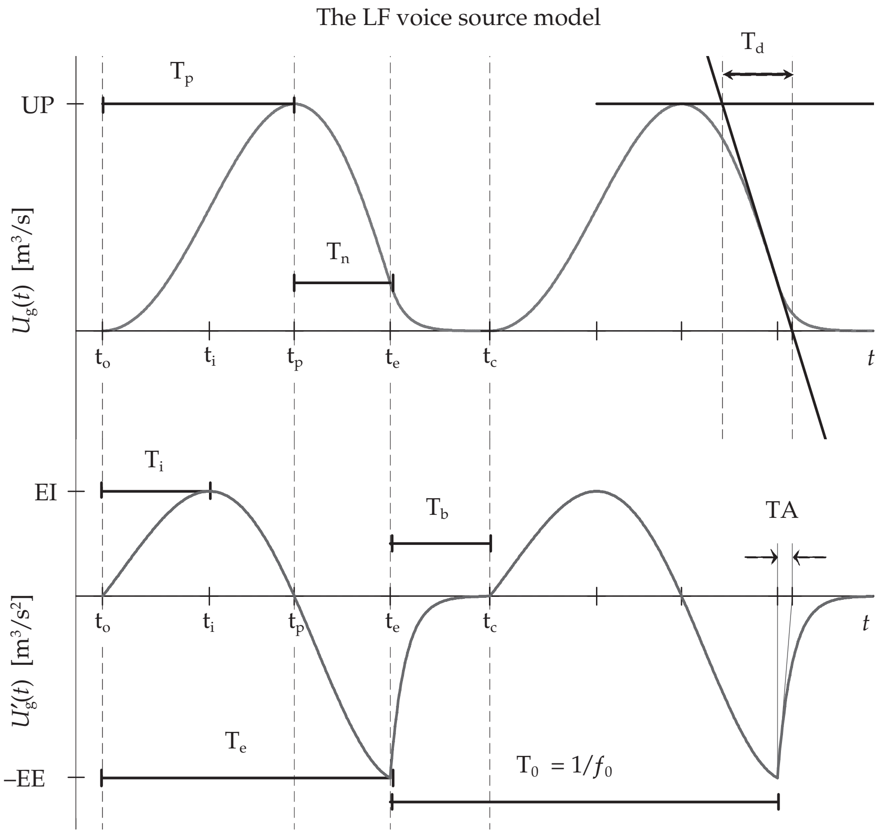
The model is made up of two segments, as determined by the mathematical expressions shown in Figure 11.4. Note that these expressions define the pulse shape in terms of differentiated glottal flow, $U_g'(t)$. However, by integrating the basic LF model expressions, we can also generate LF model pulses which relate directly to glottal airflow, $U_g(t)$. The lower waveform of Figure 11.4 shows examples of two LF model pulses of differentiated glottal flow, whereas the upper waveform shows the corresponding glottal flow pulses.

The first segment of the model is a sinusoidal function that increases exponentially in amplitude from the time point of glottal opening t_o, to the time point of main excitation t_e. Three parameters determine the shape of this segment: (1) $\omega_g = 2\pi F_g$ where F_g is the frequency of the sine function, (2) α which determines the rate of the amplitude increase and (3) E_0 which is a scale factor.

The second segment is an exponential function which is used to model the flow from the time point of the main excitation, t_e, to the time point of glottal closure, t_c. This part of the glottal cycle is termed the *return phase* and determines the residual airflow, or “dynamic leakage,” after the main excitation when the vocal folds close. In the LF model, the control parameter of the return phase is TA. TA is a measure of the “effective duration” of the return phase, and is determined by the projection on the time axis of the tangent at time t_e (see Figure 11.4).

The description of the LF model assumes that $t_c = t_o$, i.e., the time point of glottal closure is the same as the time point of glottal opening for the forthcoming pulse period. This implies that the model lacks a closed phase. In practice, for reasonably small TA values, the exponential curve will fit closely to the zero line, providing to all extents and purposes a closed phase. This saves one parameter without any significant loss in flexibility. Furthermore, in order to unambiguously determine the pulse shape, the four LF parameters E_0, α, ω_g and TA are complemented by a requirement of “area balance,” which means that the positive area of the LF pulse (from t_o to t_p), should equal the negative area (from t_p to t_c). In terms of the true glottal flow (i.e., the integrated LF pulse) this means that the base-line of consecutive pulses is kept constant. (For more detailed descriptions of this model, see Fant et al. 1985; Fant, 1995; and Gobl 2003.)

2.2.2 Other voice source models Numerous other parametric voice source models have been proposed in the literature (e.g., Rosenberg, 1971; Rothenberg et al., 1974;



$$U_g'(t) = E_0\, e^{\alpha t} \sin \omega_g t, \qquad t_o \le t < t_e$$

$$U_g'(t) = \frac{-EE}{\varepsilon T_a}\left(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon T_b}\right), \qquad t_e \le t < t_c$$

Figure 11.4 The LF voice source model.

Fant 1979a, 1979b, 1982; Ananthapadmanabha, 1984; Hedelin, 1984; Fant et al., 1985; Ljungqvist & Fujisaki, 1985; Price, 1989; Klatt & Klatt, 1990; Qi and Bi, 1994; Veldhuis, 1998). These models can be divided into two groups on the basis of whether they model the true glottal flow pulse or the differentiated glottal pulse. They also differ in the number of parameters and in the functions they use to generate the glottal pulse. Another important difference among them concerns whether or not they include a segment to model the return phase of the glottal pulse.

Fairly detailed comparisons of some of these voice source models can be found in Ananthapadmanabha (1984) and in Ljungqvist and Fujisaki (1985). Figure 11.5 summarizes some of the important features of seven different models.

	Model	Single flow derivative discontinuity	Provision for multiple flow derivative discontinuities	Provision for continuous flow derivative	Waveform realization
Amplitude, width, and skewing of the glottal flow	(a)	yes	(yes)*	no	sinusoidal
	(b)	yes	no	no	sinusoidal
Independent control of flow derivative discontinuity	(c)	yes	no	yes	sinusoidal
Modeling of activity in the glottal closed phase	(d)	yes	no	yes	
	(e)	yes	no	yes	sin+polyn.
	(f)	yes	no	yes	exp.· sin.
	(g)	yes	yes	yes	polynomial

* Rosenberg proposed several models, some of which allow multiple discontinuities.

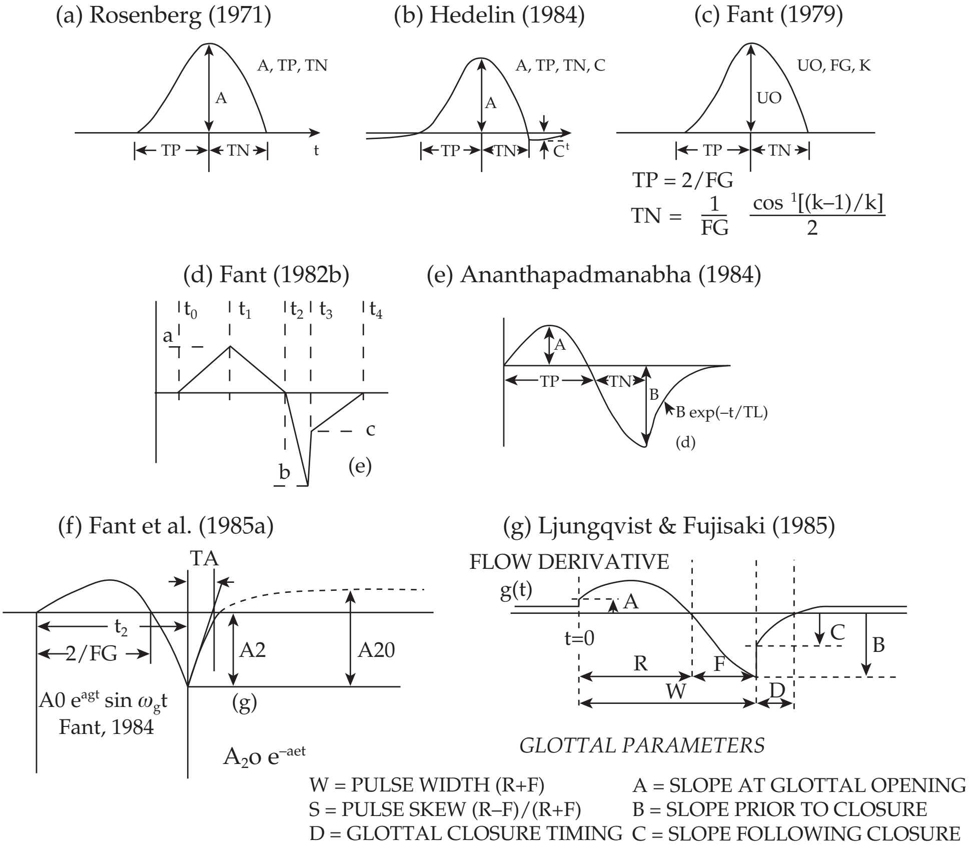


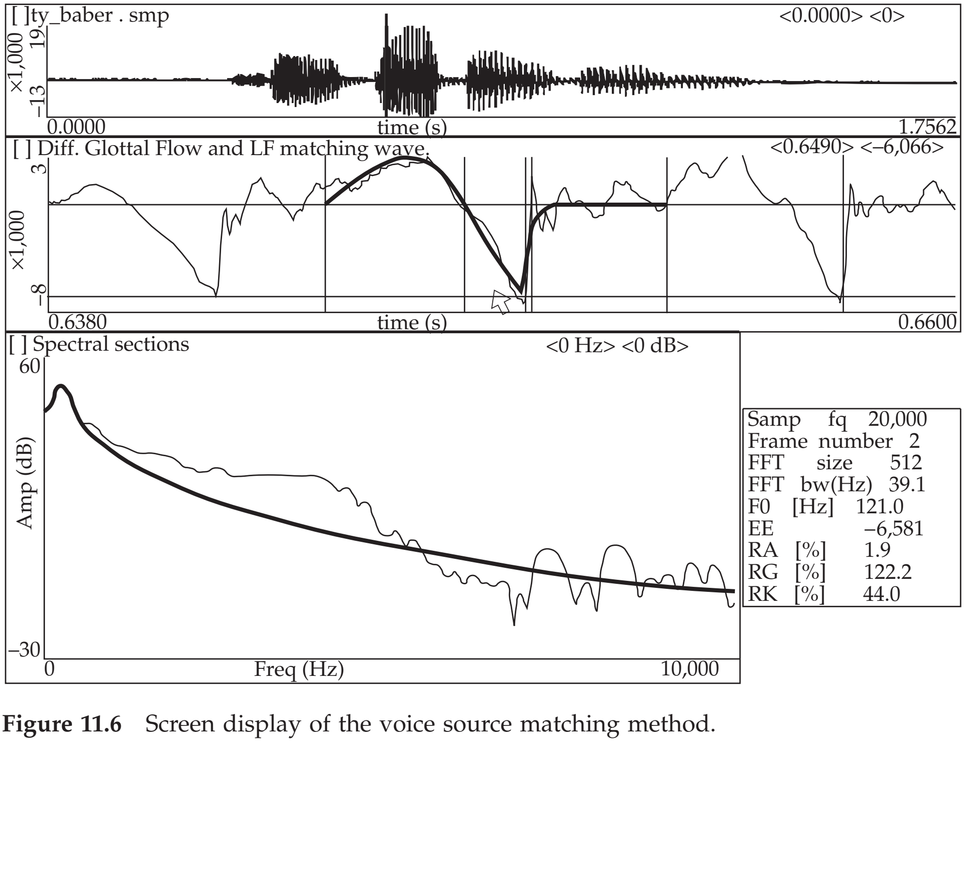
Figure 11.5 Waveforms and equations for seven proposed voice source models. (After Ananthapadmanabha, 1984; and Ljungqvist & Fujisaki, 1985)

2.3 Measuring the glottal signal: Source model matching

As mentioned earlier, a method of extracting source measurements involves matching a voice source model to the inverse filter output, and deriving the measurements from the modeled waveform. This procedure has certain advantages over measuring parameters directly from time and amplitude points of the inverse filter output. First of all, the model matching allows us to take both time and frequency domain information into account, as the spectrum of the model can be calculated. This is particularly useful for capturing features which have important spectral consequences, but which are difficult to measure accurately directly from the waveform (e.g., the return phase). A further advantage is that the modeled source signal can be quickly implemented in synthesis, and in principle this should facilitate perceptual testing of the various parameters measured.

As with inverse filtering, the matching of the model can be done automatically (e.g., Ananthapadmanabha, 1984; Chan and Brookes, 1989; Strik and Boves, 1994; Fröhlich et al., 2001), but present automatic algorithms do not always yield reliable results. Again, more accurate measurements are obtained if a manual interactive approach is adopted as can be illustrated in relation to Figure 11.6, the screen display which guides the user in the matching process.

The mid panel of this figure shows the inverse filtered waveform (differentiated glottal flow) for the pulse specified in the top panel. Superimposed on this pulse,



Samp	fq	20,000
Frame	number	2
FFT	size	512
FFT	bw(Hz)	39.1
F0	[Hz]	121.0
EE		−6,581
RA	[%]	1.9
RG	[%]	122.2
RK	[%]	44.0

Figure 11.6 Screen display of the voice source matching method.

one can also see a matched LF pulse (thick line), whose contour is determined by four time points (vertical lines) and one amplitude point (horizontal line), which are manually set by the experimenter. The four time points are: (1) the time of glottal opening, t_o; (2) the time of peak glottal flow, t_p; (3) the time of the excitation, t_e; (4) the time point on the basis of which the return phase is estimated, t_r (equals t_e + TA). The amplitude point (5) is the amplitude of the excitation, EE. The spectrum corresponding to the inverse filtered pulse is shown in the bottom panel and superimposed on it is the spectrum of the LF model pulse (thick line). The model pulse is optimized by making fine adjustments to the time and amplitude points in order to find the best overall agreement in both the time and frequency domains.

2.4 Some important voice source parameters

The LF parameters outlined in section 2.2 determine the overall shape of the glottal pulse. For our analysis, we need to measure very specific aspects of this waveform, i.e. those aspects that are thought to be acoustically and perceptually important, and which can be more readily related to the underlying physiological events. Once the matching procedure has been satisfactorily completed, these source parameters can be calculated. We outline some of the most important parameters here, illustrating in Figure 11.7 how changes in the glottal waveform affect the acoustic spectrum. One must remember however that it is difficult to give a very precise specification of the spectral consequences of the individual source parameters, as they frequently interact in complex ways. One should also remember that the precise definition to a certain extent depends on the model used. Here we define the parameters in terms of the LF model, which was used for a number of illustrations later in this chapter.

Fundamental frequency, f_0 The fundamental frequency = $1/T_0$, where T_0 is the fundamental period, which is the duration of the glottal cycle as defined by the time between the main excitation of two consecutive glottal pulses.

Excitation strength, EE The EE value is closely related to the overall strength of the glottal excitation. It is defined as the negative amplitude of the main excitation, which occurs at the time-point of maximum discontinuity of the glottal flow signal. This time-point (t_e) normally coincides with the maximum slope of the falling branch of the glottal pulse, i.e., EE is typically given by the most negative value of the differentiated glottal flow. At the production level it is determined by the speed of closure of the vocal folds and by the airflow through them. At the acoustic level it corresponds closely to the overall intensity of the source signal. A change in EE essentially causes a corresponding amplitude change in all source components, with the exception of the very lowest components, particularly the first harmonic. Thus, this parameter is the one that most resembles the amplitude parameter of the simple impulse source. However, the amplitude of the lowest components is largely determined by the pulse shape, and therefore varies less with changes in EE.

Dynamic leakage, RA The dynamic leakage is the residual flow during the return phase, which occurs from the time of the excitation to the time of complete closure (or maximum closure if there is a DC leakage). In terms of the true

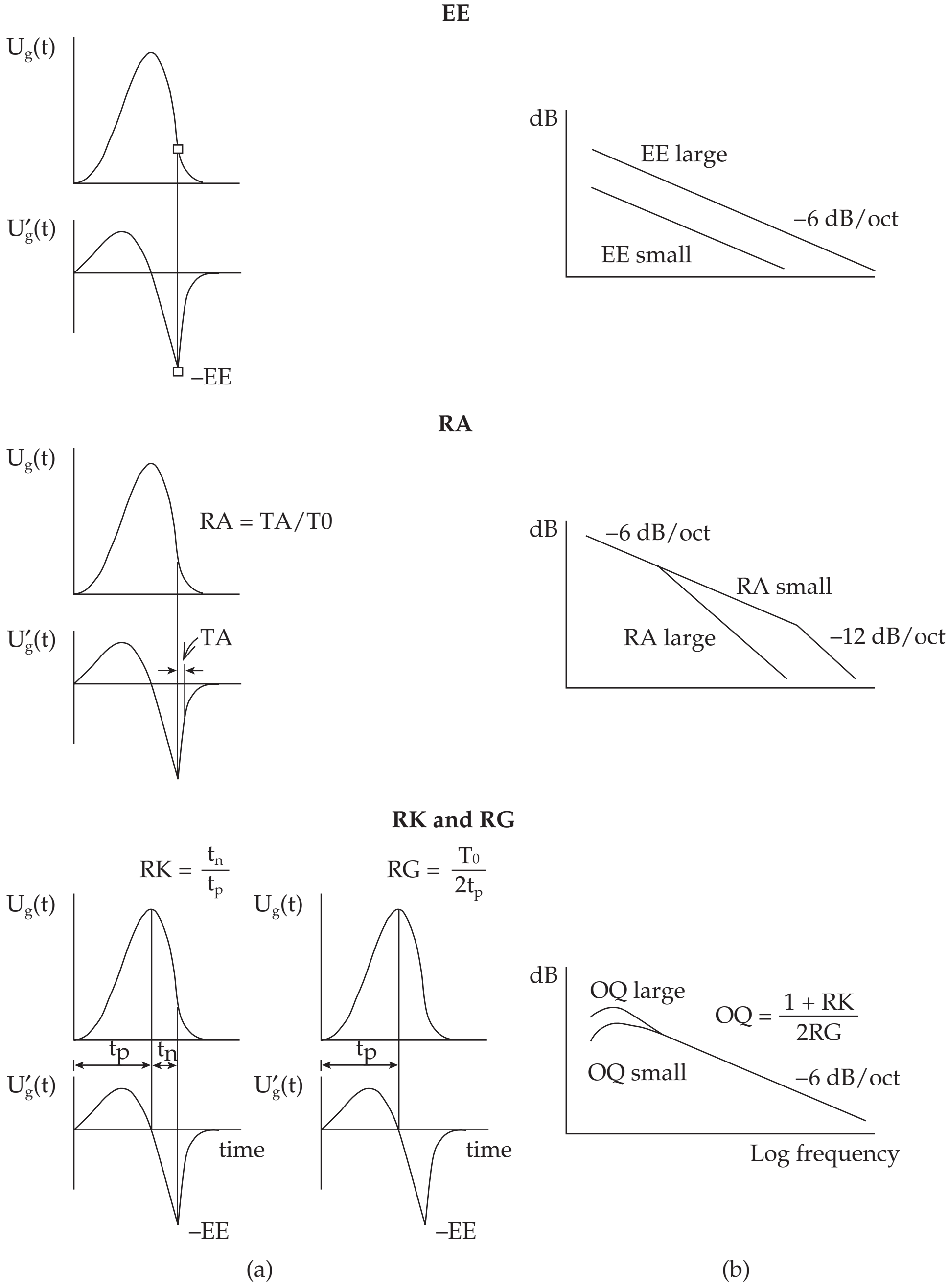


Figure 11.7 The voice source parameters EE, RA, RK, and RG, in terms of the true and differentiated glottal flow, showing how changes in these parameters affect the acoustic spectrum (for explanations, see text).

glottal flow, the return phase shows up as a "rounding of the corner" of the closing branch of the pulse. The RA value is a measure of the effective duration of the return phase, TA, normalized to the fundamental period, T_0. Thus, RA is equal to TA/T_0, and in terms of the LF model, TA is approximately the time constant of the exponential function modeling the return phase (see Figure 11.4). At the production level, RA relates to the sharpness of the glottal closure, that is, to whether the vocal folds make contact in an instantaneous way or in a more gradual fashion along their entire length and depth. Differences in dynamic leakage are important acoustically because they affect the slope of the source spectrum. The frequency characteristics of the exponential function of the return phase are approximately those of a first order low-pass filter. The cutoff frequency, FA, is inversely proportional to TA: $\mathrm{FA} = 1/(2\pi\mathrm{TA}) = f_0/(2\pi\mathrm{RA})$, i.e., the cutoff frequency of the filter is inversely correlated with the amount of dynamic leakage.

Open quotient, OQ The open quotient is the proportion of the glottal cycle for which the glottis is open. In terms of the source spectrum, it mainly affects the amplitude of the lower components. In particular, there is a close correspondence between the OQ value and the amplitude of the first harmonic: note however that the degree of correspondence varies in a way that depends on the values of the parameters RG and RK (see below). A related parameter which tends to covary with the open quotient is UP, the peak volume velocity of the glottal pulse (labeled as A, A, and U0 respectively in the first three glottal models of Figure 11.5).

Glottal frequency, FG The glottal frequency is a characteristic frequency of the glottal pulse shape during the open phase (Fant, 1979a), and in terms of the LF model, $\mathrm{FG} = \omega_g/(2\pi)$ (see Figure 11.4). FG can be determined by the time period of the opening branch of the glottal pulse as $1/(2t_p)$. An alternative expression of this parameter is RG, which is FG normalized to f_0, so that $\mathrm{RG} = \mathrm{FG}/f_0$. RG tends to vary inversely with OQ and UP, and consequently, a high RG is found with attenuated levels of the lowest end of the source spectrum. For very high RG values, the glottal frequency may approach the frequency of the second harmonic (H2), and contribute to boosting its level. A high RG and a relatively stronger H2 tend to be characteristic of tense or pressed phonation. Low RG values are found where UP is high, where it contributes to boosting H1. Thus RG contributes (with OQ and UP) to the relative amplitude of H1 and H2 in the speech output, a measure frequently used in the linguistic literature (see more on this in section 2.5).

Glottal symmetry/skew, RK In comparison to the underlying glottal area function, the glottal flow pulse is typically skewed to the right, i.e., the opening phase tends to be longer than the closing phase. This skewing of the glottal pulse would appear to be due to the inertive load of the vocal tract (Fant, 1982; Rothenberg, 1983). The acoustic consequences of pulse skewing are somewhat complex. It affects mainly the lower part of the source spectrum so that a more symmetrical pulse shape has the effect of boosting the lower harmonics. However, the degree of skewing also determines the depth of the notches (weakened or missing harmonics) in the source spectrum: the more symmetrical the pulse, the deeper the spectral dips. The locations of the notches are determined by the open quotient together with the pulse shape (cf. Flanagan, 1972, pp. 236–42). Considerable attention has been given to the effect of glottal pulse skewing. However, it may be the case that the perceptual importance of the skewing has been relatively overestimated. Skewing is typically highly correlated with the excitation strength, and its perceptual contribution is easily confused with that of the excitation strength. The risk of such confusion is particularly high if a voice source model is used which lacks direct control of the excitation strength. In other words, we would suggest that the excitation strength is fundamentally a more important parameter than the skewing of the pulse.

Aspiration noise, AH In addition to the quasi-periodic sound source produced during phonation there are also varying degrees of turbulence noise produced at or near the glottis, typically referred to as aspiration noise. The modeling of this noise source is often not explicitly included in parametric voice source models. The importance of mixed excitation (periodic excitation mixed with aspiration noise) has been mentioned on several occasions (e.g., Pandit, 1957; Dolansky & Tjernlund, 1968; Fujimura, 1968; Rothenberg, 1974; Ladefoged & Antoñanzas-Barroso, 1985; Klatt, 1986; Hunt, 1987; Gobl & Ní Chasaide, 1988; Gobl, 1989), but the noise component is difficult to estimate quantitatively. However, even when not directly part of the model, a noise generator can always be used together with it to provide the noise component of the voiced excitation. Issues like the actual spectral content of the aspiration noise, strategies for controlling the level of aspiration noise, and the question of how to modulate the noise within a glottal period have been discussed by, for example, Rothenberg (1974), Rothenberg et al. (1974), Makhoul et al. (1978), Klatt and Klatt (1990), and Gobl (2006).

The pulse-to-pulse stability of source parameters is also an important factor in determining voice quality. Traditionally, measures such as jitter and shimmer have been used to quantify pulse-to-pulse variation. Jitter is the random variation in f_0 and shimmer equals fluctuations of the pulse-to-pulse amplitude. Shimmer is often measured from the speech waveform amplitude, which can lead to errors as it is to some extent influenced by source-filter interaction effects. Ideally, shimmer should be estimated directly from amplitude measures of the glottal waveform, for example, EE. High levels of jitter and shimmer have often been found to correlate with hoarse voice. Note, however, that other source parameters are also likely to exhibit instability in certain circumstances, a fact which is probably also of perceptual importance. For examples, see the illustration of a pathological voice in Figure 11.16 and of a normal creaky voice in Figure 11.9.

Gobl (1988) has shown that many of the above mentioned source parameters tend to covary. EE is highly correlated with the negative amplitude of the speech waveform and other source parameters are often correlated with EE. For example, the return phase typically varies inversely with EE, so that if the excitation is weaker, RA is higher. There is generally also covariation between RA and RK, so that a long return phase (and a low EE) corresponds to a more symmetrical pulse shape. Several of these tendencies have been corroborated by subsequent work (Pierrehumbert, 1989; Fant, 1995, 1997) but are not invariably present as indicated in Gobl and Ní Chasaide (1992) and Ní Chasaide and Gobl (2004a).

As our state of knowledge increases, it may become possible to predict many of the source parameters from a few basic ones. For example, Fant (1995, 1997) has proposed a global pulse-shape parameter R_d, which takes into account some of the natural covariation between some of the parameters. From the R_d value, default values can be predicted for parameters such as RA and RK using formulae derived from linear regression analysis (for further details see Fant, 1995; Gobl, 2003). A similar parameter has also been proposed by Alku et al. (2002), the "normalized amplitude quotient" NAQ.

R_d and NAQ are both based on the "declination time" T_d (Fant, 1979a) normalized to the fundamental period. It has been suggested (e.g., Alku et al., 2002) that these global parameters are measures that correlate well with the tense–lax dimension of voice quality differentiation. Note that T_d is derived from amplitudes measured from the glottal flow signal and its derivative (T_d = UP/EE, see Figure 11.4). Such amplitude-based parameters may be more robust for automatic processing than time-based parameters – i.e., parameters defined by specific timing events in the glottal waveform – such as OQ, RK, and RG, which are sensitive to the errors in the estimation of the inverse filter that frequently arise in automatic source analysis systems. An extended set of amplitude-based parameters has been described in Gobl and Ní Chasaide (2003a), which includes approximations of OQ, RK, FG, and RG derived from amplitude measures and f_0.

2.5 *Spectral measurements relevant to the voice source*

In the preceding section, we have concentrated on time domain measurements of the glottal source, linking these to their expected spectral consequences. Frequency domain measurements can also be carried out on the output of the inverse filter. As is probably clear from the description of source parameters above, one may need to distinguish the very lowest frequencies from higher regions in any attempt to characterize and compare source spectra. The picture can be further complicated by the appearance of spectral notches, or even additional subglottal pole/zero pairs. Specific glottal pulse shapes (very symmetrical) can give rise to notches, and might be found, for example, in breathy voice. Furthermore, the more the glottis is abducted, the greater the coupling to the subglottal system and the greater the likelihood of subglottal resonances showing up in the source spectrum.

The spectral tilt is probably the most fundamental parameter one would want to measure in the voice source spectrum. Obtaining it is not always a simple matter as was demonstrated in early studies by Jackson et al. (1985, 1986), who explored the possibility of fitting a single regression line to source spectra. One possible method for comparing source spectra which takes account of changing levels in different frequency regions is illustrated in Figure 11.12, and explained in section 3. A complicating factor with regard to spectral tilt concerns the mixed excitation of quasi-periodic and aperiodic sources. In estimating the tilt of the voice source spectrum, we would ideally want to measure the slope due to the shape of the quasi-periodic pulses, without the influence of any other sources such as the aspiration noise source. For example, voice qualities such as breathy and whispery voice typically have a relatively steep spectral slope (see further section 3), but this increase in the tilt can be obscured by the aspiration noise. If the aspiration noise source is sufficiently strong, it may cause a relative boosting of higher-frequency components in the spectrum, thereby masking the increased spectral tilt of the quasi-periodic voice source.

Spectral measurements based on the speech output signal can also be useful. For identical speech items, differing only in voice quality, average spectra (as in Figure 11.11) or even long-term average spectra can help to demonstrate source differences. A measure frequently used is the comparison of the amplitude level of the first harmonic (H1) with the level of some higher-frequency component. A comparison of H1 and F1 levels has been used in a number of studies (see, for example, Kirk et al., 1984, 1993; Ní Chasaide & Gobl, 1993; and Figure 11.10 below). Another popular measure has involved the comparison of the level of the first two harmonics (see, for example, Fischer-Jørgensen, 1967; Bickley, 1982; Maddieson and Ladefoged, 1985; Blankenship, 2002). A very dominant H1 has been widely found to be highly correlated with a breathy mode of phonation whereas a relatively strong H2 can be correlated with tense or creaky voice.

Measurements based on the speech output waveform are particularly attractive to linguists working in the field, in that they do not require the level of technical facilities which the execution of inverse filtering and model matching require. However, it is important to bear in mind that although these types of measurements reflect differences in the source spectrum, they are also sensitive to other factors and can therefore not be used to infer the actual slope of the source spectrum. It is important for the experimenter to be aware of the other factors that can affect the level of different frequencies of the output spectrum, as the speech materials must be carefully chosen to take account of them. First of all, the frequencies of the formants affect their amplitude levels, and so a comparison of, for instance, H1 and F1 levels would clearly not be appropriate across different vowel qualities. Formant levels are also partially determined by the formant bandwidths, which reflect the degree of damping present. A high degree of damping is found where there is little or no closed phase in the glottal pulse, as, for example, in breathy voice. Supraglottal factors also affect the degree of damping, and thus the formant bandwidths. In any case, formant bandwidths affect the levels of the output spectrum in a way that does not directly reflect the levels of the source spectrum.

All of these spectral measures are also sensitive to f_0 differences or, more precisely, to any shift in the ratio of f_0 to F1 frequencies. For example, the comparison of H1 and H2 levels may be a valid measure when $F1$ is high and f_0 low. However, when $F1$ is low or f_0 is high (or in the worst case where both of these factors pertain), the levels of H1 or H2 may be boosted depending on their proximity to the F1 peak. In such cases the relative levels of H1 and H2 are influenced by filter as well as by source factors, and so are no longer reliable indicators of the mode of phonation.

By compensating for the main influence of the vocal tract, estimates can be obtained which are less sensitive to differences in vowel quality. For instance, Stevens and Hanson (1995) have proposed measures from the speech output spectrum referred to as $H1^* - H2^*$ and $H1^* - A3^*$. $H1^*$ and $H2^*$ are adjusted measures of the amplitudes of $H1$ and $H2$, which take into account the main effect of the F1 transfer function on the amplitude of these harmonics. $A3^*$ is an adjusted measure of A_3, the amplitude level of the third formant, in which the effect of F1 and F2 variation has been compensated for. In other words, these measures involve a simplified inverse filtering process, where only the main effects of the vocal tract filter are canceled. Hanson (1995, 1997) has used these measures for analyzing the glottal source characteristics of female speakers. More recently, Iseli and Alwan (2004) have presented a more comprehensive formulation for correction of harmonic amplitude levels, in order to further reduce the effects of the vocal tract filter on these measures.

3 Some Commonly Occurring Voice Qualities

As a backdrop to section 4 we present here a brief sketch of a few commonly occurring voice qualities. The aim is not only to show how these voice qualities may differ acoustically, but also to illustrate different kinds of measurements that are useful. The voice qualities we deal with are modal voice, breathy voice, whispery voice, creaky voice, tense voice, and lax voice, as described in Laver (1980, 1994). Note that although we are concerned here only with the laryngeal aspects of voice quality, the last two mentioned may involve greater or lesser degrees of tension in the entire speech apparatus, and not purely of the phonatory system.

The physiological descriptions here are in terms of three hypothesized parameters of muscular tension; adductive tension, medial compression, and longitudinal tension (see illustration in Figure 11.8 from Laver, 1980). These determine

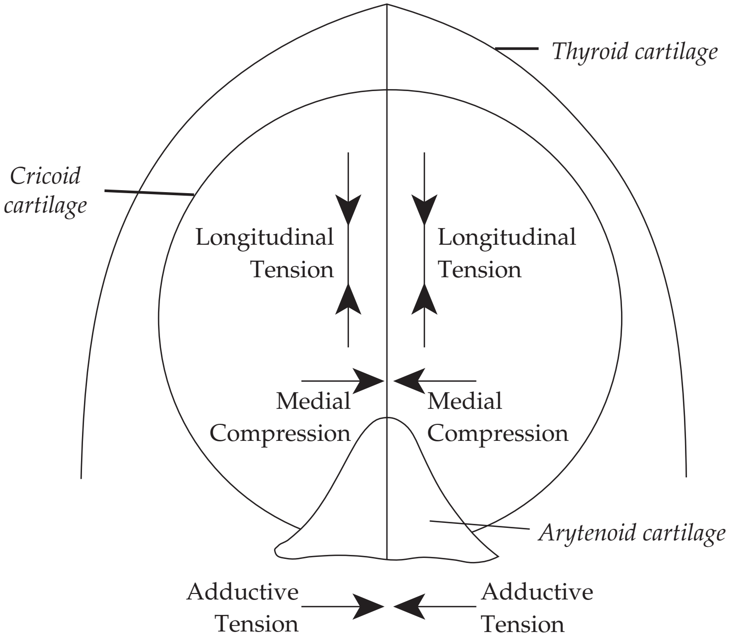


Figure 11.8 Three laryngeal parameters of muscular tension as described in Laver (1980).

the configuration and tension settings of the vocal folds, and interact with aerodynamic factors related to subglottal pressure and glottal airflow to yield a variety of voice qualities. For a fuller description the reader is referred to that text. See also the descriptions of voice quality in Catford (1964) and Ladefoged (1971). *Adductive tension* is defined as the force by which the arytenoids are drawn together, so that the cartilaginous glottis is adducted. It is controlled by the interarytenoid muscles. *Medial compression* is defined as the force by which the ligamental glottis is closed, through the approximation of the vocal processes of the arytenoids. It is primarily controlled by the lateral cricoarytenoid muscle, but the external thyroarytenoid muscle can also be involved. *Longitudinal tension* is the tension of the vocal folds, and is mediated primarily by contraction of the vocalis and of the cricothyroid muscles, whose main function is to control pitch.

Some of the acoustic characteristics of these voice qualities are illustrated in Figures 11.9–12, and were derived using the analysis techniques outlined in section 2. The speech materials were produced by a male phonetician, well acquainted with the Laver system. Figure 11.9 shows source parameter values; Figures 11.10 and 11.11 show spectral measures of the speech output signal for the

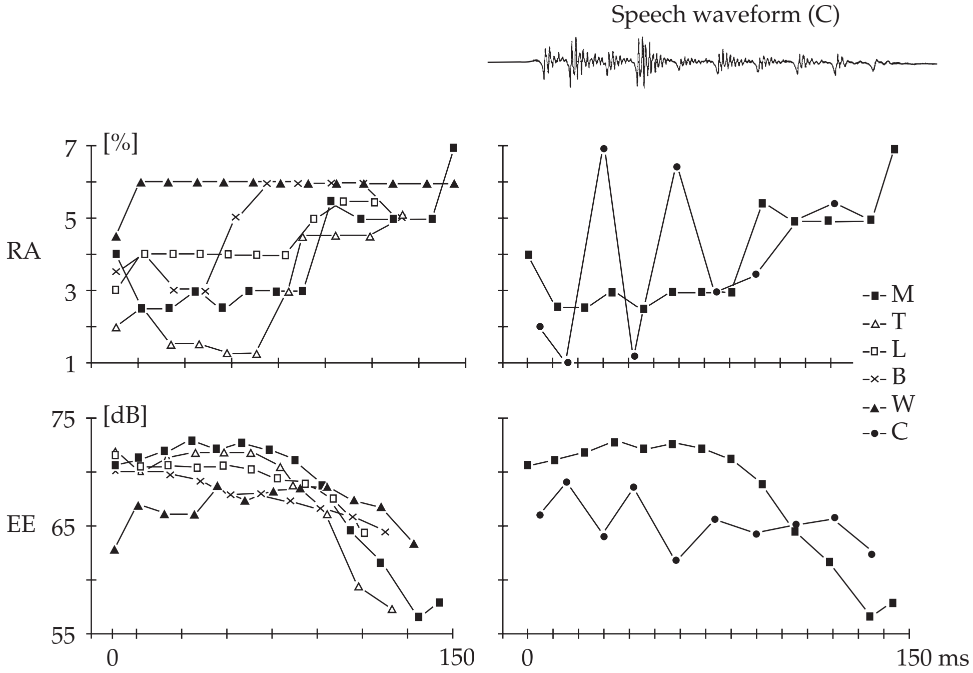


Figure 11.9 Pulse-by-pulse values for RA and EE for the voiced interval of /straiks/. Left panel shows modal (M), tense (T), lax (L), breathy (B), and whispery (W) voice. Right panel shows modal and creaky (C) voice, and additionally shows the speech waveform in the latter.

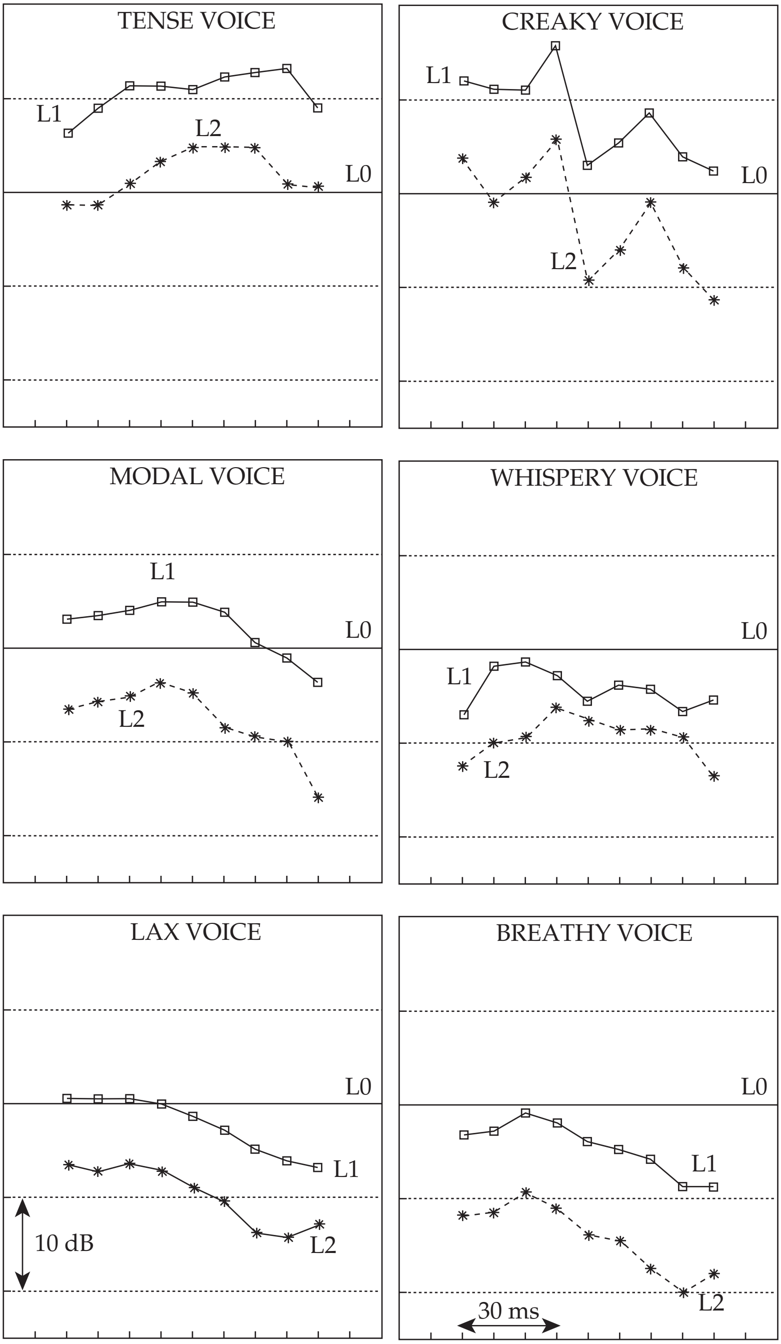


Figure 11.10 F1 and F2 levels relative to the level of H1 (L_1–L_0 and L_2–L_0) are shown for tense, modal, lax, breathy, whispery, and creaky voice qualities for a 90 ms interval in /straiks/.

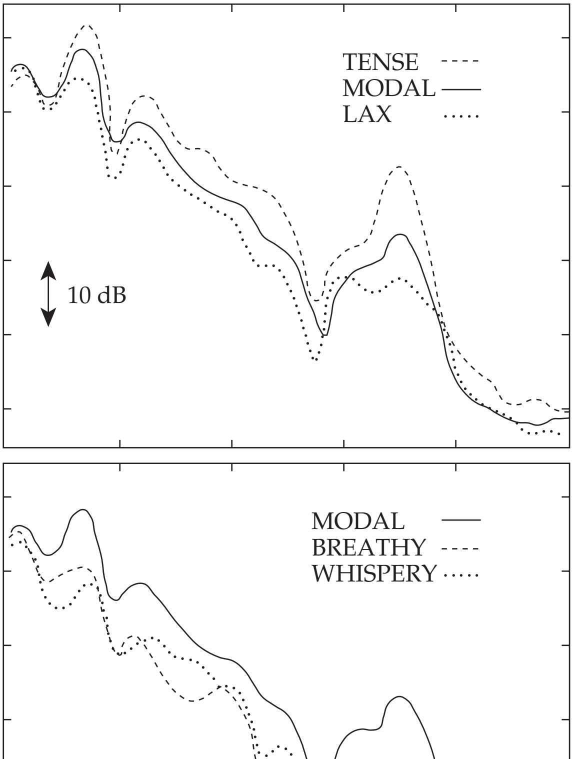


Figure 11.11 Average spectra for the voiced portion of /straiks/ shown for tense, modal, lax, breathy, whispery and creaky voice qualities.

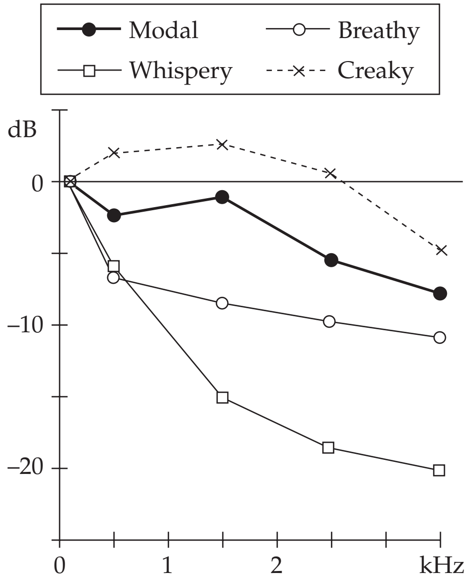


Figure 11.12 Schematic source spectra for four voice qualities in *babber* showing within four frequency bands the average deviation from a constant –12 dB/octave slope.

vocalic interval of the relatively unstressed word *strikes*. For four of the qualities, a schematic representation of the source spectra is shown in Figure 11.12, measured at the mid point of the stressed /a/ in the nonsense word *babber*. The aim here was to facilitate comparison of spectral slopes by showing the extent to which the slope for each quality deviates from the "ideal" source (i.e., –6 dB per octave for the differentiated glottal flow). To achieve this, the spectra were "flattened" by adding 6 dB per octave relative to the amplitude level of the first harmonic (L_0). The source spectrum was then divided into four frequency bands: 0–1, 1–2, 2–3, 3–4 kHz. For the vowel in question there is one formant in each band. Harmonics above 4 kHz were not measured. The mean value was then calculated of the normalized (linear) amplitudes of all harmonics within a frequency band, and plotted relative to L_0. This value represents the deviation from the "ideal" source slope, indicated by the horizontal line at 0 dB in Figure 11.12. For further details, see Gobl (1989) and Gobl and Ní Chasaide (1992).

Modal voice is the neutral mode of phonation to which other voice qualities are compared, and "which phonetic theory assumes takes place in ordinary voicing, when no specific feature is explicitly changed or added" (Laver, 1980, p. 95). For this quality, adductive tension, medial compression, and longitudinal tension are thought to be moderate. Both the ligamental and the cartilaginous part of the glottis vibrate as a single unit. The vocal fold vibration is further described by Laver as regularly periodic and efficient, with full glottal closure and thus without audible aspiration noise. Some recent studies have, however, shown that incomplete glottal closure may be very common even in what is perceived as modal voice (see, for instance, Södersten, 1994) and particularly in female speech.

The slope of the source spectrum for modal voice in Figure 11.12 is somewhat greater than the "ideal" case. Nevertheless, it is in relative terms a fairly efficient mode of phonation. It is important to remember that within utterances spoken with modal, or indeed any voice quality, there may be considerable dynamic variation of the source (see, for example, Gobl, 1988). In certain environments there may be considerable convergence of modal and breathy/whispery voice, as can be seen in the few periods preceding the voiceless consonant /k/ in Figure 11.9. This is a contextual effect which appears to affect all the voice qualities looked at, and is discussed in detail in Ní Chasaide and Gobl (1993).

Breathy voice is thought to involve minimal adductive tension, weak medial compression, and low longitudinal tension. The vocal folds vibrate very inefficiently and they never come fully together. Thus, there is a considerable constant glottal leakage with some audible aspiration noise. The high dynamic leakage of this voice quality is evidenced by high RA values. Consequently, FA is much lower than for modal voice, particularly in the stressed vowel of *babber*, where we find a value of 500 Hz for breathy voice compared to 1,500 Hz for modal. The glottal pulse is also more symmetrical (high RK) for breathy voice, and has a high open quotient, OQ. Together, these suggest a high rate of airflow through the vocal folds, as would be expected for this voice quality, and this is indeed what our calculated UP (peak glottal flow) values show. This would yield a relative boosting of the lowest harmonic, a spectral feature which has been widely reported for this voice quality. The consequent sharp slope in the lower end of the source spectrum can be seen clearly in Figure 11.12. In the speech output signal, we see in Figure 11.10 the clear dominance of H1 for breathy voice as compared to the dominance of F1 in modal voice. This particular spectral measure captures not only the relative boosting of L_0, but also the high degree of damping of F1, which would be expected for the breathy voice quality, where the vocal folds are abducted. These effects can also be observed in the average spectra of Figure 11.11.

Whispery voice is characterized by low adductive tension, moderate to high medial compression, and moderate longitudinal tension. As a consequence, there is a triangular opening of the cartilaginous glottis, whose size varies with the degree of medial compression. In weak whisper the medial compression is moderate and the opening may include a part of the ligamental glottis as well as the cartilaginous. Whisper with increasingly higher intensity has increasingly higher medial compression and smaller glottal opening, until only the cartilaginous glottis is open. Laryngeal vibration is assumed to be confined to that portion of the ligamental glottis which is adducted, and the whispery component to the triangular opening between the arytenoids. It is very inefficient and there is a considerable degree of audible aspiration noise.

As pointed out by Laver (1980, pp. 133–4), whispery and breathy voice form an auditory continuum with no clear borderline between them. In auditory terms they would be distinguished in terms of the relative dominance of the periodic and noise components: in breathy voice, the periodic component is dominant, whereas in whispery voice the noise component would be relatively greater (see further comments on these two voice qualities in section 4).

The source measurements for whispery voice are fairly similar to those of breathy voice, being in some cases more extreme deviations from modal values. Whispery voice differs mainly from breathy voice in having a lower RK and OQ, showing a more skewed glottal flow pulse, with a relatively shorter closing branch. The calculated UP values were also noticeably less than for breathy voice. As UP is highly correlated with the level of H1 (L_0), we find that there is less boosting of the fundamental than for breathy voice. Note that this difference does not show up in the source spectra of Figure 11.12, as these have been normalized to L_0. In terms of the entire source spectrum, however, the fundamental component remains very dominant, as the source spectrum has an even greater slope than does breathy voice. Probably for this reason, the measurements of L_1 and L_2 relative to L_0 in the output signal (Figure 11.10) do not look very different to those of breathy voice. Of course, bandwidth differences also affect the levels of the output spectrum.

Creaky voice is thought to involve high adductive tension and medial compression, but little longitudinal tension. Pitch has been observed to be extremely low, and would appear to be controlled by aerodynamic factors and not by varying the longitudinal tension, as in the other voice qualities. The f_0 and amplitude of consecutive glottal pulses is further known to be very irregular. Because of the high adductive tension, only the ligamental part of the vocal folds is vibrating. The folds are relatively thick and compressed, and the ventricular folds may also be somewhat adducted, so that their inferior surfaces come in contact with the superior surfaces of the true vocal folds. This would thus create an even thicker vibrating structure. The mean airflow rate has been observed to be very low.

Although every voice quality varies dynamically in the course of an utterance, creaky voice is particularly variable. Creakiness, in the sense of irregularity of successive glottal pulses, appears intermittently. It did not show up in the stressed word *babber*, but did in the relatively unstressed word *strikes* (see right hand panel in Figure 11.9 where the speech waveform for creaky voice is also shown). Here, there is an alternation of two very different types of glottal pulse. One has a reasonably high EE – which is nevertheless still lower than for modal or lax voice – a very low RA suggesting a fairly instantaneous glottal closure and strong high-frequency components. The other type of pulse shows rather opposite tendencies: EE is very low and RA is high, and consequently the spectrum of these pulses would be rather different. Both types of pulse are characterized by a low OQ, a low RK, and a relatively high RG, values being more extreme for the first type of pulse. In the stressed word *babber*, source values were not unlike those of the strong pulse described above, but differed mainly in that EE was considerably higher. The short open phase found generally for this voice quality correlates well with the known low airflow rate observed for creaky voice, and would have the consequence of reducing the levels of the lower harmonics relative to the rest of the spectrum. This effect can be seen clearly in the schematic source spectrum in Figure 11.12 and in the average spectra of the speech output signal in Figure 11.11. Similarly in Figure 11.10, we find a very dominant F1 relative to the H1 level.

Note that the pulse-to-pulse variation is to some extent smoothed out by the 30 ms Hamming window used in the spectral calculations for this figure. The relatively long closed phase of the glottal pulse should also contribute to narrow formant bandwidths, and this can probably also be inferred from the sharp peaks in the average spectra of Figure 11.11. The relatively high RG observed for this voice quality would also tend to boost the region of H2 relative to H1, a feature which has been noted in the literature as a characteristic of creaky voice.

Tense voice involves a higher degree of tension in the entire vocal tract as compared to the neutral setting. At the laryngeal level, the two parameters which show a particular increase in tension are adductive tension and medial compression. This would correspond to the term "pressed phonation," used by many authors. The increased muscular tension associated with tense voice is also likely to affect the respiratory system, causing increased subglottal pressure, as well as the supralaryngeal tract resulting in more forceful articulation. Acoustically, this voice quality – which was here analyzed only for *strikes* – exhibited a very low RA suggesting a sharp, full closure of the vocal folds. The related frequency measure FA was higher than for all the other voice qualities looked at in this context, and so one would expect a relative strengthening of high-frequency components. The glottal pulse was also more skewed for tense voice (lowest RK values), had a small open quotient (OQ) and a high RG. The picture of a relatively skewed pulse with a nearly instantaneous closure and a long closed phase accords well with the physiological description of high adductive tension and medial compression. These source values suggest that the lower harmonics should be attenuated relative to higher frequencies. This effect shows up clearly in Figure 11.11, and we observe in Figure 11.10 that F1 dominates the spectrum. As with creaky voice, the high RG will affect the ratio of H1 to H2 by relatively boosting the latter. For this speaker, there is considerable similarity between tense and creaky voice parameters, if one ignores the pulse-to-pulse variability sometimes found for the latter.

Lax voice involves a lesser degree of tension in the entire vocal tract and typically tends to have opposite characteristics to tense voice. At the laryngeal level there is a reduced degree of adductive tension and medial compression. Phonation may therefore be similar to breathy voice, sounding softer and lower pitched than modal voice: however, the amount of change in these tension parameters is often less than for breathy voice. Source measurements show parameter values similar to those of breathy voice. Excepting the extremely low RG values, differences found between lax and breathy voice suggest that, of the two, lax voice is the closer to modal voice.

For all voice qualities, the reader should bear in mind that they are not fixed entities. Nonmodal qualities may occur to a greater or lesser degree, i.e., may be further from or closer to modal voice. Voice qualities can also be of a compound type, as for example in whispery creaky voice (for more on this, see Laver, 1980, 1994).

As was mentioned for modal voice, there may be considerable dynamic variation even within utterances considered to have been spoken with a single voice quality. The same point holds across voice qualities: a nonmodal quality may be nearer or closer to the modal depending on context. For example, our data suggest that differences were greater in the stressed than in the relatively unstressed syllable. Another example is the creakiness in creaky voice, which is intermittent and appears to be associated with particular environments. It seems unlikely that a good implementation of a voice quality change in synthesis will be achieved by a single set of transformations, but will require context-sensitive rules.

4 Determinants of Voice Source Variation

Individuals differ in the voice quality that they may habitually use. Even within the speech of the individual there are considerable dynamic changes in the voice. Some aspects of source variation are within the speaker's control, and may be linked to the linguistic content of utterances or to the speaker's intention of signaling paralinguistic information on mood and attitude. Some source differences serve a sociolinguistic function insofar as social, regional, or linguistic groups may tend towards frequent use of particular voice qualities. But beyond linguistic, paralinguistic, and sociolinguistic influences, individuals' voices are shaped by many factors, some of which are not within their control, such as the physical properties of their vocal apparatus. This section presents an overview of some of these determinants of voice source variation.

4.1 Linguistically determined variation of the source

Variations in the voice source may be associated with segmental or suprasegmental aspects of the linguistic code. Some of the variations are heard as shifts in auditory voice quality, and these can have a contrastive function in a language's system. Other changes are not necessarily heard as distinct shifts in voice quality, yet they contribute to the system in a variety of ways.

Where used contrastively, the voice qualities most frequently mentioned are modal voice, creaky voice (also called laryngealized), and breathy voice (also called murmured, or in the case of consonants, voiced aspirated). For the latter quality Laver (1980) suggests, that in terms of his classifications, whispery rather than breathy voice may be involved. However, as there is considerable variability in the realizations of breathy voiced segments, it is likely that they lie at different points on the acoustic continuum from breathy to whispery voice (see below and also section 3), and they will simply be referred to here as breathy voiced.

The terms tense and lax have also been used to describe contrasts based on voice quality, but as is clear from Maddieson and Ladefoged (1985), the terms can be misleading, and likely to be used in a phonological rather than in a phonetically accurate sense. Thus, the authors speculate, tense might signify modal voice in one language – such as Wa, a Mon-Khmer language of southwest China, where it contrasts with a lax quality which may be phonetically breathy voiced – but creaky voice with raised larynx in another language – such as Yi, also spoken in southwest China, where the contrasting lax quality would appear to be modal voice.

Other more extreme voice qualities also occur, e.g., the very harsh "growl" described by Rose (1989) for the Zhenhai variety of Wu Chinese. This quality would appear to involve the ventricular folds as well as epiglottalization, and would sound like a pathological voice quality to an English speaker's ear. Judging from the descriptions in the literature, this "growl" would seem to resemble the very extreme quality referred to as "strident" which is found in vowels of !Xóõ, a Khoisan language spoken by Bushmen in southern Africa, and which involves a narrowing of the aryepiglottic folds, along with pharyngeal constriction and backing of the epiglottis (Ladefoged, 1982; Traill, 1985; Ladefoged & Maddieson, 1996).

4.1.1 The segmental level The contrastive use of voice quality for vowels or consonants is fairly common in South East Asian, South African Indo-Aryan, and Native American Languages, and these have been the focus of many studies, many of which have been carried out at UCLA. Although both vowels and consonants may employ voice quality contrasts in a given language, Ladefoged (1982) points out that it is very rare to find contrasts at more than one place in a syllable. The term register is often used to describe voice quality contrasts, but is a phonological cover term, and subsumes any other phonetic features, such as vowel quality, vowel duration, and small or exaggerated f_0 differences which are often associated with such contrasts in particular languages. Gordon and Ladefoged (2001) illustrate these other features and discuss how they can serve to reinforce a voice quality contrast.

As a practical consequence of the facilities available, especially in fieldwork, most investigations of linguistically contrastive source effects have tended to concentrate on spectral measurements based on the speech waveform, such as those outlined in section 2.5. However, for the reasons mentioned there, using these kinds of measurements to characterize an essential voice quality difference can be problematic where there are concomitant differences in formant frequencies or in f_0. For vowels, contrasts are typically of a two-way kind, e.g., the breathy voiced versus modal voiced vowels in Gujerati (for instrumental descriptions, see Fischer-Jørgensen, 1967; Bickley, 1982). A more unusual case is the six-way opposition described for !Xóõ (Ladefoged, 1982; Traill, 1985). This language distinguishes modal and breathy voiced vowels. Each of these qualities can occur with additional creakiness, to give creaky voice and breathy creaky voice. (We should note here that the latter would be termed whispery creaky voice in Laver's system, where the combination of breathiness and creakiness is regarded as an impossible combination.) Finally, both modal and breathy voiced vowels can occur with the additional extreme strident quality, mentioned above. In the case of consonants, voice quality contrasts have been reported for stops, nasals, liquids, and approximants. A modal versus breathy voice contrast of nasals is reported for Tsonga and for other Bantu languages of Moçambique and South Africa (Traill & Jackson, 1987). Breathy voiced stops are characteristic of many Indo-Aryan languages including Nepali, Gujerati, and Hindi, where they may contrast with (modal) voiced, voiceless unaspirated, and voiceless aspirated stops. For a description of the contrasts of Hindi, see Dixit (1987, 1989).

Note that where consonants are described as having contrasting voice qualities, the acoustic manifestation often appears to be primarily located at the onset or offset of the vowel. The acoustic effect in these cases is attributed to laryngeal differences associated with the consonant, but affecting the initial or final portion of the vowel. Dixit (1987) is at pains to point out that although glottal abduction in the voiced aspirated stop occurs about half way through the closure, the stop interval should not itself be regarded as different from that of the normally voiced stop. Thus the breathy voicing, though depending on the consonant, is effectively realized in the following vowel. In a similar vein, Traill and Jackson (1987) show for the breathy voiced nasals of Tsonga, that the acoustic effects are mostly associated with the vowel onset, and that vocal fold abduction for the breathy voiced nasal begins during the nasal consonant. In the case of creaky voiced sonorants in intervocalic position, there is a strong cross-linguistic tendency for the creakiness to be realized in the early part of the consonant along with the final portion of the preceding vowel. This and other findings regarding the cross-linguistic distribution of creaky voiced sonorants are reviewed by Gordon and Ladefoged (2001), who speculate that they may be motivated by the need to preserve the perceptually important place and manner cues of the CV transition.

It should be noted that even where vowels use contrastive breathy or creaky voice, the creakiness or breathiness is often localized to a portion of the vowel, rather than persisting throughout. As for consonants, it has been argued that such a tendency might also reflect a requirement to maintain other auditorily important aspects of particular phonological contrasts (see Steriade, 1999; Gordon & Ladefoged, 2001; Silverman 2003).

In view of these realization tendencies, it is not surprising that the phonological domain of particular voice quality contrasts is not always clearcut, and there may be reasons for preferring to treat a particular contrast as concerning the vowel, the consonant, or the syllable. In the Jalapa de Diaz dialect of Mazatec, Mexico, contrasts involving modal voice, breathy voice, and creaky voice qualities are found. Two possible analyses are suggested by Kirk et al. (1984). One is to view the language as having a three-way contrast at the level of the syllable. Alternatively, it can be viewed as having an opposition of modal and breathy voiced vowels, and of modal and creaky voiced consonants. For a discussion of the domain of the voice quality contrast in Wu Chinese, see Jianfen and Maddieson (1989).

As was pointed out in section 3, voice qualities differ from each other in a scalar rather than in a discrete way. For any given voice quality, e.g., breathy voice, it will occur to differing degrees across languages or even for different speakers of one language/dialect. Maddieson and Ladefoged (1985) point out that the ratio of breathiness to voicing for breathy voiced segments is greater in Hindi than in Yi and greater in Yi than in Tsonga. Wayland and Jongman (2003) suggest that in the Chanthaburi dialect of Khmer, there is a male/female difference in where the vowel contrasts are located along the tense–lax continuum. Concerning cross-speaker variation, Ladefoged hypothesizes that although speakers of a single dialect may vary in the degree of breathiness they employ for a breathy voiced segment, all speakers produce the contrast by using different degrees along a continuum. Thus, for a speaker with an intrinsically breathy voice, a modal versus breathy voiced contrast would be achieved by increasing the breathiness where relevant. It is therefore not surprising that attempts to find measures that allow classification of voice quality in absolute terms, from data illustrating linguistic contrasts, have tended to run into difficulty.

Discussion so far has only been of cases where voice quality is considered to have a contrastive function, i.e., is taken to be the main phonetic feature on which a phonological contrast rests. Differences can also occur which would not be considered phonologically distinctive. Classes of consonants differing in manner of articulation would appear to have intrinsically different voice source characteristics. For a description of differences among voiced stops, fricatives, nasals, and laterals, see Gobl and Ní Chasaide (1999b). This type of variation most likely reflects the supraglottal configuration and resulting aerodynamic conditions pertaining to the different classes of segments, and as such, one would expect it to be universal. For much the same reasons, different classes of vowels also manifest rather subtler differences in their source parameters (Ní Chasaide et al., 1994).

Voiceless consonants occasion more extensive perturbations of the source than the voiced, and this is an inevitable consequence of the devoicing gesture of the vocal folds. Although these perturbations are large, they are often of short duration, and are essentially akin to the well-described f_0 perturbations associated with voiceless consonants. Temporally much more extensive effects also do occur in some cases, so that the phonatory quality of a substantial portion of a vowel may vary considerably in a way that reflects the voiced/voiceless nature of an adjacent stop. Languages – and even dialects of a language – can differ considerably in this respect, as is illustrated by data for Swedish, English, French, and German in Ní Chasaide and Gobl (1993). The data of that study suggest underlying differences in laryngeal control that concern both the precise timing of glottal abduction and, in some cases, the laryngeal tension settings (see discussion in Gobl & Ní Chasaide, 1999b).

Although these noncontrastive source differences might seem relatively unimportant to the linguist concerned with the contrastive units of sound systems, we do need to know more about them, because they provide the baseline material for descriptive analyses and important insights into the production of speech. Furthermore, such subcontrastive features often trigger sound change. For example, voiced and voiceless consonants are known to have been an important historical source of register contrasts in vowel systems. See for example the discussion of the development of Khmer in Wayland and Jongman (2003).

4.1.2 The suprasegmental level Suprasegmental phenomena such as intonation, tone, and stress have been extensively studied, though primarily in terms of f_0 (and to a lesser degree intensity) variation. These relatively well-understood aspects are explicitly excluded from the present coverage. However, it is worth noting that when f_0 and intensity vary, many other aspects of the glottal pulse are also likely to covary. But voice source variation plays a role in the suprasegmental systems of languages quite beyond those aspects that are strictly dependent on f_0 and intensity.

Tone Among tonal languages it is not unusual to find that a particular voice quality is associated with specific tones. Three of the seven tones of Green Mong (a dialect of Hmong, spoken in Southeast Asia) have rather similar f_0 contours, and would appear to be primarily distinguished on the basis of phonatory quality (Andruski & Ratliff, 2000). In Mandarin tones, creaky voice appears to be often associated with the third low-dipping tone and the fourth falling tone (Davison, 1991; Belotel-Grenié & Grenié, 2004). The yin and yang tones of Wu Chinese described by Jianfen and Maddieson (1989) are also characterized by specific voice qualities. The yang tones differ from the yin in that they employ breathy phonation and begin with a lower f_0 onset. See also Rose (1989) on the rather different realizations of yin/yang tones in the Zhenhai variety of Wu.

As f_0 differences tend to be associated with different voice qualities in any case, it is hardly surprising to find register correlates of tonal contrasts and vice versa. Despite such correlations, many authors are at pains to point out that f_0 and voice quality are separately controllable, and that variation in one does not allow prediction of variation in the other. This point is illustrated in the cross-language data on Tamang, Naxi, and Vietnamese, presented in Michaud and Mazaudon (2006).

The link between voice quality and f_0 may have historical implications, and the likelihood of tonal contrasts evolving from earlier voice quality contrasts has been discussed by Maddieson and Hess (1987), Jianfen and Maddieson (1989), Rose (1989), Abramson et al. (2004), and Michaud and Mazaudon (2006).

As might be expected, there are also cases where contrasts in specific languages are open to competing analysis as involving primarily register or tone. See for example the lively debate concerning the so-called register contrast of Mon (Lee, 1983; Diffloth 1985; Thongkum, 1987). Maddieson and Hess (1987) suggest that the six tones of Lisu should be interpreted as a four-way tonal contrast, with a register contrast in two of the tones. Rose (1989) argues for an interpretation of the yin/yang difference in tones of the Zhenhai dialect of Wu as a register contrast, which interacts with a three-way tonal contrast. Clearly f_0 and voice quality features are two aspects of the voice source which can work individually or in a complementary way in the formation of tonal contrasts.

Intonation Although one normally thinks of intonation in terms of pitch dynamics, we have argued in Ní Chasaide and Gobl (2004a, 2004b) that variations in source parameters other than f_0 contribute importantly. Figure 11.13 shows the dynamic variation of f_0, EE and RK for the Swedish utterance "Inte i detta århundrade" [ɪntɪ ^{x}dɛtːa oːrhɵndradə]. Declination, typically thought of as the gradual drop in f_0 in the course of an utterance, is also clearly evidenced in the declining strength of the glottal pulse excitation (EE) and in the increasingly symmetrical glottal pulse shape (RK). Linguists tend to describe phrase boundaries in terms of phrase and boundary tones (as well as temporal features), yet voice source features are also important. The sharp rise in RK towards the phrase end in Figure 11.13, indicative of a breathy voice offset, is a typical marker of the phrase boundary. Episodes of creakiness towards the end of the utterances are also often

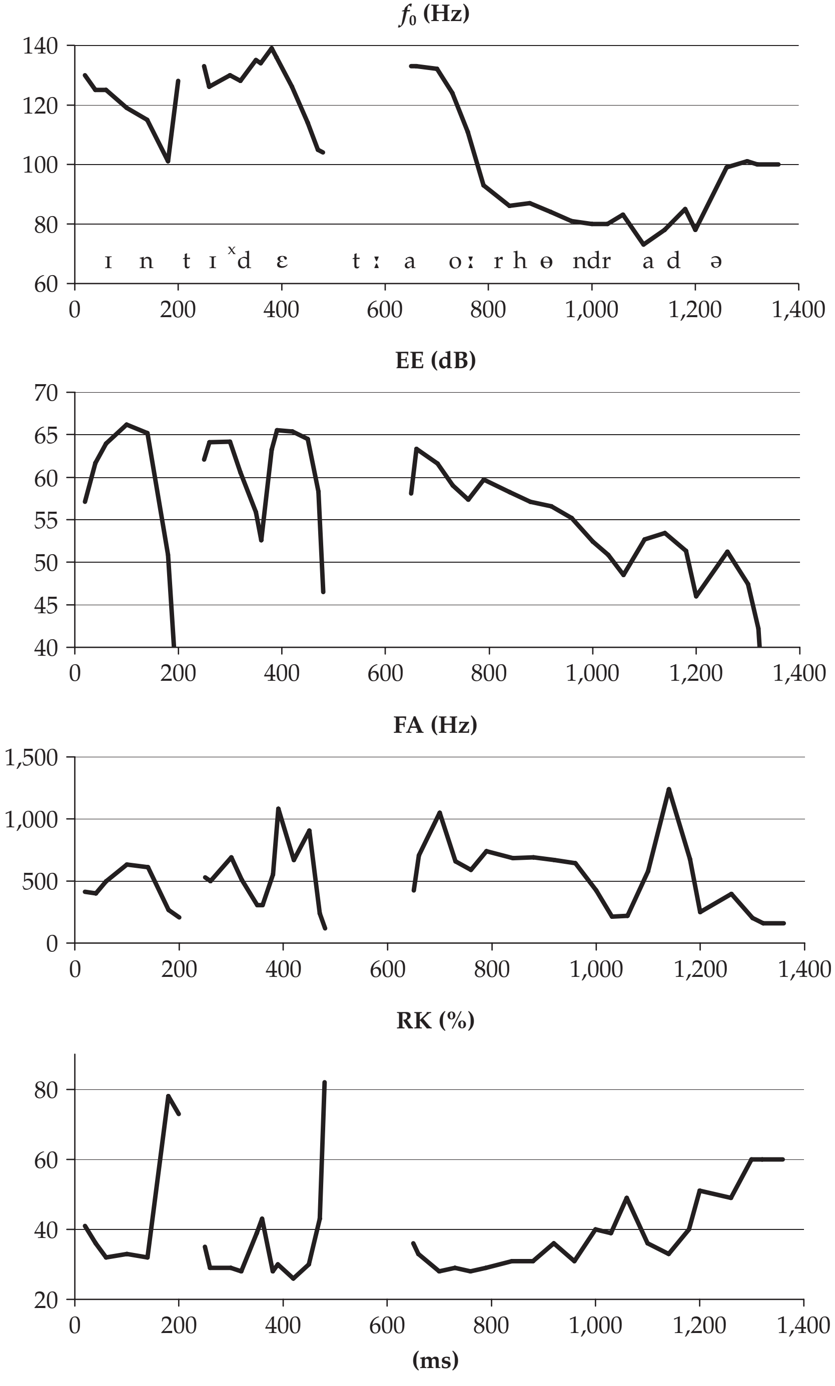


Figure 11.13 Voice source dynamics in the Swedish utterance "Inte i detta århundrade" [ɪntɪ ˣdɛtːa oːrhɵndradə].

associated with the phrase boundaries of declarative sentences of some languages, such as Swedish and certain varieties of English (see Fant & Kruckenberg, 1989; Epstein, 2003; Ní Chasaide & Gobl, 2004a).

Voice source features also contribute to the prominence of accented syllables. The fragment in Figure 11.14 illustrates the variations in f_0, EE, FA, and RG for the final two words in an intonational phrase of Swedish, "stor sal" [stuːr sɑːl], where the nuclear accent is on "sal." Note that prominence of the accented syllable [sɑːl], is not just a matter of the salient low tone in the early part of the vowel, but is clearly contributed to by the tenser phonatory settings, which we can infer from the combination of source parameters; the high excitation strength (EE), the relative strengthening of the higher source components (high FA), and the increase in the glottal frequency (high RG). Note that the boosting of higher harmonics as a correlate of accentuation has been suggested in a number of other studies (e.g., Sluijter & van Heuven, 1996; Campbell & Beckman, 1997; Fant et al., 2002; Epstein, 2003).

Further observations concerning focal stress were made by Gobl (1988), where source characteristics of a word were analyzed in focal, prefocal, and postfocal positions of an utterance. The dynamic range of the source excitation (EE) was considerably greater when the word was in focal position than in the other environments, being stronger for the vowel and weaker for the surrounding voiced consonants. This is effectively an enhancement of the vowel/consonant distinction in the stressed syllable. This enhancement may be even more important than absolute levels of the source excitation.

Some studies have focused on the possibility that high and low pitch accents might be differentiated by source characteristics other than pitch. Pierrehumbert (1989) reported source correlates of high pitch accent not otherwise found with high pitch, but more recently, Epstein (2003) did not find consistent source differences between high and low pitch accented tones. This question is complex, due to the fact that there are in any case interactions between f_0 and other voice source parameters, which are still not well understood.

It should be emphasized that, with the possible exception of creaky episodes at phrase boundaries, the kinds of source correlates of intonational categories referred to here are not heard by the listener as changes in voice quality. Quite simply, within any utterance spoken with, say, modal voice, there is ongoing extensive modulation of source parameters as part of the utterance's prosodic expression.

This is an essentially unresearched aspect of prosody, but as argued in Ní Chasaide and Gobl (2004a, 2004b), a holistic approach to prosodic analysis is called for, where linguistic prosody – of tone or intonation – is treated as the dynamic, temporal variation of the entire source signal, and not just as the frequency of the first harmonic, f_0.

4.2 *Paralinguistic aspects of voice source variation*

Speakers use temporary shifts in voice quality to signal affect, i.e., their mood and emotion, as well as their attitude to the interlocutor and to the context of the interaction. In spoken interaction, meaning can broadly be viewed as transmitted

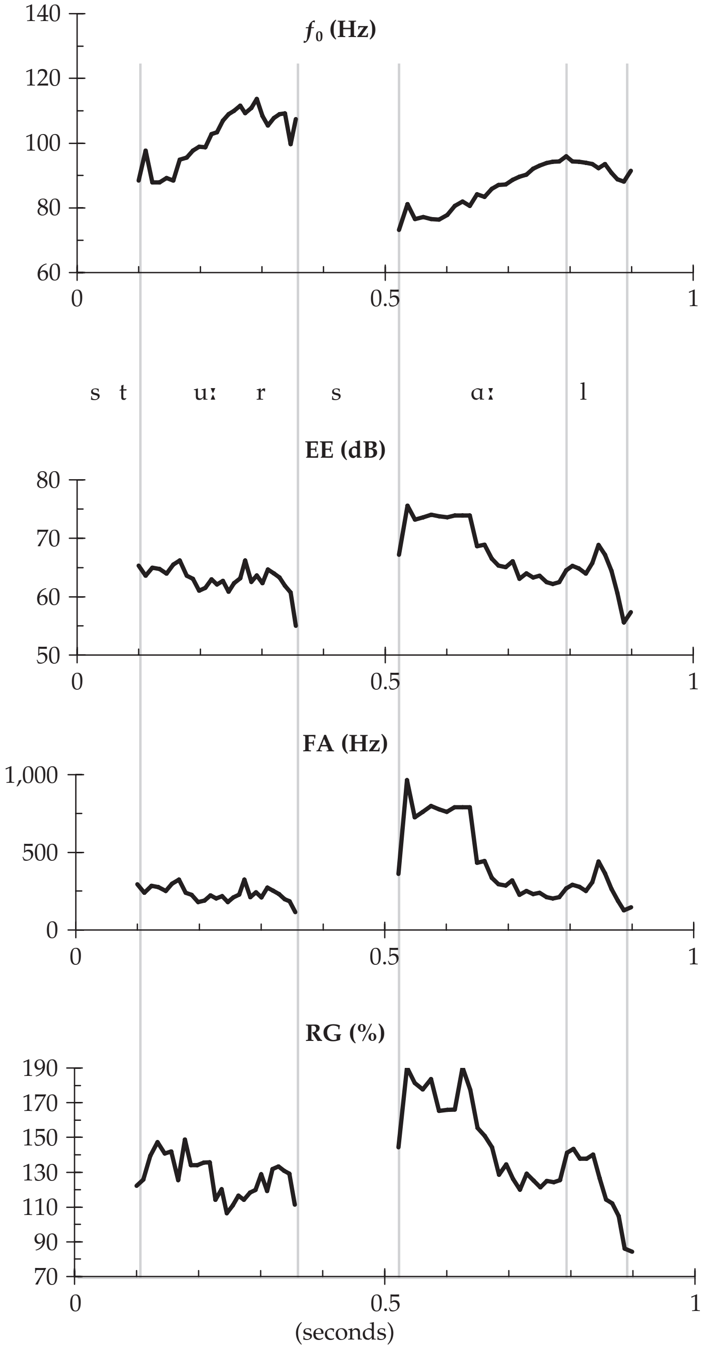


Figure 11.14 Variation in the voice source parameters EE, RA, and RG for the Swedish utterance fragment "stor sal" [stuːr sɑːl].

on two levels: firstly in terms of the verbal content, which can be more or less represented in text; secondly, in terms of the tone of voice which carries the affective content, and which cannot be represented in text. Tone of voice can reinforce or run contrary to the verbal content. For example, speakers might sound interested or indifferent in a true reflection of their attitudes, but equally, they might adopt a tone of voice that signals interest in a topic which they find truly boring, or might feign indifference to cover a very real but controllable interest, anger, or other emotion. It is worth remembering that voice quality changes associated with extremes of emotion are likely to be involuntary, arising from physiological changes brought about by the emotional state itself. As such, they are extralinguistic and presumably universal, and do not belong to the conventional, learnt system of affect signaling. Voluntary modulations of the voice to signal affect, though undoubtedly often rooted in the correlates of involuntary emotional expression, are nonetheless more prone to being conventionalized within a language or cultural group, and hence, less likely to be universal.

Impressionistic observations in the literature have traditionally associated specific voice qualities with particular affective states, e.g., for speakers of English, creaky voice is said to be associated with boredom, breathy voice with intimacy, whispery voice with confidentiality, and harsh voice with anger (see, for example, Laver, 1980). The fact that there is little quantitative information on this important aspect of spoken communication mainly reflects the many technical and methodological difficulties such research presents (see discussion in Gobl & Ní Chasaide, 2003b).

As a consequence of the inherent difficulty in obtaining reliable acoustic measures of the voice source, most research in the field has tended to focus on the more readily measurable parameters of f_0, intensity, and tempo. Extensive studies, carried out by Scherer and colleagues on the vocalization of emotion (e.g., Scherer, 1986; Kappas et al., 1991; Scherer et al., 2003) have provided a wealth of information on the variation of f_0 (changes in the mean value, range, contour type, variability), intensity (changes in the mean value), and temporal patterning. (See also related research, for instance Mozziconacci, 1998, and Carlson et al., 1992.) While these studies yield many important insights, we still lack basic knowledge on the crucial dimension of voice quality. A pilot study of voice source correlates of some portrayed affective states (Yanushevskaya et al., 2007) highlights the fundamental importance of voice quality and illustrates that many source parameters differ in rather complex ways.

A further problem concerns the speech samples one should analyze. Most studies to date have focused on strong emotions such as anger, fear, joy, and sadness and have employed portrayals of these emotions, often by actors, of some "semantically neutral" utterances. The considerable intersubject variability frequently found in these studies suggests that individuals may vary not only in their style of acting/portrayal, but also in their ability to simulate emotions. Furthermore, actors' portrayals, though interesting in themselves, are likely to be quite different from spontaneous, naturally occurring emotions. Departing from the typical paradigm, Campbell and Mokhtari (2003) have used a very large corpus of spontaneous speech, recorded for a single speaker over a number of years. Measures of NAQ (see section 2) illustrate that there is systematic variation of the voice source, depending on the relationship to the interlocutor (mother, infant, friends, etc.) and the nature of the interaction between them. This work has served to highlight the role of voice quality in providing the expressive, human dimension to the verbal content of utterances.

In a rather different approach, the authors and colleagues have used synthesis and perceptual experiments to explore the mapping of affect to voice quality. Parametric source-filter synthesis was used to produce stimuli differing in terms of voice quality. (For a full description of the basic methodology, see Gobl & Ní Chasaide, 2003b.) A series of perception tests elicited whether and to what extent the different voice qualities evoke different affective colourings. Each individual test examined a single pair of opposite affective attributes, e.g., bored/interested. A seven-point scale was used where 0 represented no affective colouring, and deviations from 0 (from –3 to +3) indicated the strength of any perceived affect. The full set of attribute pairs tested included *relaxed/stressed*, *content/angry*, *friendly/hostile*, *sad/happy*, *bored/interested*, *intimate/formal*, *timid/confident*, and *afraid/unafraid*.

Results of an initial experiment are illustrated in Figure 11.15 (Gobl & Ní Chasaide, 2003b), and show for each test the mean rating for each voice quality tested. They illustrate how altering voice quality alone can be potent in imparting different affective colorings to an otherwise neutral utterance. And, while offering broad support for traditional wisdoms, the results also suggest some refinements concerning the association of voice quality and affect. For example, lax-creaky

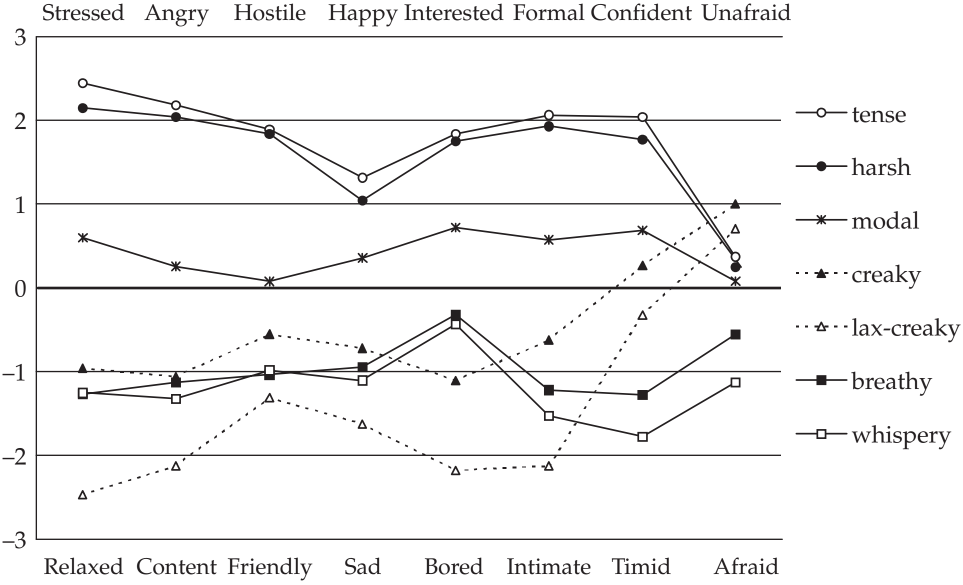


Figure 11.15 Perceived strength of affect signaling for seven voice qualities. 0 = no affective content and ±3 = maximally perceived.

voice rather than creaky voice yields the highest ratings for boredom; and this same voice quality is here more potent than breathy voice in evoking intimacy. They also indicate that, rather than a one-to-one mapping of voice quality to affect that is typically implied, a particular quality tends to be associated with a cluster of (often, though not necessarily, related) affective states. So, for example, lax-creaky voice yields not only high ratings for *bored* and *intimate*, but also for *relaxed* and *content*.

Further experiments (Gobl et al., 2002; Yanushevskaya et al., 2005) have explored how voice quality and f_0 combine in cueing affect. Pitch modulation also contributes to tone of voice and, as mentioned above, there is a considerable body of literature on the f_0 correlates of certain emotions. For these experiments, in addition to the voice quality stimuli of the earlier tests (*VQ* stimuli), further stimulus sets were included, involving different or further modifications of the utterance used in the first experiment. One set (f_0 stimuli) included different affective f_0 contours, with unchanging (modal) voice throughout. The f_0 contours used were based on the affect-related contours provided in Mozziconacci (1995). A further group of stimuli included the same array of voice qualities as the first set, but this time each affective f_0 contour was combined with one or other of the voice qualities ($f_0 + VQ$ stimuli).

Results of these experiments showed that the f_0 stimuli were rather ineffective in evoking affect, compared to stimuli with voice quality manipulations or a combination of voice quality and f_0 variation. These results may explain why attempts to replicate affective speech in synthesis by incorporating the f_0 variations described in production studies have not met with success.

This same experimental paradigm has also been used to explore cross-language/culture differences in voice-to-affect mapping, by running the same tests on subjects from different language backgrounds, including English and Japanese (see Yanushevskaya et al., 2006). While the associations of particular stimuli with certain affects were in many cases rather similar for the different language groupings, some striking differences did emerge. For example, a tense voice quality with a high and dynamically varied f_0 contour was associated with indignation by English subjects, while for Japanese subjects this quality was associated with intimacy.

4.3 Sociolinguistic and cross-language variation

Voice quality may also have a sociolinguistic function, and serves to differentiate among linguistic, regional, and social groups. Supralaryngeal as well as phonatory features may be used to this end. As anyone who has taught a foreign language will attest, cross-language differences in voice quality are an important aspect of a convincing accent, but difficult to teach as they are virtually never described in the linguistic or applied linguistic literature. This can lead to cultural misperceptions, as the native speaker is likely to interpret the foreigner's voice quality in terms of his/her own paralinguistic system for affect or attitude signaling. The difference mentioned above between Japanese and English listeners serves to illustrate just how such misperceptions are likely to arise.

Within a particular language or dialect group, voice quality features may demarcate social subgroups. In Edinburgh English a greater incidence of creaky voice is associated with a higher social status, whereas whispery and harsh qualities are linked to a lower social status (Esling, 1978). Similar findings have been reported for Glasgow English (Stuart-Smith, 1999). In Norwich, working- and middle-class accents are differentiated on the basis of habitual phonatory and supralaryngeal settings (Trudgill, 1974). Other social groupings may also tend towards different voice qualities. Rose (1989) suggests that the extremely harsh "growl" mode of phonation found in the Zhenhai dialect of Wu Chinese differs in terms of the sex and age of the speaker, being least harsh for women, and most harsh for old men. In investigating differences which are correlated with sex and age, it is of course important to distinguish between truly sociolinguistic markers and differences which are due to laryngeal anatomy.

4.4 Extralinguistic determinants of the voice source

There are other factors which determine the quality of the voice, many of them beyond the control of the speaker. Differences in the size, shape, and muscular tone of the laryngeal structures play a major role. The voices of men, women, and children reflect mostly anatomical differences, although intrinsic, anatomy-based features may be enhanced or reduced depending on the socio-cultural context. For example, women working and competing in a male environment may choose to adopt a mode of phonation more similar to that of the male. Our understanding of the female and child voice lags behind that of the male voice. For some descriptions of the female voice see Karlsson (1992), Hanson (1995, 1997), Holmberg et al. (1988), and for child versus adult male differences, see Gobl (1988).

Voice quality is also affected by the individual's physical and mental health. There has been considerable study of certain acoustic correlates (mostly f_0 and intensity) of psychiatric illnesses such as depression and schizophrenia. For a summary of studies on the vocal indicators of depression, see Scherer (1987). Pathologies of the laryngeal structures also affect vocal quality and many studies of the voice have been medically motivated. Figure 11.16 illustrates a number of source parameters (f_0, EE, RA, RK) for a female speaker with vocal fold nodules, as compared to a normal speaker matched for age, sex, and accent. The pathology appeared to be particularly associated with the initiation of voicing, where the vibratory cycle was grossly perturbed in a number of ways. f_0 was very high, the glottal excitation was weak (low EE), there tended to be a considerable degree of dynamic leakage (high RA), and the pulse shape was generally more symmetrical (high RK). Probably as important as the actual values was the unstable nature of the glottal pulse during this initial interval, evident from the considerable pulse-to-pulse fluctuations for all parameters. At a certain point, somewhere between the 10th and 20th glottal cycle, the phonatory pattern switched abruptly. f_0 dropped by about an octave to normal values and the other source parameters indicated a more normal and stable glottal pulse. For further details on this study, see Kane and Ní Chasaide (1992).

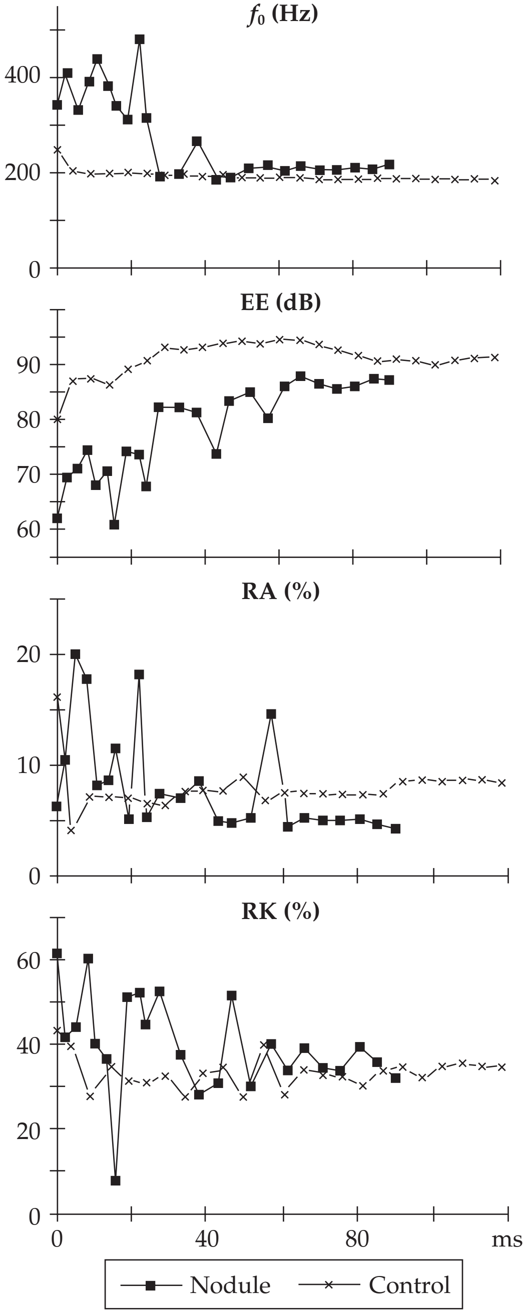


Figure 11.16 Values of f_0, EE, RA, and RK compared for a speaker with vocal fold nodules and a normal speaker, from the initiation of voicing (following stop release) in the nonsense word *baa*.

Over and above the linguistic and nonlinguistic factors mentioned so far, voice quality also carries uniquely personal information and serves an important function in allowing us to identify speakers and tell them apart.

5 Future Research

The human voice has evolved as a vehicle for conveying many different types of information, and human listeners have developed the ability to detect very small, very subtle voice quality changes, and interpret their function. In spite of improvements in recent years in the techniques for describing and modeling voice source variation, our abilities lag far behind what the human ear can effortlessly do. At this stage of play, we can but appreciate the scale and complexity of the research that will be needed to gain a full understanding of how the voice is used in speech communication. Most of the studies to date have been very limited, either in the quantity of data analyzed, or in the kinds of source measures made.

Because the voice source simultaneously imparts information at many different levels, our understanding of any one aspect will depend on our grasp of the whole picture, so that we can properly interpret one dimension of variation relative to the other dimensions. To take an obvious example, cross-speaker differences are often quite significant, and this can be greater than the (sometimes subtle) types of within-speaker variations that impart linguistic or paralinguistic information.

Perhaps less obviously, we need research that spans what at the moment might appear to be separate fields of research. Particularly, linguistic prosody and the paralinguistic signaling of affect by prosody are in our view two aspects of a single system. Although, as is usual, we have treated them here under separate headings, we would hold that these can only be properly understood relative to each other. Currently, these two aspects of spoken communication are pursued as two distinct entities, mostly within different disciplines. Within linguistics, the focus is mainly on the phonological, contrastive aspects of intonation. There is little attention to the role of intonation in affect signaling, even though this was seen as a central function of intonation in earlier linguistic treatments (e.g., O'Connor & Arnold, 1961). Research on the vocal expression of affect has been largely conducted within psychology, taking little account of linguistic prosody. In our view, the paralinguistic signaling of affect involves, not a separate system, but rather systematic modifications and enhancements to the underlying elements of linguistic prosody.

The key to understanding how these two dimensions of prosody collaborate may be to provide a more holistic description of voice dynamics. In most of the current literature, the prevailing wisdom is that linguistic prosody is about f_0 dynamics, and paralinguistic signaling depends crucially on voice quality variation. As discussed above, voice source parameters vary importantly for both linguistic *and* paralinguistic aspects of prosody. Likewise, f_0 dynamics contribute importantly to both. Ideally, we need to include in our descriptive framework how the entire source signal – f_0 *and* the other important source parameters – varies dynamically over time. As argued in Ní Chasaide and Gobl (2004a), such a holistic approach will hopefully, by providing a fuller understanding of the workings of linguistic prosody, also lay the foundations for a better understanding of paralinguistic prosody. By removing the unhelpful dichotomies between linguistic and paralinguistic, and between f_0 and voice quality, we should be able to work towards a unified account of voice prosody.

Progress towards such a holistic account of prosody will inevitably depend on the availability of suitable analysis tools. The manual interactive techniques outlined in section 2, permit a fine-grained analysis of the source, but because of their labour-intensive nature, they are not suitable for the large-scale studies that would be ideally needed for progress in this field. The research agenda must therefore be directed not only at descriptive studies, but also at devising new techniques or enhancing current ones to automate the acquisition of accurate source data. Many new automatic algorithms have been proposed, and while they may be useful for estimating certain global, long-term aspects of the source, they do not yet generate the accuracy needed to provide the finer-grained analyses we would require for many purposes. Most techniques to date focus on signal processing solutions: it is our belief that in order to make significant progress in the area of source-filter decomposition, new approaches will be required which are to a greater extent guided by knowledge of speech production. And as in other areas of speech research, acoustic analysis must be supplemented by physiological experiments, to elucidate underlying production processes, using for example high-speed digital imaging of vocal fold vibration, as described by Hirose (this volume).

An improved understanding of all aspects of the voice source signal and of how it varies in speech would open the door to many applications. The most immediate application of providing more expressive and potentially more flexible voices in speech synthesis would greatly enhance the acceptability of synthesis-based devices. One can envisage at some future date the possibility of customized voices in aids for the vocally impaired, designed to match the original voice quality of the user, and allowing affect-conveying voice modulations. Furthermore, when reasonably accurate automatic analysis procedures become available, one can envisage many other applications in areas such as speaker recognition and verification. And with an increased understanding of the range and types of variation found in normal and pathological voices, such techniques might also facilitate more advanced methods for acoustic screening and for diagnostic procedures for voice disorders.

NOTE

The authors would like to acknowledge that some of their work discussed in this chapter was carried out within the framework of a number of research projects, including the EU-funded ESPRIT/Basic Research Actions ACCOR I and SPEECH MAPS, the working groups ACCOR II and VOX, and the Network of Excellence on Emotion, HUMAINE, as well as the projects Prosody of Irish Dialects (funded by the Irish Research Council for the Humanities and Social Sciences) and Emovoice (funded by the Irish Research Council for Science, Engineering and Technology).

REFERENCES

Abramson, A., Thongkum, T. L., & Nye, P. W. (2004) Voice register in Suai (Kuai): An analysis of perceptual and acoustic data. *Phonetica*, 61, 147–71.

Alku, P., Bäckström, T., & Vilkman, E. (2002) Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112, 701–10.

Ananthapadmanabha, T. V. (1984) Acoustic analysis of voice source dynamics. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm), 2–3, 1–24.

Andruski, J. & Ratliff, M. (2000) Phonation types in production of phonological tone: The case of Green Mong. *Journal of the International Phonetic Association*, 30, 37–61.

Badin, P., Hertegård, S., & Karlsson, I. (1990) Notes on the Rothenberg mask. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 1, 1–7.

Belotel-Grenié, A. & Grenié, M. (2004) The creaky voice phonation and the organization of Chinese discourse. Paper presented at the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing.

Blankenship, B. (2002) The timing of nonmodal phonation in vowels. *Journal of Phonetics*, 30, 163–91.

Bickley, C. A. (1982) Acoustic analysis and perception of breathy vowels. *Speech Communication Group Working Papers I* (Research Lab of Electronics, MIT, Cambridge), 71–82.

Campbell, N. & Beckman, M. E. (1997) Stress, prominence, and spectral tilt. *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*, University of Athens, 67–70.

Campbell, N. & Mokhtari, P. (2003) Voice quality: The 4th prosodic dimension. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2417–20.

Carlson, R., Granström, B., & Nord, L. (1992) Experiments with emotive speech, acted utterances and synthesized replicas. *Speech Communication*, 2, 347–55.

Catford, J. C. (1964) Phonation types: The classification of some laryngeal components of speech production. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (eds.), *In Honour of Daniel Jones* (pp. 26–37), London: Longman.

Chan, D. S. F. & Brookes, D. M. (1989) Variability of excitation parameters derived from robust closed phase glottal inverse filtering. *Proceedings of the European Conference on Speech Communication and Technology*, Paper 33.1. Paris: Eurospeech.

Davison, D. S. (1991) An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics*, 78, 50–7.

Diffloth, G. (1985) The registers of Mon vs. the spectrographist's tones. *UCLA Working Papers in Phonetics*, 60, 55–8.

Dixit, R. P. (1987) Mechanisms for voicing and aspiration: Hindi and other languages compared. *UCLA Working Papers in Phonetics*, 67, 49–102.

Dixit, R. P. (1989) Glottal gestures in Hindi plosives. *Journal of Phonetics*, 17, 213–37.

Dolansky, L. & Tjernlund, P. (1968) On certain irregularities of voiced speech waveforms. *IEEE Transactions*, AU-16.

Epstein, M. (2003) Voice quality and prosody in English. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2405–8.

Esling, J. H. (1978) Voice quality in Edinburgh: A sociolinguistic and phonetic study. Doctoral dissertation, University of Edinburgh.

Fant, G. (1960) *The Acoustic Theory of Speech Production*. The Hague: Mouton. (2nd edn. 1970.)

Fant, G. (1979a) Glottal source and excitation analysis. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 1, 85–107.

Fant, G. (1979b) Vocal source analysis: A progress report. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 3–4, 31–54.

Fant, G. (1982) The voice source: Acoustic modeling. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 4, 28–48.

Fant, G. (1995) The LF-model revisited: Transformations and frequency domain analysis. *STL-QPSR* (Speech, Music and Hearing, KTH, Stockholm, Sweden), 2–3, 119–56.

Fant, G. (1997) The voice source in connected speech. *Speech Communication*, 22, 125–39.

Fant, G. & Kruckenberg, A. (1989) Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm), 2, 1–83.

Fant, G., Kruckenberg, A., Gustafson, K., Liljencrants, J., & Botinis, A. (2002) Individual variations in prominence correlates: Some observations from lab-speech. *Proceedings of FONETIK 2002*, *TMH-QPSR* (Speech, Music and Hearing, KTH, Stockholm), 44, 177–80.

Fant, G., Liljencrants, J., & Lin, Q. (1985) A four-parameter model of glottal flow. *French-Swedish Seminar on Speech*, Grenoble; also in *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm), 4, 1–13.

Fischer-Jørgensen, E. (1967) Phonetic analysis of breathy (murmured) vowels in Gujerati. *Indian Linguistics*, 28, 71–139.

Flanagan, J. L. (1972) *Speech Analysis Synthesis and Perception*, 2nd edn. New York: Springer.

Fröhlich, M., Michaelis, D., & Strube, H. W. (2001) SIM: Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *Journal of the Acoustical Society of America*, 110, 479–88.

Fujimura, O. (1968) Approximation to voice aperiodicity. *IEEE Transactions on Audio and Electroacoustics*, AU-16, 68–73.

Gobl, C. (1988) Voice source dynamics in connected speech. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 1, 123–59.

Gobl, C. (1989) A preliminary study of acoustic voice quality correlates. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 4, 9–22.

Gobl, C. (2003) The voice source in speech communication: Production and perception experiments involving inverse filtering and synthesis. Doctoral dissertation, KTH (The Royal Institute of Technology), Stockholm.

Gobl, C. (2006) Modelling aspiration noise during phonation using the LF voice source model. *Proceedings of the 8th International Conference on Spoken Language Processing, INTERSPEECH 2006*, Pittsburgh, 965–8.

Gobl, C., Bennett, E., & Ní Chasaide, A. (2002) Expressive synthesis: How crucial is voice quality? *Proceedings of the IEEE Workshop on Speech Synthesis*, Santa Monica, California, paper 52, 1–4.

Gobl, C., Monahan, P., Fitzpatrick, L., & Ní Chasaide, A. (1994) A new approach to source-filter decomposition. *Proceedings of the 2nd Review Meeting of the Esprit/Basic Research Action no. 6975: SPEECH MAPS*, KTH, Stockholm.

Gobl, C. & Ní Chasaide, A. (1988) The effects of adjacent voice/voiceless consonants on the vowel voice source: A cross language study. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm), 2–3, 23–59.

Gobl, C. & Ní Chasaide, A. (1992) Acoustic characteristics of voice quality. *Speech Communication*, 11, 481–90.

Gobl, C. & Ní Chasaide, A. (1999a) Techniques for analysing the voice source. In W. J. Hardcastle & N. Hewlett (eds.), *Coarticulation: Theory, Data and Techniques* (pp. 300–20). Cambridge: Cambridge University Press.

Gobl, C. & Ní Chasaide, A. (1999b) Voice source variation in the vowel as a function of consonantal context. In W. J. Hardcastle and N. Hewlett (eds.), *Coarticulation: Theory, Data and Techniques* (pp. 122–43). Cambridge: Cambridge University Press.

Gobl, C. & Ní Chasaide, A. (2003a) Amplitude-based source parameters for measuring voice quality. *Proceedings of the ISCA VOQUAL'03 Workshop on Voice Quality: Functions, Analysis and Synthesis*, Geneva, 151–6.

Gobl, C. & Ní Chasaide, A. (2003b) The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189–212.

Gordon, M. & Ladefoged, P. (2001) Phonation types: A cross-linguistic overview. *Journal of Phonetics*, 29, 383–406.

Hanson, H. (1995) Glottal characteristics of female speakers. Doctoral dissertation, Harvard University.

Hanson, H. (1997) Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101, 466–81.

Hedelin, P. (1984) A glottal LPC-vocoder. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Diego, 1.6.1–1.6.4.

Hedelin, P. (1986) High quality glottal LPC-vocoder. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, 9.9.1–9.9.4.

Hertegård, S. & Gauffin, J. (1991) Insufficient vocal fold closure as studied by inverse filtering. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 243–50). San Diego: Singular Publishing Group Inc.

Hertegård, S. & Gauffin, J. (1992) Acoustic properties of the Rothenberg mask. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm), 2–3, 9–18.

Holmberg, E. B., Hillman, R. E., & Perkell, J. S. (1988) Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America*, 84, 511–29.

Holmes, J. N. (1975) Low-frequency phase distortion of speech recordings. *Journal of the Acoustical Society of America*, 58, 747–9.

Hunt, M. J. (1978) Automatic correction of low-frequency phase distortion in analogue magnetic recordings. *Acoustic Letters*, 2, 6–10.

Hunt, M. J. (1987) Studies of glottal excitation using inverse filtering and an electroglottograph. *Proceedings of the 11th Int. Congress of Phonetic Sciences*, Tallinn, 3, 23–6.

Huffman, M. K. (1987) Measures of phonation type in Hmong. *Journal of the Acoustical Society of America*, 81, 495–504.

Iseli, M. H. & Alwan, A. (2004) An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. *Proceedings of ICASSP*, Montreal, Canada, 669–72.

Jackson, M., Ladefoged, P., Huffman, M. K., & Antoñanzas-Barroso, N. (1985) Measures of spectral tilt. *UCLA Working Papers in Phonetics*, 61, 72–8.

Jackson, M., Ladefoged, P., Huffman, M. K., & Antoñanzas-Barroso, N. (1986) Automated measures of spectral tilt. *UCLA Working Papers in Phonetics*, 62, 77–88.

Jianfen, C. & Maddieson, I. (1989) An exploration of phonation types in Wu dialects of Chinese. *UCLA Working Papers in Phonetics*, 72, 139–60.

Kane, P. & Ní Chasaide, A. (1992) A comparison of the dysphonic and normal voice source. *Journal of Clinical Speech and Language Studies*, 2, 17–29.

Kappas, A., Hess, U., & Scherer, K. R. (1991) Voice and emotion. In R. &d B. Rimé (eds.), *Fundamentals of Nonverbal Behaviour* (pp. 200–38). Cambridge: Cambridge University Press.

Karlsson, I. (1992) Analysis and synthesis of different voices with emphasis on female speech. Doctoral dissertation, KTH, Stockholm.

Kirk, P., Ladefoged, P., & Ladefoged, J. (1984) Using a spectrograph for measures of phonation types in a natural language. *UCLA Working Papers in Phonetics*, 59, 102–13.

Kirk, P. L., Ladefoged, J., & Ladefoged, P. (1993) Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In A. Mattina & T. Montler (eds.), *American Indian Linguistics and Ethnography in Honor of Laurence C. Thompson* (pp. 435–50). Missoula: University of Montana Press.

Klatt, D. H. (1980) Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67, 971–95.

Klatt, D. H. (1986) Detailed spectral analysis of a female voice. *Journal of the Acoustical Society of America*, 81, S80(A).

Klatt, D. H. & Klatt, L. C. (1990) Analysis, synthesis, and perception of voice quality variation among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820–57.

Ladefoged, P. (1971) *Preliminaries to Linguistic Phonetics*. Chicago: The University of Chicago Press.

Ladefoged, P. (1982) The linguistic use of different phonation types. *UCLA Working Papers in Phonetics*, 54, 28–39.

Ladefoged, P. & Antoñanzas-Barroso, N. (1985) Computer measures of breathy phonation. *UCLA Working Papers in Phonetics*, 61, 79–86.

Ladefoged, P. & Maddieson, I. (1996) *The Sounds of the World's Languages*. Oxford: Blackwell.

Ladefoged, P., Maddieson, I., Jackson, M., & Huffman, M. K. (1987) Characteristics of the voice source. *UCLA Working Papers in Phonetics*, 67, 119–25.

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1997) On the perception of emotions in speech: The role of voice quality. *Scandinavian Journal of Logopedics, Phoniatrics and Vocology*, 22, 157–68.

Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

Lee, T. (1983) An acoustical study of the register distinction in Mon. *UCLA Working Papers in Phonetics*, 57, 79–96.

Liljencrants, J. (1969) Speech synthesizer control by smoothed step functions. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 4, 43–50.

Ljungqvist, M. & Fujisaki, H. (1985) A comparative study of glottal waveform models. *Technical Report of the Institute of Electronics and Communications Engineers*, Japan, vol. EA85-58, 23–9.

Löfqvist, A. & McGowan, R. S. (1991) Voice source variations in running speech. In J. Gauffin & B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms* (pp. 113–20). San Diego: Singular Publishing Group Inc.

Maddieson, I. & Ladefoged, P. (1985) "Tense" and "lax" in four minority languages of China. *UCLA Working Papers in Phonetics*, 60, 59–83.

Maddieson, I. & Hess, S. A. (1987) The effect on F0 of the linguistic use of phonation type. *UCLA Working Papers in Phonetics*, 67, 112–18.

Mahshie, J. & Gobl, C. (2003) Estimating glottal parameters in nasalized speech: An analysis by synthesis. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2193–6.

Makhoul, J., Vishwanathan, R., Schwartz, R., & Huggins, A. W. F. (1978) A mixed-source model for speech compression and synthesis. *Journal of the Acoustical Society of America*, 64, 1577–81.

Michaud, A. & Mazaudon, M. (2006) Pitch and voice quality characteristics of the lexical word-tones of Tamang, as compared with level tones (Naxi data) and pitch-plus-voice-quality tones (Vietnamese data). *Proceedings of the 3rd International Conference on Speech Prosody*, Dresden, Germany, 823–26.

Mozziconacci, S. (1995) Pitch variations and emotions in speech. *Proceedings of* *the 13th International Congress of Phonetic Sciences*, Stockholm, 1, 178–81.

Mozziconacci, S. (1998) Speech variability and emotion: Production and perception. Doctoral dissertation, Technische Universiteit Eindhoven, Eindhoven.

Ní Chasaide, A. & Gobl, C. (1993) Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech*, 36, 303–30.

Ní Chasaide, A. & Gobl, C. (2004a) Voice quality and f_0 in prosody: Towards a holistic account. *Proceedings of the 2nd International Conference on Speech Prosody*, Nara, Japan, 189–96.

Ní Chasaide, A. & Gobl, C. (2004b). Decomposing linguistic and affective components of phonatory quality. *Proceedings of the 8th International Conference on Spoken Language Processing, INTERSPEECH 2004*, Jeju Island, Korea, 2, 901–4.

Ní Chasaide, A., Gobl, C., & Monahan, P. (1992) A technique for analysing voice quality in pathological and normal speech. *Journal of Clinical Speech and Language Studies*, 2, 1–16.

Ní Chasaide, A., Gobl, C., & Monahan, P. (1993) Dynamic variation of the voice source in VCV sequences: Intrinsic characteristics of selected consonants. *Proceedings of the 1st Review Meeting of the Esprit/Basic Research Action no. 6975: SPEECH MAPS*, Grenoble, Institut de la Communication Parlée, 2, 44 pp.

Ní Chasaide, A., Gobl, C., & Monahan, P. (1994) Dynamic variation of the voice source: Intrinsic characteristics of selected vowels and consonants. *Proceedings of the Speech Maps Workshop, Esprit/Basic Research Action no. 6975*, Grenoble, Institut de la Communication Parlée, 2, 35 pp.

O'Connor, J. D. & Arnold, G. F. (1961) *The Intonation of Colloquial English*. London: Longman.

Pandit, P. B (1957) Nasalization, aspiration and murmur in Gujarati. *Indian Linguistics*, 17, 165–72.

Pierrehumbert, J. B. (1989) A preliminary study of the consequences of intonation for the voice source. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 4, 23–36.

Price, P. J. (1989) Male and female voice source characteristics: Inverse filtering results. *Speech Communication*, 8, 261–77.

Qi, Y. & Bi, N. (1994) A simplified approximation of the four-parameter model of voice source. *Journal of the Acoustical Society of America*, 96, 1182–5.

Rose, P. (1989) Phonetics and phonology of Yang tone phonation types in Zhenhai. *Cahiers Linguistiques Asie Orientale*, 18, 229–45.

Rosenberg, A. E. (1971) Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49, 583–98.

Rothenberg, M. (1973) A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America*, 53, 1632–45.

Rothenberg, M. (1974) Glottal noise during speech. *STL-QPSR* (Speech Transmission Laboratory, KTH, Stockholm, Sweden), 2–3, 1–10.

Rothenberg, M. (1983) An interactive model for the voice source. In D. M. Bless & J. H. Abbs (eds.), *Vocal Fold Physiology* (pp. 155–65). San Diego: College-Hill Press; also in *STL-QPSR*, Speech, Music and Hearing, KTH, Stockholm, 4/1981, 1–17.

Rothenberg, M., Carlson, R., Granström, B., & Lindqvist-Gauffin, J. (1974) A three-parameter voice source for speech synthesis. *Proceedings of the Speech Communication Seminar*, Stockholm, 2, 235–43.

Scherer, K. R. (1986) Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143–65.

Scherer, K. R. (1987) Vocal assessment of affective disorders. In J. D. Maser (ed.), *Depression and Expressive Behavior* (pp. 57–82). Hillsdale, N: Lawrence Erlbaum.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003) Vocal expression of emotion. In R. J. Davidson, H. Goldsmith, & K. R. Scherer (eds.), *Handbook of the Affective Sciences* (pp. 433–56). New York/Oxford: Oxford University Press.

Scherer, K. R., Ladd, R. D., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, 76, 1346–56.

Silverman, D. (2003) Pitch discrimination during breathy versus modal phonation. In J. Local, R. Ogden, & R. Temple (eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI* (pp. 293–304). Cambridge: Cambridge University Press.

Sluijter, A. M. C. & Heuven, V. J. van (1996) Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471–85.

Södersten, M. (1994) Vocal fold closure during phonation. Doctoral dissertation, Studies in Logopedics and Phoniatrics No. 3, Huddinge University Hospital, Stockholm.

Steriade, D. (1999) Phonetics in phonology: The case of laryngeal neutralization. *UCLA Working Papers in Phonology*, 3, 25–146.

Strik, H. & Boves, L. (1994) Automatic estimation of voice source parameters. *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, 155–8.

Stuart-Smith, J. (1999) Glasgow: Accent and voice quality. In P. Foulkes & G. Docherty (eds.), *Urban Voices: Variation and Change in British Accents* (pp. 203–22). London: Arnold.

Thongkum, T. L. (1987) Another look at the register distinction in Mon. *UCLA Working Papers in Phonetics*, 67, 132–65.

Traill, A. (1985) *Phonetic and Phonological Studies of !Xóõ Bushman*, Quellen zur Khosian-Forschung, 1. Hamburg: Helmut Buske.

Traill, A. & Jackson, M. (1987) Speaker variation and phonation types in Tsonga nasals. *UCLA Working Papers in Phonetics*, 67, 1–28.

Trudgill, P. (1974) *The Social Differentiation of English in Norwich*, Cambridge: Cambridge University Press.

Veldhuis, R. (1998) A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, 103, 566–71.

Wayland, R. & Jongman, A. (2003) Acoustic correlates of breathy and clear vowels: The case of Khmer. *Journal of Phonetics*, 31, 181–201.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2005) Voice quality and f_0 cues for affect expression: Implications for synthesis. *Proceedings of the 9th European Conference on Speech Communication and Technology, INTERSPEECH 2005*, Lisbon, 1849–52.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2006) Mapping voice to affect: Japanese listeners. *Proceedings of the 3rd International Conference on Speech Prosody*, Dresden, Germany, paper OS4-4-265, 4 pp.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2007) Time- and amplitude-based voice source correlates of emotional portrayals. *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, ACII 2007*, Lisbon, 159–70.

12 Articulatory–Acoustic Relations as the Basis of Distinctive Contrasts

KENNETH N. STEVENS AND
HELEN M. HANSON

This chapter reviews the current state of the quantal/enhancement theory of speech production. Special emphasis is placed on new insights about the physical principles that lead to quantal relations between the articulatory configuration of the vocal tract and its acoustic output. Evidence that supports these principles is discussed. In particular, we postulate that there are two sources of quantal relations, one being the principles that govern the interaction of aerodynamic forces with the vocal tract surfaces, and the other being the principle of coupled resonators. We propose that these two principles lead to (1) a natural division of the distinctive features into two groups, articulator-free and articulator-bound, and (2) natural constraints among the features within these two groups for any segment, such that the featural representation of a segment is rather sparse. We conclude by comparing quantal/enhancement theory with other theories of speech production and perception, and suggest areas of research that explore the consequences of quantal/enhancement theory for unresolved issues in English, and for languages other than English.

1 The Question: How Is the Discrete Linguistic Representation of an Utterance Related to the Continuously-Varying Speech Signal?

Speech communication involves the generation of sound by a speaker and interpretation of that sound by a listener. In preparation for the production of sound, the speaker plans the utterance in a linguistic form, one component of which is a concatenation of words that are organized into phrases and other units, both larger and smaller. The linguistic representation then initiates (possibly via the generation of additional intermediate representations) commands to the muscles that are responsible for respiration and for manipulating the various laryngeal and supraglottal structures. When the resulting sound reaches the ear of the listener, it is processed by the peripheral auditory system and by the higher levels of the auditory and speech-processing system. The output of this processing forms the basis for lexical access and, ultimately, interpretation of the utterance.

It is assumed here that words are represented in the memory of speakers and listeners in terms of sequences of segments, each of which consists of a bundle of binary distinctive features (Jakobson, 1928). The features are distinctive in the sense that changing the value of one such feature in a segment within a word creates a different potential word that contrasts with the original word. A pair of words that differ only in one feature in any one of the segments is called a minimal pair. Examples are the minimal pairs *pat/bat*, *bait/bet*, and *pat/pad*, where the feature that generates the minimal pair is in, respectively, the initial consonant, the vowel, and the final consonant. It is generally assumed that there is a universal set of such features, and any given language makes use of a subset of these universal features to distinguish among and define its words. The presence of features in the mental lexicon provides a straightforward way of defining phonological contrasts.

Evidence to support the concept of features comes in part from the field of phonology, in which the distinctive features play a fundamental role in the phonological rules that are part of the knowledge possessed by speakers of a language. For example, Chomsky and Halle (1968) compiled a number of rules involving distinctive features – rules that capture phonological patterns observed in English. A simple example is the rule for describing the English past-tense suffix, in sequences like *place-ed*, *work-ed*, *grabb-ed*, *fill-ed*, and *play-ed*. This type of assimilation is commonly characterized by postulating that the regular past-tense marker is voiced /d/, which assimilates voicelessness when affixed to a verb ending in a voiceless consonant.

Further indirect evidence for the role of distinctive features in speech comes from perception experiments that examine consonant confusions made by listeners when they identify the consonants in consonant–vowel syllables in noise and with band-pass filtering (Miller & Nicely, 1955). The results of these experiments show that the patterns of confusion are organized along featural lines. In the presence of noise and certain types of bandpass filtering, some features are poorly identified while others are robust. Any one feature is known to be identified by a listener based on multiple cues; some of the cues may be masked by noise or filtering and others may be more robust in these environments.

If one accepts the discrete linguistic representation of words, segments, and features in the lexicon, one of the central problems of research in speech science is to model the relation between this discrete and nonvarying abstract representation and the continuously varying speech signal that is produced by a speaker. That is, how can the seemingly continuous signal, output by the apparently continuous-time articulatory system, convey both the discreteness of successive segments in an utterance and the discreteness of phonological contrasts? It is natural, then, to question whether there is some link between the inventory of features that languages use and the acoustics of speech production.

The first attempt to provide a connection between distinctive features and acoustic theories of speech production was made in 1952 by Jakobson et al. The proposed distinctive features were described for the most part by perceptually oriented terms such as *strident* and *mellow*. Some years later, Chomsky and Halle (1968) proposed a revision of the inventory of distinctive features with names oriented to articulation.

Attempts to link the discrete underlying representation to the continuous acoustic signal have perhaps been held up by the assumption that the vocal tract should be modeled as a system in which both the input and output parameters vary continuously. In this chapter we argue for a different view: while the positions and movements of the articulators (and other mechanisms), i.e., the input parameters, vary more or less continuously, the output parameters of the system have discrete aspects. Such a model provides a way to neatly link the discrete features to acoustics. Evidence for this claim is given in the form of articulatory–acoustic relations that have quantal properties (section 2). In addition to providing a basis for discreteness in sequencing (segments), it is claimed that the quantal nature of the relationship between the speech production mechanism and its acoustic output forms the basis for the discreteness of phonological contrasts. That is, each distinctive feature in the universal set is assumed to be grounded in a particular "defining" acoustical/perceptual consequence of an articulatory configuration or gesture. Note that we do not claim that everything in the acoustics is quantal. Rather, our claim is that some aspects of the system and its output (or the perception of its output) are quantal, and these quantal aspects can be linked to the emergence of a discrete representation of features and segments in the lexicon.

Furthermore, quantal theory and the notion of defining gestures coupled with defining acoustic cues for each contrast do not imply invariance in the signal when a token of one of these contrastive categories is produced. It is well known that the acoustic signal is highly variable. Although it could be argued that this variability in speech production is evidence against universal features and even segments, we claim the reverse: that there exists an underlying quantalness combined with systematic and contextually predictable modifications, that (taken together) result in a highly variable acoustic speech signal that nevertheless successfully specifies the invariant contrastive feature categories. Because the modifications are systematic, the underlying discrete representation should be recoverable from the acoustic signal.[1]

There are many sources of systematic variation in the cues to distinctive feature contrasts. For example, the variation among speakers in vocal-tract characteristics such as length will result in variation of the defining acoustic correlates. In addition, we note in section 3 that, during the production of a word or sound in context, the defining gestures and acoustic correlates associated with the features of a segment may be weakened or even obliterated, due to overlap of gestures from adjacent segments or to prosodic influences. This overlap is another source of variability.

Further variability arises when additional acoustic attributes are introduced in certain contexts through the action of articulators other than the defining ones. Some of these actions may be introduced in order to enhance the perceptual contrast carried by the basic or defining gestures. These enhancing gestures may simply amplify a contrast, or may shore up a contrast that is weakened by the previously mentioned overlap. It is important to note that enhancing gestures and their associated acoustics (1) need not be quantal in nature, (2) are not subject to weakening or obliteration (as are defining gestures), and (3) are not secondary in any way to defining gestures in the perceptual process; they simply function differently than defining gestures, and it is this difference that gives them their explanatory power. The differences between features and their defining gestures, on the one hand, and enhancing gestures, on the other, is crucial to understanding or modeling some of the phenomena we observe in speech.

It was proposed by Clements (1985) that the features of a speech segment are not just a simple list without structure, but are organized into a hierarchical tree structure. This tree structure was an attempt to account for the observed limitations on types of phonological assimilation. Clements' proposal was expanded by Halle and Stevens (1991), who observed that an anatomical organization of features was also necessary to account for different phonological rules. The constraints imposed by the hierarchical organization of features mean that not all features need to be defined for every segment, resulting in "sparse" definitions of segments. We argue in section 2.1.4 that the features can also be organized according to constraints imposed by the physics of the speech-production mechanism. A topic for future research is to determine whether such an organization will converge with an organization based on phonological observations.

In sum, we suggest in this chapter a model of speech production in which:

1 The phonological representation of a word consists of a sequence of segments, with each segment formed by a bundle of distinctive features. Because there are constraints on the combinations of features that can form a bundle, the number of features that define each segment is sparse.
2 The features of a given language are drawn from a universal set of features, each of which is specified by the relationship between a defining gesture and an acoustic correlate; this relationship is based on the quantal nature of the speech production mechanism.
3 Feature-defining gestures and their associated acoustics can be imperiled by overlap with neighboring segments.
4 Enhancing gestures, which are particular to a given language, are introduced to enhance contrasts or to shore up contrasts that are weakened or obliterated by overlap.

In our concluding remarks (section 4), we will contrast the quantal/enhancement theory with other models of speech production.

2 One Answer: Quantal Theory and Distinctive Features

Sounds produced by the vocal tract can be described in terms of a number of acoustic parameters which change as the positions and states of the various

Table 12.1 A list of distinctive features discussed in this chapter. The features are divided into two groups, articulator-free and articulator-bound.

Articulator-free	*Articulator-bound*
[sonorant]	[high]
[consonantal]	[low]
[continuant]	[back]
[syllabic]	[round]
[glide]	[labial]
[strident]	[coronal]
[delayed release]	[anterior]
[spread glottis]	[velar]
[constricted glottis]	[nasal]
	[rhotic]
	[lateral]
	[tense]
	[stiff vocal folds]

articulators are manipulated. The articulatory structures that are controlled by speakers, and the states of these structures when they produce speech, appear to be capable of being varied through a continuous range, although there are endpoints in these movements as the structures come in contact with fixed surfaces or as the displacements reach the ends of their ranges. As discussed in section 1, the linguistic representation of speech sounds, for both speakers and listeners, is assumed to be based on distinctive features that (we argue) arise from certain discontinuities in the articulatory–acoustic relationship. In Table 12.1 we give the set of distinctive features discussed in this chapter. We note here that these features are a subset of some universal set, from which are drawn the features on which any spoken language is based. We stress at this point that, while we may make reference to other features, our discussion will focus on English, the American dialect in particular, because this is a language that has been well studied. It is our belief, however, that the theories presented here can be extended or adjusted to apply to the universal set of distinctive features and all of the languages based on them.

The features in Table 12.1 have been divided into two groups, the articulator-free and articulator-bound features, as suggested by Halle (1992). Features in either set have their origin in particular articulatory actions that give rise to basic acoustic and perceptual attributes. In the case of articulator-free features, the articulatory actions are classified in terms of the type of constriction or narrowing that is produced in the vocal tract, without specifying which articulator is creating the constriction. Examples are the features [sonorant], [continuant], and [strident]. Articulator-bound features, on the other hand, specify which articulator forms the constriction, how that articulator is shaped or positioned, and the actions of other articulators which do not themselves create the constriction but which influence the acoustic pattern that emerges when the constriction is formed. The features [anterior] and [back] are examples of articulator-bound features. It should be said that the assignment of features to the articulator-free or articulator-bound categories is in flux, and remains a topic of some debate.

As briefly reviewed in section 1, various attempts have been made to link the distinctive features with the acoustic properties of speech production. In recent years a "quantal" concept of speech production has been developed (Stevens, 1972, 1989; Stevens & Keyser, in press), showing that a feature can be defined by a quantal relation between an articulatory parameter and an acoustic parameter. As the articulatory parameter is manipulated through its range of values (and other parameters are kept constant), the changes in the acoustic parameter are often not monotonic. That is, within some regions of articulatory space, the acoustic parameter is relatively insensitive to articulatory change, whereas outside these regions the acoustic parameter is quite sensitive to articulatory change. This concept is illustrated in Figure 12.1, in which one can see that setting the articulatory parameter to be within either of ranges I or III will result in an acoustic parameter with a relatively constant value, while setting the articulatory parameter to be within range II results in an acoustic parameter that changes rapidly as the articulatory parameter moves through that range.

The property that there are articulatory regions where the acoustic attributes are relatively stable and other regions where the acoustics undergo significant changes for small articulatory differences suggests that the human vocal tract and its associated acoustic sources are characterized by a set of articulatory–acoustic "states." It is hypothesized that these states provide some basis for the two types of discreteness noted in section 1, segmental and phonological. That is, as an articulatory structure is displaced through a range of positions or configurations during continuous speech, there will be points in time where acoustic discontinuities or dislocations occur, and other points where there are extrema (maxima or minima) in some acoustic parameters such as formant frequencies. These points

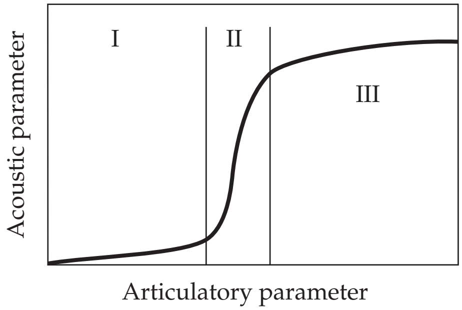


Figure 12.1 Hypothetical acoustic–articulatory relation showing two relatively stable regions (I and III) and a region II where there is a rapid change in an acoustic parameter for a relatively small change in the articulatory parameter. (From Stevens, 1989, copyright Elsevier)

in time can be regarded as sequential markers for segmental units. The quantal acoustic properties in the vicinity of these discontinuities or extrema help to define the features of these segmental units. The relatively stable acoustic region defines the acoustic and articulatory parameters for a feature, and the region where there is an abrupt acoustic change is a region that is avoided in the defined implementation of a feature. The acoustic attribute that is stable in such a relation between articulation and sound provides a "cue" to a potential listener for identifying the feature. For example, in the first segment of the word *ben* the defining acoustic cue for the feature [labial] is an acoustic energy burst with a particular spectrum shape (relatively flat in this case).

2.1 Where quantal relations come from

There are two general physical properties of the speech production system which generate the special kinds of articulatory–acoustic relations that give rise to the various distinctive features. One of these properties stems from interaction between airflows and pressures on the one hand and mechanical properties of articulatory structures on the other hand. The other physical property is the result of acoustic coupling between resonators. These properties will be referred to as AMI (aero-mechanical interaction), and as ARC (acoustic resonator coupling).

Aero-mechanical interactions arise because the nature of the interaction of airflow with the compliant mechanical structures that form the surfaces of the vocal tract can change abruptly as an articulatory parameter changes continuously. As a result, the nature of the generated acoustic source changes abruptly. Examples of AMIs are given in section 2.1.1.

Acoustic resonator coupling is due to the nature of the acoustic filter formed by the vocal-tract structures. If the structures simply formed a single cavity, e.g., something approximating a lossless uniform tube, the transfer function from an acoustic source to the vocal-tract output would vary smoothly as the length of the vocal tract varied or as the vocal-tract shape changed, within limits. However, the vocal tract, together with adjacent structures, can create several cavities that can be coupled and uncoupled, particularly if narrow constrictions separate these cavities. As individual cavities couple or uncouple the vocal-tract transfer function can show an abrupt discontinuity as a consequence of the rapid movement of a zero in it. Examples of ARC are given in section 2.1.2.

The division of features into those defined by AMI and those defined by ARC coincide roughly with Halle's (1992, 1995) division of features into articulator-free and articulator-bound, respectively (Table 12.1). The features that are defined by AMI are related largely to manner of articulation. Features that stem from ARC are features specifying place of articulation. Because the features in the two groups are defined by different physical principles (AMI vs. ARC), relations among the features in these groups are quite different: the articulator-free features, being based on aerodynamic conditions in the vocal tract, are constrained into a hierarchy, while there are fewer constraints among the articulator-bound features. Because of the various constraints among features (discussed in section 2.1.4), the featural representation of any segment or feature bundle is relatively sparse, sometimes requiring only 4–6 features to be defined, rather than the 20 or so features that are used across all segments in English.

Each of these two general physical principles, AMI and ARC, has several subclasses. In the following sections, an overview of these classes is given, using simple models for illustration. A more quantitative analysis of some examples drawn from some of these subclasses is also reviewed, together with acoustic data.

It should be noted that quantal relations have not yet been studied for all features that play a role in defining phonological contrasts in English, and certainly there is a need to examine features in other languages. For example, there are languages in Australia that appear to have four different "places of articulation" for coronal fricatives (Butcher, 2006); to account for such contrast patterns, possible quantal relations that include not only place of articulation but also tongue-blade shape need to be examined.

In addition to the quantal relations observed between articulatory and acoustic parameters, a similar type of quantal relation is also observed between certain acoustic parameters and some aspects of the auditory responses to them. That is, as an acoustic parameter is manipulated, there are abrupt changes in the auditory response for certain values of the parameter. Thus the articulatory system for generating speech can be regarded as a generator of sounds that lead to categorical acoustic and perceptual properties. We do not, however, discuss acoustic–auditory relations in this chapter.

2.1.1 Aero-mechanical interactions (AMI) The articulator-free features listed in Table 12.1 have their origin in abrupt changes that occur in the aerodynamics, and thus acoustics, of the vocal tract as the articulatory configuration changes smoothly. With all speech sounds, the respiratory system creates a pressure differential between the air inside the lungs and that outside the lungs. This pressure differential results in airflow through the cavities that make up the vocal tract. In most cases speech is produced during an exhalation, i.e., when the lung pressure P_S is positive relative to atmospheric pressure. Different configurations of the articulators will result in different aerodynamic conditions, as the airflow is shaped by factors such as whether there are constrictions, which articulators form the constrictions, and the nature of the constrictions.

The basic aero-mechanical interaction can be illustrated by a simple model in which a pressure P_S, representing the subglottal pressure, is applied at one constricted end of a tube, as shown in Figure 12.2a. The right end represents the output of the model at the lips. The walls of the constricted portion of the tube, representing the laryngeal area, have certain physical properties that can be modeled by an impedance consisting of a simple acoustic compliance, mass, and resistance. For this simple model the air pressure in the larger volume is essentially atmospheric pressure. This action of creating the condition of zero pressure relative to atmospheric in the oral cavity is in effect the defining articulator action for the feature [+sonorant]. The defining acoustic attribute is a sound source only in the vicinity of the glottis. Through various manipulations of the dimensions

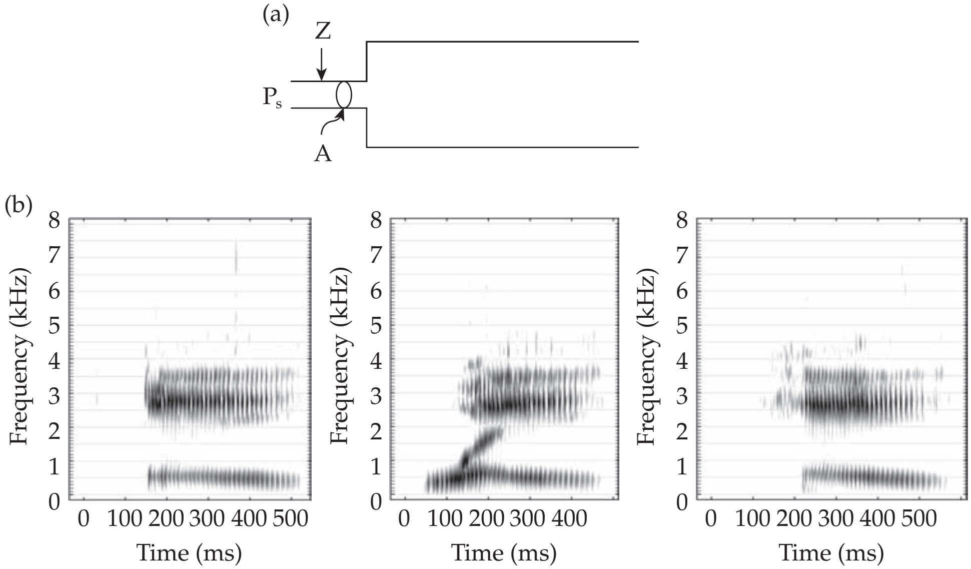


Figure 12.2 (a) A model of the basic aero-mechanical interaction, in which a tube is constricted at one end. The constricted portion represents the vocal folds, P_S the subglottal pressure, Z the impedance of the vocal folds, and A the cross-sectional area between the folds. The lack of a narrow constriction in the larger volume means that the air pressure there is essentially atmospheric pressure, and is, in effect, the defining articulator action for the feature [+sonorant]. (b) The spectrograms show three utterances (produced by a male speaker) that represent different configurations of the glottal source in a syllable with the vowel [e]: a glottal stop, with an abrupt, somewhat irregular onset (left); the word/we/ illustrating a glide (middle); and the word /he/ illustrating a spread glottis (right).

of the constriction and the physical properties of the surfaces of the constriction, the vocal folds can be caused to respond in different ways that have different quantal characteristics: (1) a constriction in the airway downstream from the glottis is not narrow enough to produce a significant sound source due to turbulence in the air stream, and thus the vocal folds vibrate normally (the feature [+sonorant]); (2) vocal-fold vibration does not occur but aerodynamic noise is generated within or downstream from the constriction (the feature [+spread glottis]); or (3) the constriction size is reduced to zero and then released to generate a stop-like acoustic property corresponding to a glottal stop (the feature [+constricted glottis]). All three of these configurations satisfy the articulatory definition of the feature [+sonorant] given above. The spectrograms in Figure 12.2b show three utterances representing different configurations of the glottal source in a CV syllable with the vowel [e]: the syllable/ʔe/, with an abrupt, somewhat irregular onset (left); the syllable /we/illustrating a glide with normal (modal) vibration (middle); and the syllable /he/illustrating a spread glottis (right).

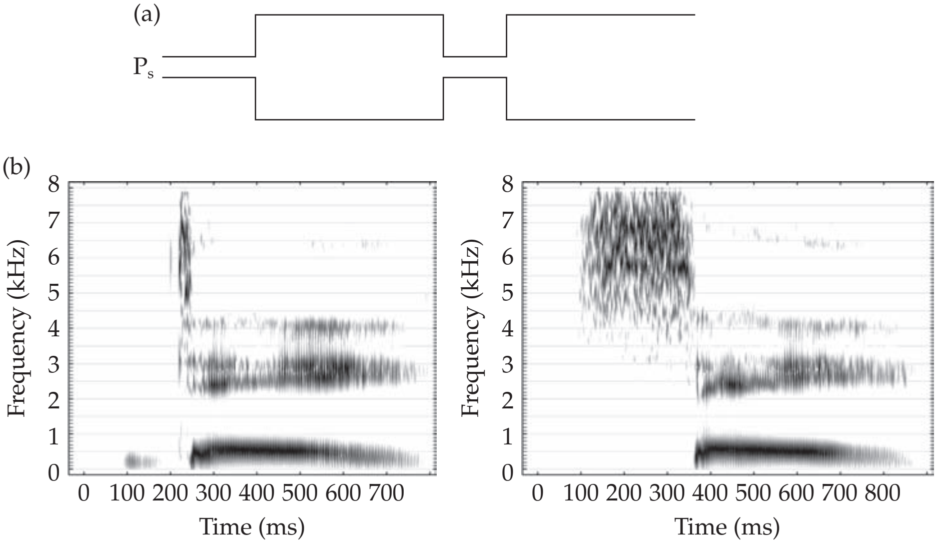


Figure 12.3 (a) The simple model of Figure 12.2 is expanded by introducing a narrow constriction in the large volume, dividing this space into two smaller volumes. Pressure builds up behind the constriction to produce turbulence noise. The formation of such a constriction is the defining articulator action for the feature [–sonorant]. (b) The spectrograms illustrate two consonants that are [–sonorant]: a stop consonant in the syllable *day* (left) and a fricative consonant in the syllable *say* (right), produced by a female speaker.

A greater degree of complexity is obtained if the simple model in Figure 12.2 is expanded by introducing a second narrow constriction downstream from the first constriction, as in Figure 12.3a. This model can be manipulated in a number of different ways that lead to distinctively different acoustic outputs, all of which can be classified as [–sonorant]. The variable parameters (in addition to those in Figure 12.2a) include the dimensions of the space between the two constrictions, the physical properties of the walls of the second constriction, and the dimensions of this constriction. The quantal states that can be achieved by this configuration include the following: (1) the production of voiceless or voiced fricative consonants; (2) the production of voiceless or voiced stop consonants; (3) voiceless and voiced aspirated stop consonants; and (4) certain trills and taps. In addition, the place of articulation can be changed by adjusting the location of the constriction, as will be discussed in section 2.1.2. The spectrograms in Figure 12.3b illustrate two consonants that are [–sonorant]: a stop consonant in the syllable *day* (left) and a fricative consonant in the syllable *say* (right).

The contrast between the acoustic behavior for the configurations in Figures 12.2a and 12.3a illustrates the quantal differences between the features [+sonorant] and [–sonorant]. Pressure builds up behind the constriction in Figure 12.3a to produce turbulence noise in front of the constriction, but there is no constriction

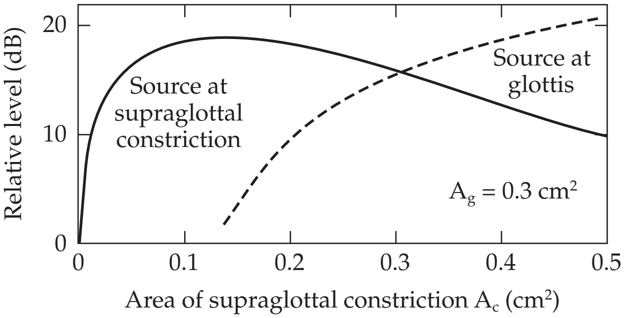


Figure 12.4 Calculated relative levels at the two constrictions in the model of Figure 12.3. The cross-sectional area of the glottal constriction A_g is fixed at 0.3 cm^2. The level of the noise source at the supraglottal constriction increases from zero to a maximum in the range from below 0.1 to about 0.2 cm^2. (From Stevens, 1998, reprinted by kind permission of MIT Press)

in Figure 12.2a (except for the glottal constriction). Within the configuration of Figure 12.3a, there is a contrast between complete closure of the constriction for the feature [–continuant] (stop consonant) and a narrow constriction for [+continuant] (fricative consonant) (Stevens, 1998; Stevens & Keyser, in press). This distinction between a [–continuant] segment and a [+continuant] segment is illustrated in Figure 12.4. For the stop consonant, there is complete closure and no airflow, with an abrupt release. For the fricative, on the other hand, air flows through the consonantal constriction as well as through the glottal opening. Turbulence noise is generated at the consonantal constriction and also at the glottal constriction. For maximum noise amplitude at the consonantal constriction, its area should be adjusted to be roughly equal to or somewhat less than the glottal constriction. This adjustment should lead to a plateau in the articulatory–acoustic relation that underlies the feature [sonorant] when the consonantal constriction is narrow, and the feature [continuant] when the constriction approaches completeness.

In some languages, [spread glottis] (sometimes called aspiration) is a distinctive feature, with quantal defining attributes. In those languages, the defining acoustic attribute is an interval of turbulence noise generated primarily at the glottis, extending past the frication noise interval for the consonant. In English, aspiration is present for stop consonants in some contexts, but because [spread glottis] is not a distinctive feature in English, this aspiration is considered to be an enhancing attribute. (See section 3 for further discussion of enhancing gestures vs. defining gestures.)

Trills and taps are produced with the tongue blade as the active articulator, with its stiffness and shaping adjusted to produce the appropriate interaction between the airflow and the compliance and shaping of the tongue blade. These actions may be distinctive in some languages, but not in English.

An additional smaller set of possible quantal articulatory–acoustic relations due to AMI can be simulated by adding still another constriction downstream from

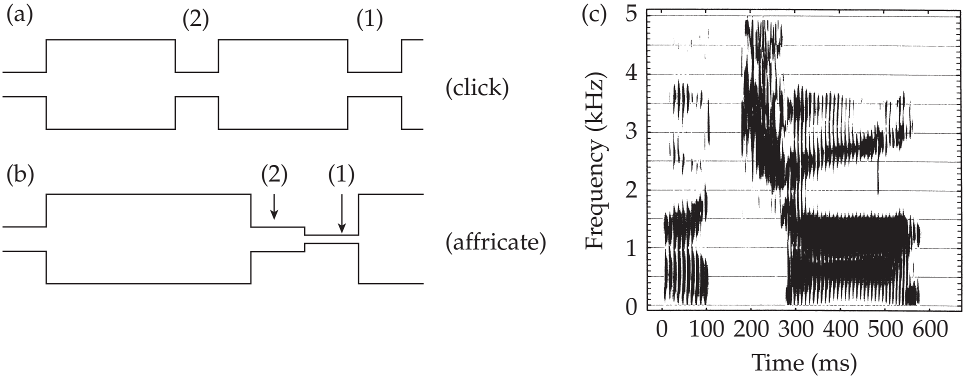


Figure 12.5 Addition of a second narrowing in the supraglottal cavity leads to a set of quantal relations, in which one of the constrictions is released later than the other. This configuration is the basis of the feature [delayed release]. (a) Velar closure at (2), a further narrowing downstream at (1), and expansion of the volume between (1) and (2) leads to production of a click when constriction (1) is released prior to constriction (2). (b) The two constrictions are formed one next to the other. Constriction (1) is released prior to constriction (2), resulting in a noise burst followed by frication, leading to production of an affricate. (c) Spectrogram of a voiceless affricate in English (see text). (Spectrogram from Stevens, 1998, reprinted by kind permission of MIT Press)

the two constrictions in Figure 12.3a, as in Figures 12.5a and 12.5b. The inclusion of a third component in the vocal-tract airway with appropriate characteristics for each of these three components simulates the production of clicks observed in !Xóõ and some other African languages (Traill, 1985; Ladefoged and Traill, 1994). These sounds are produced with a velar closure (the second constriction in the model) and with an expansion of the volume of the oral cavity to produce a large negative pressure in this cavity, as in Figure 12.5a. Some clicks are produced with a complete closure of the most anterior constriction (usually with the tongue blade) followed by a release, whereas others are produced with this constriction narrowed, leading to a click with frication noise. The velar constriction is opened after the release of the tongue blade.

Another action is shown in Figure 12.5b, which is similar to that in Figure 12.5a, but with quite different dimensions and placements for the second and third constrictions in the model. This configuration simulates the articulatory and aerodynamic events for affricate consonants. In English, affricates are distinctive only with an alveolar-palatal place of articulation. There is also a voicing distinction for this affricate. Affricates are more frequent in some other languages, sometimes with several places of articulation, including a labio-dental place /pf/ and a dental-alveolar place /ts/.

A spectrogram of a voiceless affricate in intervocalic position in English (Figure 12.5c) illustrates the sequence of acoustic events, and provides some insight into the articulatory sequence that gives rise to these events. The consonant is produced by initially closing and then releasing the anterior constriction labeled (1) in Figure 12.5b. The release is characterized by an initial noise burst whose spectrum reflects the resonance of the cavity anterior to the initial tongue-blade closure. In this example, the trajectory of the frequency of a spectrum peak following the initial noise burst reflects this movement of the front-cavity resonance (F3 or F4) from about 3,500 Hz to about 2,300 Hz. Following this phase, the movement of F3 becomes slower, and there is a brief interval of frication noise at this value of about 2,300 Hz before the onset of voicing.

As suggested by the model in Figure 12.5b and by the spectrogram, two articulatory components appear to be involved in the production of an affricate. One is the articulator labeled (1) in Figure 12.5b. This articulator forms the initial release. The other "articulator," which is posterior to the initial one, is shaped to produce the fricative portion of the release. This two-articulator group is apparently controlled with a single gesture from front to back, from which the two-component acoustic output emerges (Stevens, 1993).

The affricate consonant contrasts with both the [–continuant] stop and the [+continuant] fricative. The feature that specifies this contrast has been called [delayed release] (Chomsky & Halle, 1968). A quantitative articulatory–acoustic specification of this feature and a clear description of a quantal articulatory–acoustic relation are topics for future research.

2.1.2 Acoustic resonator coupling (ARC) We turn now to the second mechanism that is potentially responsible for quantal articulatory–acoustic relations: acoustic resonator coupling. As an introduction to this topic, it is noted first that the acoustics of the production of non-nasal vowels is usually viewed simply as the excitation of an all-pole transfer function by a quasi-periodic glottal source. Consequently, articulatory changes simply create movements of the formants without abrupt acoustic changes that could potentially define a quantal articulatory–acoustic relation. Close analysis, however, shows that there are a number of speech sounds, including vowels, for which the transfer function includes zeros as well as poles. The presence of zeros in the transfer function creates spectra that can exhibit abrupt changes that are potential sources of quantal relations. Four examples of such zero-induced relations are discussed here.

(1) A fixed cavity coupled to a variable cavity via a short, narrow tube: Non-nasalized vowels. The resonances of a fixed cavity (in this case, the subglottal resonances) are coupled through the glottis to the resonances of the vocal tract proper, as schematized in Figure 12.6. For an adult speaker the lowest three subglottal resonances are about 600 Hz, 1,400 Hz, and 2,100 Hz, on average (Cranen & Boves, 1987). When the vocal-tract resonances interact with those of the subglottal system, there can be significant modification to the spectrum expected for the vocal-tract resonances (Hanson & Stevens, 1995). This configuration is an example of one in which a fixed subglottal impedance is below the glottal source, and exerts an influence on the transfer function of the vocal tract, particularly when a vocal-tract formant (i.e., a natural frequency) is close to a subglottal resonance.

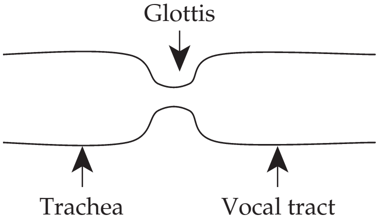


Figure 12.6 Tracheal resonances are coupled to vocal-tract resonances through the glottis, leading to acoustic effects underlying the features [low] and [back].

We focus first on the influence of the second subglottal resonance, F2sub. This second subglottal resonance contributes to a definition of the feature [back] for vowels. An explanation of the relation of F2sub to quantal aspects of the vowel feature [back] can be developed by examining the articulatory−acoustic relation when the articulation of a vowel is manipulated from a backed position to a fronted position, as in the diphthong in the word *hide*. If there were no acoustic coupling between the vocal-tract resonances and the subglottal resonances, the vocal tract formant F2 would rise smoothly and continuously, as in the straight line in Figure 12.7a. However, when there is acoustic coupling, experimental data and theoretical analysis show that its effect on the frequency and amplitude of the F2 prominence can be significant when F2 is in the vicinity of F2sub. The effect of the coupling on the F2 frequency is schematized in Figure 12.7a. As F2 passes near F2sub (1,400 Hz in this case), the frequency of the F2T prominence makes an abrupt change, as shown by the short-dashed line. There is also an abrupt dip in the amplitude of the F2T prominence (not shown in the figure). Experimental data from a number of speakers show shifts of the order of 100 Hz at this point, and dips in amplitude of several dB (Chi & Sonderegger, 2007). Thus the frequency and amplitude of the F2 prominence are somewhat unstable as the tongue-body position passes through a region in which the original F2 is close to F2sub. It is hypothesized that this region of F2 instability corresponds to a boundary that separates vowels that are [+back] from those that are [–back]. Examination of vowel systems for several speakers of English, together with measurements of the subglottal resonances for those speakers, shows that F2 for the monophthongal vowels tends not to occur in this region (Sonderegger, 2004; Chi & Sonderegger, 2007).

A similar influence has been observed when F1 for a vowel is in the vicinity of the first subglottal resonance F1sub, which is at about 600 Hz. Experimental data show that this region of F1 tends to be avoided by vowels in English: for [+low] vowels, F1 is higher than F1sub, and for [–low] vowels, F1 is lower than F1sub (Jung & Stevens, 2007). Figure 12.7b shows formant trajectories for the word *bide* produced by a female speaker. There are two breaks indicated, in the trajectories for both F1 and F2. Both of these breaks occur in the vicinity of where the first two subglottal resonances occur for female speakers.

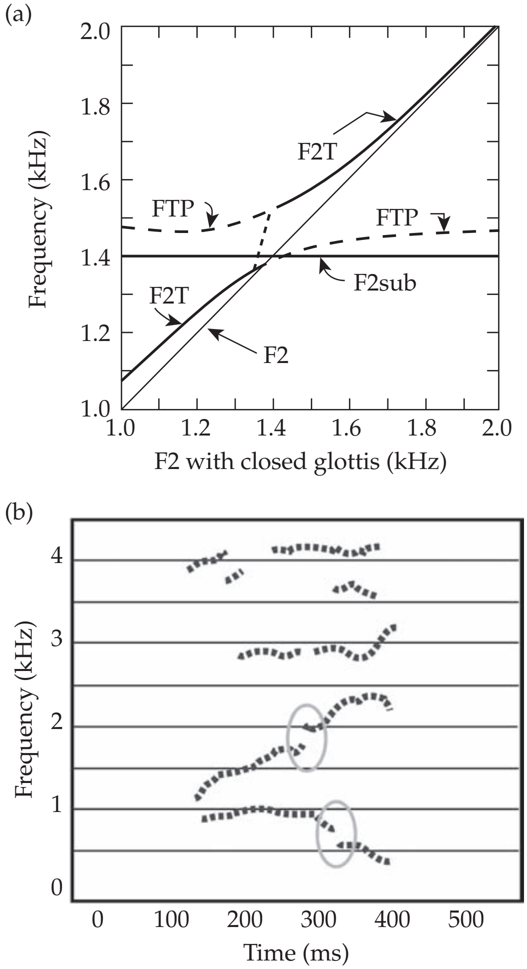


Figure 12.7 (a) Calculations of the frequencies of the two poles and zero in the vicinity of F2 when there is coupling to the trachea through a partially open glottis. The abscissa (F2 with closed glottis) is the formant frequency that would exist if the glottis were closed and there was no acoustic coupling to the trachea. F1 and F3 are held constant at 400 and 2,500 Hz. The frequency of the tracheal zero (F2sub) is assumed to be fixed at 1,400 Hz. The thin straight line labeled F2 represents the pole corresponding to F2 with no coupling to the trachea. The heavy solid line F2T represents the pole corresponding to F2, shifted by the influence of the tracheal system. FTP is the tracheal pole. The frequency of the most prominent spectral peak (F2T) shows an abrupt jump in frequency (short-dashed line) when F2 is just below 1,400 Hz. The long-dashed lines (FTP) represent the poles that are less prominent in the spectrum (adapted from Stevens, 1998, reprinted by kind permission of MIT Press). (b) Measured trajectories of F1 and F2 (and higher formants) for a female speaker saying *bide*. Two discontinuities are circled representing the influence of subglottal resonances. F1sub and F2sub are at about 600 Hz and 1,800 Hz, respectively.
(Adapted from Hanson & Stevens, 1995.)

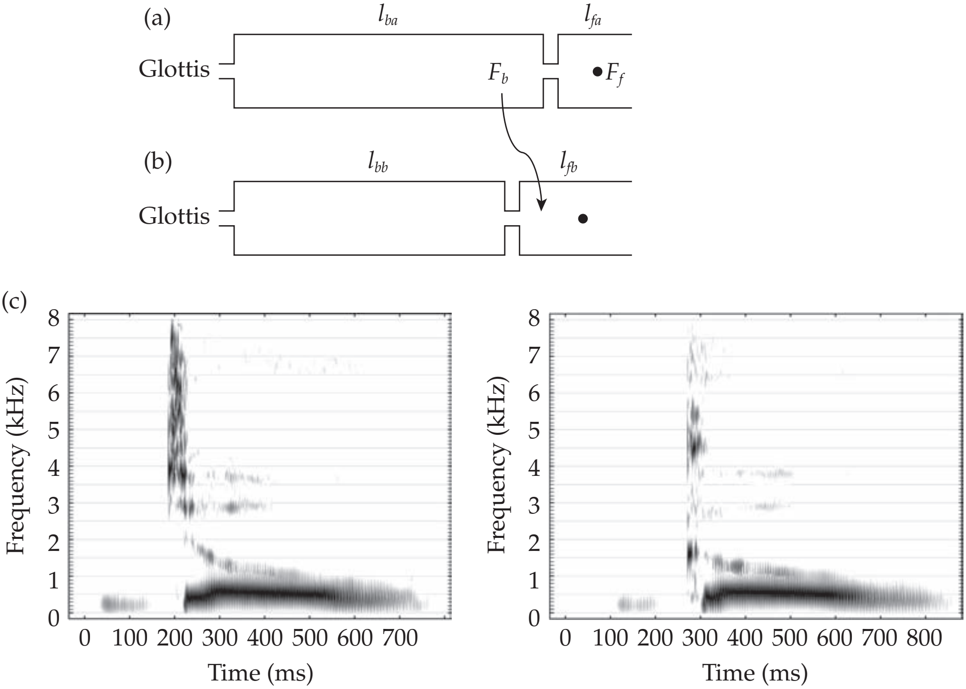


Figure 12.8 As a supraglottal constriction shifts back in the vocal tract, back-cavity resonances move to the front cavity, and their poles are no longer obscured by zeros. This shift in cavity affiliation is indicated by the arrow linking part (a) to (b). (c) The spectrograms illustrate the shift of back cavity resonances to front. On the left is a spectrogram for the syllable [do] produced by a female speaker, and on the right is the syllable [go] produced by the same speaker. F2 is not prominent in the burst following the alveolar release, but it is following the velar release.

In both of these examples relating to the features [back] and [low], a subglottal influence that appears to be small over much of the frequency range for F1 and F2 can have a strong influence on the placement of acoustic prominences corresponding to the vowel formants.

(2) Two variable cavities coupled by a short, narrow tube: Obstruents. A second configuration that can involve rapid spectrum changes due to displacements in the components of pole–zero pairs is sketched in Figure 12.8a–b. A supraglottal narrowing results in a noise source downstream from the narrowing. Back cavity resonances are only weakly excited by the source, i.e., poles or natural frequencies associated with the back cavity are obscured by adjacent zeros. As the location of the constriction moves forward in the oral cavity, back-cavity resonances shift to the front cavity, and associated poles are no longer obscured by zeros. This configuration is an example of quantal relations arising from concomitant changes

of back-cavity and front-cavity lengths (e.g., from alveolar to velar consonants). This process in which poles and zeros are manipulated by shifting the place of articulation plays an important role in defining quantal aspects of place of articulation for obstruent consonants.

Figure 12.8a is a schematic representation of a vocal-tract shape for an obstruent consonant with a constriction near the front of the vocal tract. The back-cavity length l_{ba} is much greater than the front-cavity length, as might be expected for an alveolar consonant. There is a front-cavity resonance F_f whose frequency is approximately $c/(4l_{fa})$, where c equals the velocity of sound. For example, if $l_{fa} = 2.5$ cm, as it might be for an alveolar stop or fricative, then F_f would be about 3,500 Hz. There is usually a strong spectrum peak at this frequency in the sound output, constituting an important acoustic cue for perception of place of articulation. The next lowest natural frequency in Figure 12.8a, F_b, is associated with the cavity behind the constriction, which has a length l_{ba}. This frequency is excited only weakly by the turbulence noise source downstream from the constriction; that is, in the transfer function from source to output there is a zero that is very close in frequency to this back cavity resonance. If the length l_{fa} of the front cavity is now increased, as in Figure 12.8b, the frequency of the front cavity resonance decreases. At some point during this increase in l_{fa}, a pole F_b from the upstream portion of the vocal tract appears now as a downstream spectrum prominence. This shift of affiliation from back cavity to front is indicated by the arrow between Figures 12.8a and 12.8b. Another way of stating this change is that the zero that was formerly close to the pole at F_b in the transfer function is now shifted away from the pole. The prominence in the acoustic spectrum is thus shifted downwards in frequency. This abrupt shift in the acoustics as the length of the front cavity increases constitutes a "quantal" change in the output spectrum, and defines a change in a distinctive feature for place of articulation for an obstruent consonant (Stevens, 2003). If in the example given above the front-cavity length increases by 1 cm, and if what was F_b now becomes a front-cavity resonance, its frequency is $c/(4 \times 3.5) = 2{,}500$ Hz. This increase in l_{fa} would put this configuration in the range of a velar stop or fricative instead of an alveolar consonant. Therefore, there are clear articulatory regions which speakers can reliably use for place features, and other regions which speakers should avoid, and these delineate features such as [alveolar] or [velar]. Figure 12.8c shows spectrograms for the syllables [do] and [go] produced by a female speaker, illustrating the shift of back cavity resonances to front. On the left is a spectrogram for the syllable [do], and on the right is the syllable [go] produced by the same speaker. F2 is not prominent in the burst following the alveolar release, but it is following the velar release.

(3) A fixed cavity coupled to a variable side branch: Nasal consonants. The places of articulation for nasal consonants in English are similar to those for stop consonants. The articulatory events are not, however, identical since aerodynamic forces on the articulators for the [–sonorant] stop consonants have an effect on the details of the articulator movements (Stevens, 2001), whereas there is no increase in intraoral pressure for the nasals. The acoustic descriptions for the place features

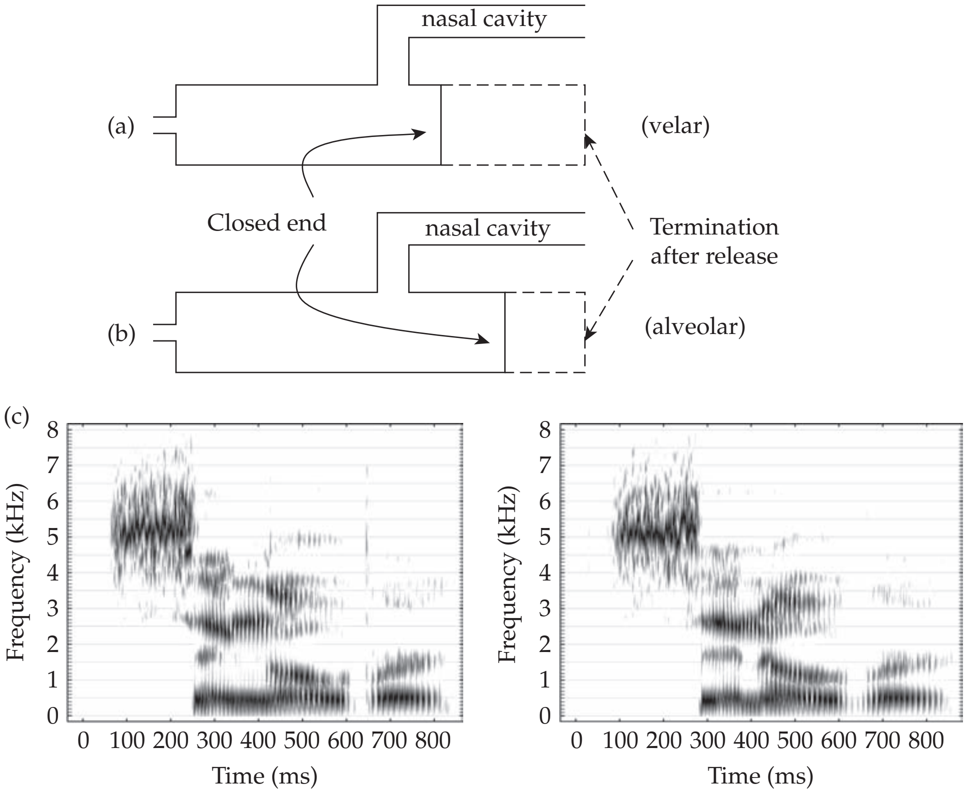


Figure 12.9 A model for nasal consonants, for two places of articulation: (a) velar, and (b) alveolar. (c) Spectrograms illustrating the alveolar (left) and velar (right) nasal consonants in intervocalic position. The F2 onset for postvocalic /n/ is lower than that for postvocalic /ŋ/, indicating the larger back cavity. The high-frequency energy onset (above 3,000 Hz) at the release for /n/ is higher than for /ŋ/. These indicate the place of articulation.

are quite different for the two classes of consonants: nasal consonants are [+sonorant] with no turbulence and stop consonants are [–sonorant].

The acoustic events that define the place-of-articulation feature for a stop consonant are the spectral characteristics of the noise burst at the consonant release. For a particular place of articulation, these spectral characteristics arise from the abrupt excitation of a particular natural frequency (e.g., the third or fourth formant) at the end of a silent interval.

The model for nasal consonants is shown in Figures 12.9a–b. The connection between the vocal tract and the nasal cavity is represented, including the output at the nose and the closure in the oral cavity. Two places of articulation are depicted: a velar nasal in Figure 12.9a and an alveolar nasal in Figure 12.9b. For these consonants, the front of the oral cavity is not excited during the oral closure for the consonant, and direct acoustic information about the length of the front cavity is not available in the acoustic signal.

Defining acoustic information concerning the place of articulation for a nasal consonant is expected to be centered on the sequence of acoustic events that occur as the consonant release occurs, much like the acoustic events at the release of a stop consonant. For a nasal consonant, this sequence consists of the initial release of the consonant closure, as labeled in Figure 12.9, coordinate with the introduction of acoustic excitation of the previously quiescent front cavity (shown by the dashed lines in the figure). The initial release involves the rapid shifting of a zero in the transfer function as the opening to the front cavity is released. The front cavity, which was previously characterized by a coincident pole–zero pair now becomes acoustically excited as the zero shifts abruptly from its previous position that canceled the pole determined by the front-cavity resonance.

This way of describing a nasal consonant release does not propose well-specified quantal articulatory–acoustic relations that describe the defining acoustic attributes for place of articulation features. However, much as the spectral characteristics of the noise burst following release define the place of articulation for stops, we expect that spectral prominences that arise immediately following the release of nasal consonants will define place features for nasals. And, as for stops (and fricatives), we expect that the places of articulation will be delimited by regions rendered unstable by the rapid changes in frequency and amplitude of the prominences. These speculations have yet to be examined in detail. Spectrograms in Figure 12.9c illustrate the alveolar (left) and velar (right) nasal consonants in intervocalic position. The utterances are "sin over" and "sing over," produced by a male speaker. The F2 offset for postvocalic /n/ (at about 420 ms) is lower than that for postvocalic /ŋ/ (also at about 420 ms), indicating the larger back cavity. The high-frequency onset (above 3,000 Hz) at the release for /n/ is higher than for /ŋ/. These differences following the nasal release are similar to those discussed above for stop releases.

(4) A variable cavity coupled to a fixed side branch: Nasalized vowels and approximants. The fourth type of vocal-tract configuration that illustrates quantal properties based on acoustic resonator coupling involves sonorant consonants or vowels with side branches in the area function (Figure 12.10a). In English, this class of segments consists of a rhotic consonant, a lateral consonant, and nasalization of a vowel. Nasalization is not an attribute that is contrastive for vowels in English, but it occurs frequently as an "enhancing" gesture that provides cues for the presence of a nasal consonant. In many languages, of course, there is a distinctive feature [+nasal] that contrasts nasal and non-nasal vowels. A common attribute that distinguishes these three types of segments from non-nasal vowels is that the transfer function from the glottal volume-velocity source to the output volume velocity contains zeros as well as poles. For a non-nasal vowel, the transfer function is all-pole (ignoring the fixed zeros introduced by the subglottal resonances). A consequence is that the amplitudes of some formant prominences differ from the amplitudes that are expected for vowels. In particular, there may be some formant prominences that are weakened or obscured by zeros.

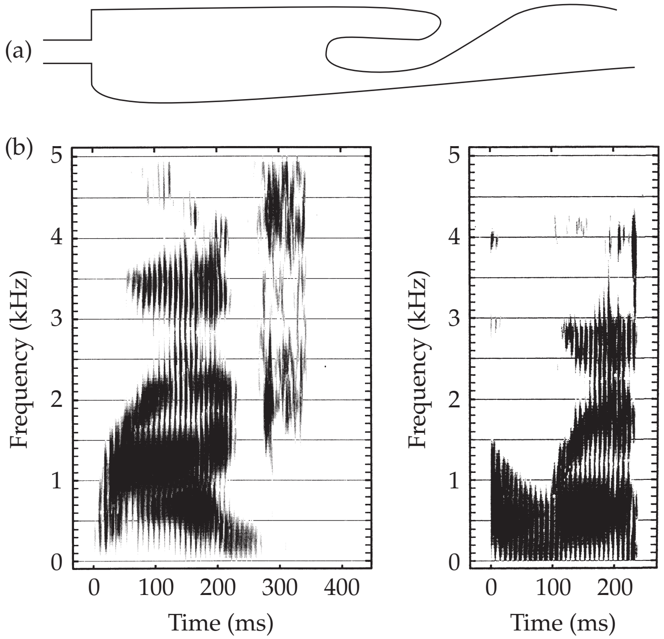


Figure 12.10 (a) There is a continuous path between the glottis at the left and the lips or nostrils on the right, but over another portion of the vocal tract, there is a complete closure. The side branch introduces zeros and poles to the transfer function, which lead to abrupt changes in spectral prominences. (b) The spectrograms are for the utterances "rug" and "a let," illustrating the lack of high-frequency energy for sounds produced with a side branch. (Spectrograms from Stevens, 1998, reprinted by kind permission of MIT Press)

In a rhotic segment in English, the zeros (together with additional poles) arise due to a cavity introduced in a space under the tongue blade (Espy-Wilson et al., 2000). Thus in the transition between a vowel and a rhotic the shaping of the tongue blade creates this extra cavity, which introduces an extra pole and zero in addition to the already existing poles. A consequence is that acoustic energy in the F3 region for an adjacent vowel becomes weakened by the presence of this zero. The new pole appears at a lower frequency, leaving reduced energy in the frequency region normally occupied by F3 for a vowel. The spectrogram on the left of Figure 12.10b is for the word *rug* and illustrates the lack of energy in the normal F3 region.

The defining articulatory gesture for this type of rhotic can be described as a shaping of the tongue blade in a way that creates this extra acoustic cavity under the blade. The corresponding defining acoustic attribute is a weakening of the spectrum prominence in the F3 region, accompanied by the introduction of an additional spectrum peak well below the normal F3 region.

For the lateral consonant (in English) the deviation from an all-pole spectrum is achieved by a tongue-blade configuration that divides the area function into two unequal paths. These paths can be of unequal lengths, they can have different cross-sectional areas, or one of the paths can be closed (Narayanan et al., 1997). The general influence of these deviations in the area function is again to introduce pole–zero pairs in the transfer function. The consequence of this type of spectrum perturbation, coupled in this case with a backing of the tongue body, is shown experimentally to be a modification of the spectrum balance of the output to produce a weakening of energy in the frequency range of about 1,500 to 2,500 Hz or higher. That is, a low-frequency F2 with the next highest spectrum peak being somewhat above the range normally expected for vowels. As a consequence there is a lack of spectrum energy over a relatively wide frequency range (Prahler, 1998). The spectrogram on the right of Figure 12.10b is for the utterance "a let," and illustrates the weakening of energy in the 1,500−2,500 Hz region.

It is possible, then, with our present knowledge, to specify the defining acoustic attributes for rhotics and laterals, but a close link between these acoustic properties and well-specified quantal articulations has not been adequately worked out.

2.1.3 The enigma of [stiff vocal folds] Halle and Stevens (1971) have argued that the voicing contrast observed for obstruent (i.e., [–sonorant]) consonants be represented by a feature [stiff vocal folds]. Evidence to support this view is both physiologic (Löfqvist et al., 1989) and acoustic (Hanson, 2009). The feature is [+stiff vocal folds] for voiceless and [–stiff vocal folds] for voiced obstruents. There are several acoustic correlates that contribute to this contrast (Keyser & Stevens, 2006), but the acoustic attribute that defines the quantal distinction is the presence or absence of vocal-fold vibration during the obstruent interval. The defining articulatory attribute is the vocal-fold stiffness, as illustrated in Figure 12.11. As this articulatory parameter is manipulated for a given value of intraoral pressure, the acoustic result passes through a threshold where vibration is inhibited on one

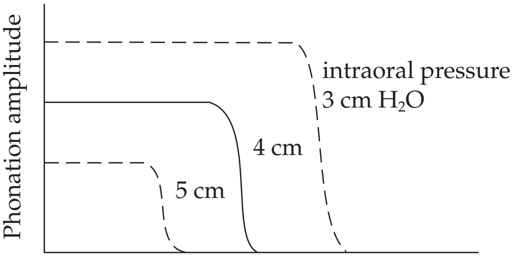


Figure 12.11 Schematic representation of relation between amplitude of glottal vibration and stiffness of the vocal folds for three different intraoral pressures. The figure shows an abrupt change of glottal source amplitude at a stiffness threshold that depends on intraoral pressure.

side ([+stiff vocal folds]) and facilitated on the other ([–stiff vocal folds]).[2] The stiffness threshold depends on the intraoral pressure.

Classification of the feature [stiff vocal folds] as articulator-free or articulator-bound is, however, a bit of a puzzle. It seems clear that the quantal nature of the feature is based on AMI principles. However, while other AMI-based features are easily seen to be articulator-free, [stiff vocal folds] better fits the definition of articulator-bound: a particular articulator and its action are specified. Furthermore, in terms of constraints among features, it behaves more like the articulator-bound features, in that it is not strongly constrained by the aerodynamic state of the vocal tract; any of the [–sonorant] sounds can include the feature [stiff vocal folds]. We have thus included [stiff vocal folds] in the articulator-bound column of Table 12.1.

2.1.4 A proposed hierarchy of features When the vocal tract is in a particular aerodynamic state, there are constraints on the additional aerodynamic events that can occur. (To some extent there are also constraints on the articulator-bound features.) For example, the feature [+continuant] (as we have defined it in section 2.1.1) can only be implemented in segments that are [–sonorant]. This example and others suggest that there is a natural hierarchy of the articulator-free features, such that specification of a feature that is low in the hierarchy implies specification of the features above it in the hierarchy. In an examination of the articulator-free features, from the point of view of the quantal acoustic and aerodynamic aspects, the feature [sonorant] is at the top of one branch of such a hierarchy. A possible hierarchy is illustrated in Figure 12.12, with the [+sonorant] branch of the tree in Figure 12.12a and the [–sonorant] branch in Figure 12.12b. The need for specification of features like [continuant], [strident], or [delayed release] is dependent on the value of the feature [sonorant]. For [+sonorant], those three features need not be specified. For [–sonorant], [continuant] would need to be specified, but specification of [delayed release] is dependent on the value of [continuant], and so forth, leading to a sparse specification of the articulator-free features for a segment. We emphasize that the hierarchy in Figure 12.12 is speculative, and is included here only for illustrative purposes. But we believe that the general idea behind this hierarchy, which is based on the physics of speech production, bears further study.

Because their definitions are based on aero-mechanical interactions, the articulator-free features are particularly constrained. Specification of the articulator-bound features, on the other hand, is less constrained, although which features need to be specified will depend to some extent on values of the articulator-free features. For example, in English the features [high], [low], and [back] are only specified for [+sonorant] sounds, and usually only for vowels (i.e., [–glide] according to the hierarchy in Figure 12.12). Also in English, the features [nasal], [rhotic], and [lateral] form a cluster of features that are only specified when the articulator-free feature [+consonantal] is specified. Another cluster of features in English is comprised of [labial], [coronal], [anterior], and [velar], which are only specified for [–sonorant] segments. Further constraints may occur within some of these clusters;

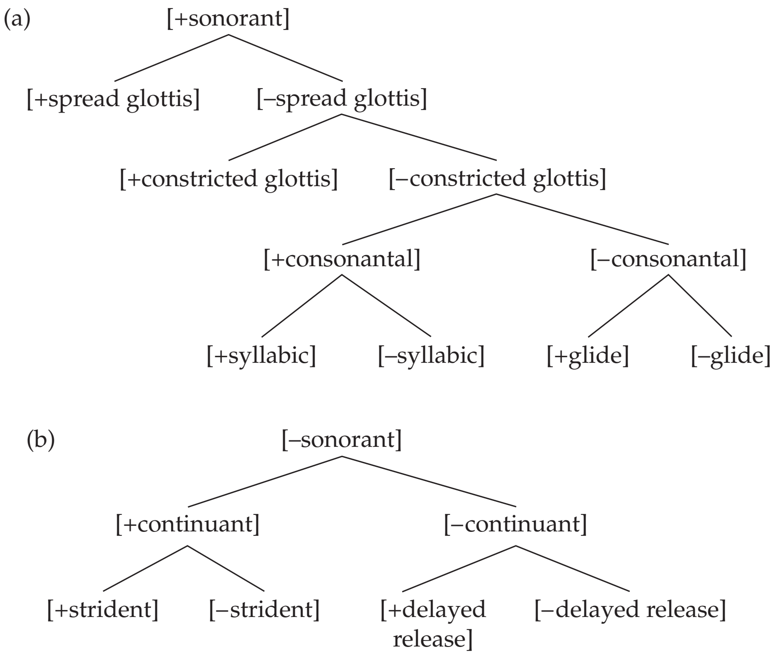


Figure 12.12 A possible hierarchy of articulator-free features, based on principles of aero-mechanical interactions. (a) the [+sonorant] branch; (b) the [–sonorant] branch.

for example, specification of [anterior] implies the feature [+coronal], while [–labial] and [–velar] do not need to be specified when [anterior] is given.

The sparseness of a feature-based representation is illustrated in Table 12.2. The word *seem* is represented by three segments, or three bundles of features. For each segment there is a list of several features, some articulator-free and some articulator-bound. The initial segment is [–sonorant, +continuant, +strident, +stiff vocal folds, +anterior]. However, since the feature [strident] implies [–sonorant] and [+continuant], it is not necessary to specify those features. With regard to the articulator-bound features in the first segment, the feature [+strident] implies a tongue-blade consonant, and the feature [+anterior] unambiguously specifies the place of articulation (at least in English). The second segment is a vowel with the articulator-free feature [–glide], from which the feature [+sonorant] can be implied. The articulator-bound features [–back], [+high], and [+tense] identify the vowel, since [+high] is automatically [–low]. In addition, these three articulator-bound features imply the articulator-free feature [–glide], because the features [high] and [low] are only specified for vowels. For the final segment, the articulator-free segment feature is [–syllabic], and this places limits on the articulator-bound features, one of which is [+nasal]. This feature requires a place designation, which in this case is [+labial].]

Table 12.2 Distinctive feature bundles that are required for the lexical representation of the word *seem*.

/s/	/i/	/m/
+strident	**–glide**	**–syllabic**
+stiff vocal folds	–back	+nasal
+anterior	+high	+labial
	+tense	

The hierarchies discussed here are based on articulatory–acoustic relations that underlie the quantal-enhancement theory. Other feature hierarchies, or adjustments of the feature inventories, have been proposed based on the patterns of features that play a role in phonological rules across languages. One example, proposed by Halle (2005), is to add to the inventory of features a set of unary features that specify a list of designated articulators. He shows the role of these features in several phonological rules. Still other hierarchies might emerge from models of human processing of speech – models that propose strategies used by listeners in uncovering information about features and lexical items from the acoustic signal and by speakers in the planning of utterances. A topic for future research is to examine the relations between these various hierarchies, and hopefully to bring together the disciplines of phonology, acoustic phonetics, and human speech production and perception.

3 Enhancement and Overlap: Introducing Variation to the Defining Acoustic Cues

Quantal theory seeks to explain why the inventory of distinctive features that make up the phonological contrasts in the languages of the world is what it is. It is not intended to be the principal basis of a model that describes how human listeners extract words from continuous speech or how speakers generate running speech. The surface representation of words and word sequences includes not only the *feature-defining* acoustic and articulatory attributes but also an array of articulatory gestures (and their acoustic consequences), including prosodic events, that enhance the perceptual saliency of the defining attributes.

There are two general ways in which enhancement gestures may be added to a defining gesture for a particular feature in a language: (1) An articulatory gesture is superimposed on the defining gesture, and thereby enhances the perceptual saliency of the defining attribute. In effect, the acoustic attribute resulting from the enhancing gesture increases the perceptual distance between the feature and other features for which the attributes (or cues) might be similar (Liljencrants & Lindblom, 1972; Diehl, 1991). (2) A new acoustic attribute is introduced that is separate from the defining acoustic attribute for the feature. This new attribute introduces new perceptual cues for the feature.

An example of the first type of enhancement in English is the rounding of the lips in the production of /ʃ/. This rounding tends to lower the natural frequency of the front portion of the vocal tract, causing the frequency of the lowest major spectrum prominence in the fricative spectrum to be in the F3 range, well below the F4 or F5 range for the lowest spectrum prominence for the contrasting fricative consonant /s/. Other examples can be observed in vowels. In a five-vowel system, the nonlow back vowels are often produced with lip rounding, presumably to enhance the contrast with vowels having the feature [–back] (Keyser & Stevens, 2006). Similarly, the nonlow front vowels are often produced with lip spreading, thereby enhancing the acoustic attribute that defines [–back]. These enhancements are usually implemented for all contexts in which the feature appears.

For the second type of enhancement, the additional perceptual cues for a feature can depend on the context in which the feature occurs. This type of enhancement is introduced in regions of the speech signal that are *adjacent* to the times when the defining acoustic attributes for the feature appear, rather than being superimposed on the defining attributes. A typical example of this second type of enhancement is the frequency of F2 or its movement immediately after the release burst of a stop consonant. The voicing feature for obstruent consonants in English is enhanced by several kinds of cues, including the increased fundamental frequency in a vowel following a voiceless obstruent, or the aspiration that occurs in the initial part of a vowel that follows a voiceless stop consonant. A number of enhancements of this kind are discussed in Keyser and Stevens (2006).

The ubiquitous presence of these enhancements of defining attributes for features is evidence for redundancy in the acoustic manifestations of individual features. This redundancy helps to provide additional cues for the listener, particularly in running speech, in which there is likely to be overlap of articulatory gestures for some features. Often this overlap can cause masking or obliteration of the defining properties for a feature. Enhancement cues can help to provide the listener with additional cues. Although particular quantal articulatory–acoustic relations form the basis for the existence of phonological contrasts or features, the listener is often provided with enhancements which are no less important for perception than are defining aspects.

The quantal approach to distinctive contrasts is not proposed to extend to contrasts between prosodic categories of prominences and constituents, which provide structural information to listeners at both the phrase and word levels of an utterance. The articulatory–acoustic relations that lead to definitions of features have to do with "target" acoustic attributes, and never time-varying parameters or trajectories, such as f_0 movements, durations, or formant movements. These time-varying parameters, however, can certainly qualify as enhancements. An interesting question is how to handle the fact that some acoustic attributes have been treated as cues to both prosodic and segment-level contrasts, including, e.g., the generally accepted correlates of duration, f_0, and vowel amplitude.

Thus, in quantal/enhancement theory, prosodic aspects of spoken utterances, which generally involve time or time variation, are considered to be enhancing, and are therefore not necessarily based on a quantal relationship between acoustics and articulation. On the other hand, duration and f_0 have also been considered to be segment-level contrastive cues, i.e., to vowel length and tone, respectively. An approach to these two phenomena that is more consistent with quantal/enhancement theory views vowel length and lexical tone as word-level prosodic phenomena, rather than as cues to segmental contrasts. Thus, inspired by recent work of Okobi (2006), we suggest that duration and f_0 at the word level might be regarded as an aspect of word-level prosody or lexical stress, rather than as contrastive segmental features, and that they may operate very differently in different language systems. That is, it might make sense to consider f_0 and duration to be word-level prosodic cues, even in cases where they have traditionally been considered to be features of segments. For example, in tone languages, rather than consider that the tones are features that are specified for segments, the tones could be word-level prosodic events that result in words with contrastive meanings. Similarly, vowel duration contrasts might be cues to contrasts in lexical stress (e.g., based on differences in number of moras) rather than to feature contrasts at the segmental level (see, for example, Clements & Keyser, 1983).

4 Concluding Remarks

4.1 Other models of speech processing

There are other models of speech production and perception that propose somewhat different approaches toward the representation of phonetic or phonological units. In common with the quantal/enhancement model proposed here, these models also provide ways of accounting for the variability in the relation between the apparently continuous properties of the acoustic signal and the more discrete phonological units. In one such model, the inventory of units is expanded to a set of "phones," in which the variation of a given phonological unit in different contexts is represented by several different "allophones" (see, for example, O'Shaughnessy, 2000). For example, the voiceless stop consonant /p/ might be represented by an aspirated /p/ in some contexts and an unreleased /p/ in other contexts. By expanding the inventory of allophones, this kind of representation captures some of the acoustic variability that occurs in the surface acoustic representation of particular segmental units. This allophonic approach, however, contains little theoretical underpinning that estimates the inventory of phones that may be necessary. The selection of allophones is based largely on listener judgments; because it is not tied to distinctive contrasts, the set often varies considerably from one study to the next.

Another model has been proposed to account for certain types of variability that occur in the production of coronal consonants (Lahiri and Reetz, 2002). For example, if the word *ben* discussed above occurred in a sequence "ben goes . . . ," one might imagine that /n/ could be represented by an alternative pronunciation as *beng*, or, as proposed by Lahiri and Reetz (2002), coronal consonants might be represented lexically as underspecified. That is, this place of articulation could be regarded as a kind of default place. This approach may have merit for coronals in certain contexts (e.g., in syllable-final stop consonants), although it is not clear what advantage it might have for other contexts. Another approach that is oriented more to articulatory units represents an utterance in terms of inventories of articulatory gestures (Browman & Goldstein, 1992). This method of representation provides a way of accounting for the various kinds of gestural overlap that result in variability of the speech signal in various phonetic and prosodic contexts. It is based on the premise that a listener has the capacity to interpret the signal in terms of the articulatory gestures that produced it. The details of the process whereby this process is carried out have not been spelled out.

The quantal/enhancement approach has elements of both the "phone" model and the articulatory-oriented model. The units of representation in the speaker-listener are distinctive features, as noted above. These features are universal, and are defined by quantal relations based on both articulatory descriptions and the corresponding acoustic consequences. A "segment" consists of a bundle of distinctive features, but these features are independent of the context. However, in different contexts the distribution of acoustic cues for the individual features may be different. Almost always, however, some of these cues for the features that define a segment are still present, while others may be weakened or obliterated. In most cases, it appears that enough information remains in these cues to permit the features to be identified.

Still another component of human speech perception is emphasized in research on an "indexical" model, which is based on the observation that the perception of an utterance by a listener can be influenced by the listener's knowledge of the speaker (see, for example, Luce & McLennan, 2005, for a discussion). In terms of the quantal/enhancement model proposed here, this observation emphasizes the importance of the fact that the defining acoustic attribute for a distinctive feature as produced by a particular speaker may depend on the speaker's anatomical details. More importantly, however, the details of the enhancing attributes attached to a feature are probably more speaker dependent.

4.2 Problems for future research

The quantal/enhancement theory proposed here appears to provide a physical basis for a number of distinctive features, both articulator-free and articulator-bound. However, there are a number of features for which defining attributes have not been clearly worked out. In English, these include the features [high], [glide], and the feature [tense] for vowels (sometimes called advanced tongue root, or ATR). Many features that define contrasts in other languages have yet to be examined within the quantal/enhancement framework. These include the pharyngeal consonants, features that define different tongue-blade shapes as well as several places of articulation, and others. This research requires that the quantal relations be defined, and that enhancing gestures for each feature be catalogued. This exercise could lead to some modification or revision of some classical inventories of features.

Another topic that has received considerable attention from phonologists, but not from an acoustical and aerodynamic point of view, is research on the study of the hierarchical organization of features. In the formulation of feature geometries it could be worthwhile to examine the classification of features based on acoustic and aerodynamic observations in addition to the role of features in assimilation processes. For example, in section 2.1.2 there is a collection of the features [lateral], [rhotic], and [nasal], on the grounds that all of these features are [+consonantal, +sonorant, and +coronal] (if only the coronal nasal is included). All of these segments come together since their articulatory description involves a side branch in the area function and the common quantal acoustic attribute is the action of zeros in the transfer function. One might ask if the segments in this group play a role in some assimilation process.

While aerodynamics and acoustics have helped to provide a basis for phonological contrasts in language, they have been less successful in providing a framework for quantifying the role of prosody in human speech production and perception. As has been noted in this chapter, prosody clearly provides additional cues to a listener, but these are different from the defining attributes for features. Prosody is concerned with temporal or time-varying properties rather than "target" attributes. One might hope, however that with future research these two sides of human speech production and perception might be brought closer together with quantitative tools of acoustics, physiology, and aerodynamics.

NOTES

Development of some of the concepts in this chapter has profited from discussions with Nick Clements, Morris Halle, and Jay Keyser, whose insights we acknowledge with thanks. We are greatly indebted to Xuemin Chi for her assistance in the generation of the figures. We also thank Stefanie Shattuck-Hufnagel for valuable discussion and comments on earlier versions of this manuscript. This research was supported in part by NIH grant DC-00075.

[1] The question of how a listener disentangles and uses variability in the speech signal to extract the discrete linguistic representation has been addressed by Stevens (2005).

[2] Like many distinctive features, there are several acoustic attributes that contribute to listener perception of the contrasts described here. Usually one of these attributes exhibits a clear quantal articulatory–acoustic relationship. Other attributes (called enhancements) may contribute to perception of the contrast in some contexts, and do not exhibit quantal characteristics. See section 3 on enhancement; also Keyser and Stevens (2006) and Stevens and Keyser (in press).

REFERENCES

Browman C. P. & Goldstein, L. (1992) Articulatory phonology: An overview. *Phonetica*, 49, 155–80.

Butcher, A. (2006) Australian aboriginal languages: Consonant-salient phonologies and the "place-of-articulation imperative." In J. Harrington & M. Tabain (eds.), *Speech Production: Models, Phonetic Processes and Techniques* (pp. 187–210). New York: Psychology Press.

Chi, X. & Sonderegger, M. (2007) Subglottal coupling and its influence on vowel formants. *Journal of the Acoustical Society of America*, 122, 1735–45.

Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York: Harper & Row.

Clements, G. N. (1985) The geometry of phonological features. *Phonology Yearbook*, 2, 223–50.

Clements, G. N. & Keyser, S. J. (1983) *CV Phonology: A Generative Theory of the Syllable*. Cambridge, MA: MIT Press.

Cranen, B. & Boves, L. (1987) On subglottal formant analysis. *Journal of the Acoustical Society of America*, 81, 734–45.

Diehl, R. (1991) The role of phonetics within the study of language. *Phonetica*, 48, 120–34.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000) Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America*, 108, 343–56.

Halle, M. (1992) Features. In W. Bright (ed.), *Oxford International Encyclopedia of Linguistics*, vol. 3 (pp. 207–12). New York: Oxford University Press.

Halle, M. (1995) Feature geometry and feature spreading. *Linguistic Inquiry*, 26, 1–46.

Halle, M. (2005) Palatalization/velar softening: What it is and what it tells us about the nature of language. *Linguistic Inquiry*, 36, 23–41.

Halle, M. & Stevens, K. N. (1971) A note on laryngeal features. *MIT Research Laboratory of Electronics Quarterly Progress Report* 101, 198–213.

Halle, M. & Stevens, K. N. (1991) Knowledge of language and the sounds of speech. In J. Sundberg, L. Nord, & R. Carlson (eds.), *Music, Language, Speech and Brain* (pp. 1–19). London: Macmillan.

Hanson, H. M. (2009) Effects of obstruent consonants on fundamental frequency at vowel onset, *Journal of the Acoustical Society of America*, 125, 425–41.

Hanson, H. M. & Stevens, K. N. (1995) Sub-glottal resonances in female speakers and their effect on vowel spectra. In *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 3, 182–5.

Jakobson, R. (1928) Quelles sont les méthodes les mieux appropriées à un exposé complet et pratique de la grammaire d'une langue quelconque? *Actes du Premier Congrès International de Linguistes*, (Leiden, 1930), 33–6. (Reprinted 1962 in *Selected Writings I,* 3–6. The Hague: Mouton.)

Jakobson, R., Fant, G., & Halle, M. (1952) Preliminaries to speech analysis: The distinctive features and their correlates. *MIT Acoustics Laboratory Technical Report* 13. (Reprinted 1967, Cambridge, MA: MIT Press.)

Jung, Y. & Stevens, K. N. (2007) Acoustic-articulatory evidence for quantal vowel categories: The feature [low]. *Journal of the Acoustical Society of America*, 122, 3029.

Keyser, S. J. & Stevens, K. N. (2006) Enhancement and overlap in the speech chain. *Language*, 82, 33–63.

Ladefoged, P. & Traill, A. (1994) Clicks and their accompaniments. *Journal of Phonetics*, 22, 33–64.

Lahiri, A. & Reetz, H. (2002) Underspecified recognition. In C. Gussenhoven, N. Warner, & T. Rietveld (eds.), *Laboratory Phonology 7* (pp. 637–862). Mouton: Berlin.

Liljencrants, J. & Lindblom, B. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839–62.

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989) The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America*, 85, 1314–21.

Luce, P. A. & McLennan, C. T. (2005) Spoken word recognition: The challenge of variation. In D. Pisoni & R. Remez (eds.), *The Handbook of Speech Perception* (pp. 591–60). Oxford: Blackwell.

Miller, G. A. & Nicely, P. E. (1955) An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27, 339–52.

Narayanan, S., Alwan, A., & Haker, K. (1997) Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data, part I: The laterals. *Journal of the Acoustical Society of America*, 101, 1064–77.

Okobi, T. (2006) Acoustic correlates of word stress in American English. Doctoral dissertation, Massachusetts Institute of Technology.

O'Shaughnessy, D. (2000) *Speech Communications: Human and Machine.* New York: IEEE Press.

Prahler, A. (1998) Analysis and synthesis of the American English lateral consonant, masters dissertation, Massachusetts Institute of Technology.

Sonderegger, M. (2004) Subglottal coupling and vowel space: An investigation in quantal theory. BS thesis, Massachusetts Institute of Technology.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In P. B. Denes & E. E. David, Jr. (eds.), *Human Communication: A Unified View* (pp. 51–66). New York: McGraw-Hill.

Stevens, K. N. (1989) On the quantal nature of speech. *Journal of Phonetics*, 17, 3–46.

Stevens, K. N. (1993) Modelling affricate consonants. *Speech Communication*, 13, 33–43.

Stevens, K. N. (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. (2001) The properties of the vocal-tract walls help to shape several phonetic distinctions in language. In N. Grönnum & J. Rischel (eds.), *To Honour Eli Fischer-Jörgensen: Festschrift on the Occasion of Her 90th Birthday, February 11th, 2001,* (Travaux du Cercle Linguistique de Copenhague; pp. 285–97). Copenhagen: C. A. Reitzel.

Stevens, K. N. (2003) Acoustic and perceptual evidence for universal phonological features. In M. J. Solè, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 199–202). Barcelona: Causal Productions.

Stevens, K. N. (2005) Features in speech perception and lexical access. In D. Pisoni & R. Remez (eds.), *The Handbook of Speech Perception* (pp. 125–55). Oxford: Blackwell.

Stevens, K. N. & Keyser, S. J. (in press) Quantal theory, enhancement, and overlap. *Journal of Phonetics*.

Traill, A. (1985) Phonetic and phonological studies of !Xóõ Bushman. *Quellen zur Khoisan-Forschung*, 1. Hamburg: Helmut Buske.

FURTHER READING

Halle, M. (2002) *From Memory to Speech and Back: Papers on Phonetics and Phonology, 1954–2002*. Berlin: Mouton de Gruyter.

Ladefoged, P. & Maddieson, I. (1996) *The Sounds of the World's Languages*. Oxford: Blackwell.

Perkell, J. S. & Klatt, D. H. (1986) *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum.

13 Aspects of Auditory Processing Related to Speech Perception

BRIAN C. J. MOORE

1 Introduction

This chapter reviews selected aspects of auditory processing, chosen because they play a role in the perception of speech. The review is concerned with relatively basic processes, many of which are strongly influenced by the operation of the peripheral auditory system and which can be characterized using simple stimuli such as pure tones and bands of noise. Although there are certainly specialized brain mechanisms for speech perception, the initial analysis of speech and non-speech sounds is probably similar, with many processes being shared between the two. It turns out that the resolution of the auditory system in frequency and time, as measured in psychoacoustic experiments, usually markedly exceeds the resolution necessary for the identification or discrimination of speech sounds. This partly accounts for the fact that speech perception is robust, and resistant to distortion of the speech and to background noise.

2 Frequency Selectivity

2.1 *The concept of the auditory filter*

Frequency selectivity refers to the ability to resolve the sinusoidal components in a complex sound, and it plays a role in many aspects of auditory perception, including the perception of loudness, pitch, and timbre. Fletcher (1940), following Helmholtz (1863), suggested that frequency selectivity can be modeled by considering the peripheral auditory system as a bank of bandpass filters, with overlapping passbands. These filters are called the "auditory filters." Fletcher thought that the basilar membrane within the cochlea provided the basis for the auditory filters, and this view is widely accepted. Each location on the basilar membrane responds to a limited range of frequencies, so each different point corresponds to a filter with a different center frequency.

Frequency selectivity can be most readily quantified by studying masking, which is the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound. The following assumptions are made about the process of detecting a sinusoidal signal in a broadband noise background:

1 The listener makes use of an auditory filter with a center frequency close to that of the signal. This filter passes the signal but removes a great deal of the noise.
2 Only the components in the noise which pass through the filter have any effect in masking the signal.
3 The threshold for detecting the signal is determined by the amount of noise passing through the auditory filter; specifically, threshold is assumed to correspond to a certain signal-to-noise ratio at the output of the filter.

This set of assumptions is known as the "power spectrum model" of masking (Patterson & Moore, 1986), since the stimuli are represented by their long-term power spectra, i.e., the short-term fluctuations in the masker are ignored. Although the assumptions of the model are not always valid (Moore, 2003a), stimuli can be found for which the assumptions are not strongly violated.

The question considered next is "What is the shape of the auditory filter?" In other words, how does its relative response change as a function of the input frequency? Most methods for estimating the shape of the auditory filter at a given center frequency are based on the assumptions of the power spectrum model of masking. The threshold of a signal whose frequency is fixed is measured in the presence of a masker whose spectral content is varied. It is assumed, as a first approximation, that the signal is detected using the single auditory filter which is centered on the frequency of the signal, and that threshold corresponds to a constant signal-to-masker ratio at the output of that filter. The methods described below both measure the shape of the filter using this technique.

2.2 *Psychophysical tuning curves*

One method of measuring the shape of the auditory filter involves a procedure which is analogous in many ways to the determination of a neural tuning curve, and the resulting function is called a psychophysical tuning curve (PTC). To determine a PTC, the signal is fixed in level, usually at a very low level, say, 10 dB above absolute threshold (called 10 dB Sensation Level, SL). The masker can be either a sinusoid or a narrow band of noise.

For each of several masker frequencies, the level of the masker needed to just mask the signal is determined. Because the signal is at a low level it is assumed that it produces activity primarily at the output of a single auditory filter. It is assumed further that at threshold the masker produces a constant output power from that filter, in order to mask the fixed signal. Thus the PTC indicates the masker level required to produce a fixed output power from the auditory filter as a function of frequency. Normally a filter characteristic is determined by plotting the output from the filter for an input varying in frequency and fixed in level.

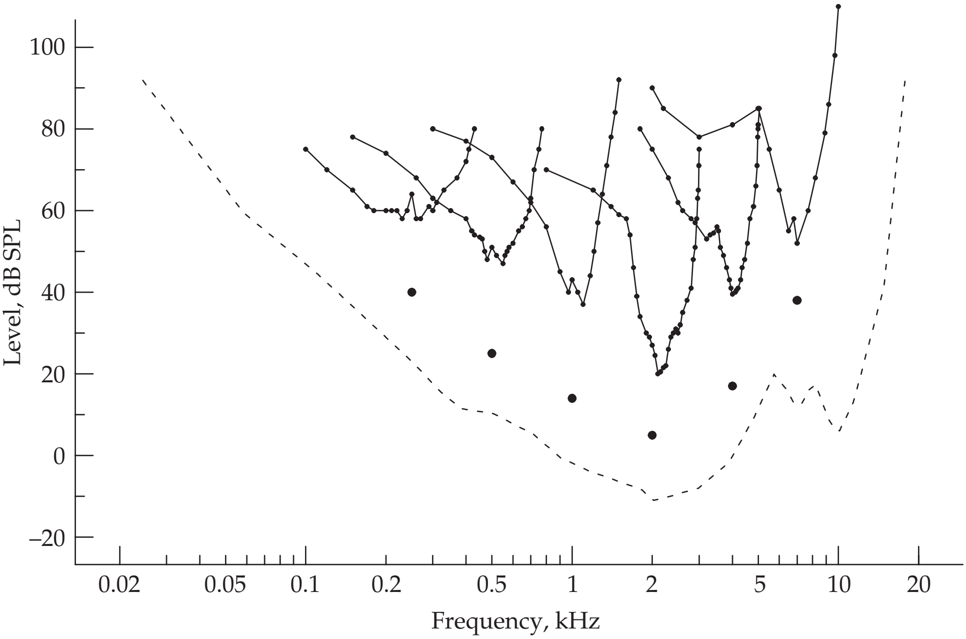


Figure 13.1 Psychophysical tuning curves (PTCs) determined in simultaneous masking, using sinusoidal signals at 10 dB SL. For each curve, the solid circle below it indicates the frequency and level of the signal. The masker was a sinusoid which had a fixed starting phase relationship to the 50 ms signal. The masker level required for threshold is plotted as a function of masker frequency on a logarithmic scale. The dashed line shows the absolute threshold for the signal. (Data from Vogten, 1978)

However, if the filter is linear the two methods give the same result. Thus, assuming linearity, the shape of the auditory filter can be obtained simply by inverting the PTC. Examples of some PTCs are given in Figure 13.1.

One problem in interpreting PTCs is that, in practice, the listener may use the information from more than one auditory filter. When the masker frequency is above the signal frequency, the listener might do better to use the information from a filter centered just below the signal frequency. If the filter has a relatively flat top, and sloping edges, this will considerably attenuate the masker at the filter output, while only slightly attenuating the signal. By using this filter the listener can improve performance. This is known as “off-frequency listening” or “off-place listening,” and there is good evidence that humans do indeed listen “off-frequency” when it is advantageous to do so (Johnson-Davies & Patterson, 1979; O’Loughlin & Moore, 1981b). The result of off-frequency listening is that the PTC has a sharper tip than would be obtained if only one auditory filter were involved (O’Loughlin & Moore, 1981a).

Another problem with PTCs is that they can be influenced by the detection of beats, which are amplitude fluctuations caused by the interaction of the signal and the masker. The rate of the beats is equal to the difference in frequency of the signal and masker. Beats with a low rate are more easily detected than beats with a rate above about 120 Hz (Kohlrausch et al., 2000), and slow beats provide a detection cue which results in an increase in the masker level required for threshold for masker frequencies adjacent to the signal frequency. This results in a PTC which has a sharper tip than the underlying auditory filter (Kluk & Moore, 2004). This sharpening effect is greatest when a sinusoidal masker is used, but it occurs even when the masker is a narrowband noise (Kluk & Moore, 2004).

2.3 *The notched-noise method*

Patterson (1976) described a method of determining the auditory filter shape which limits off-frequency listening and appears not to be influenced by beat detection. The method is illustrated in Figure 13.2. The signal (indicated by the bold vertical line) is fixed in frequency, and the masker is a noise with a spectral notch centered at the signal frequency. The deviation of each edge of the notch from the center frequency is denoted by Δ*f*. The width of the notch is varied, and the threshold of the signal is determined as a function of notch width. Since the notch is symmetrically placed around the signal frequency, the method cannot reveal asymmetries in the auditory filter, and the analysis assumes that the filter is symmetric on a linear frequency scale. This assumption appears to be reasonable, at least for the top part of the filter and at moderate sound levels, since PTCs are quite symmetric around the tips. For a signal symmetrically placed in a notched noise, the optimum signal-to-masker ratio at the output of the auditory filter is achieved with a filter centered at the signal frequency, as illustrated in Figure 13.2.

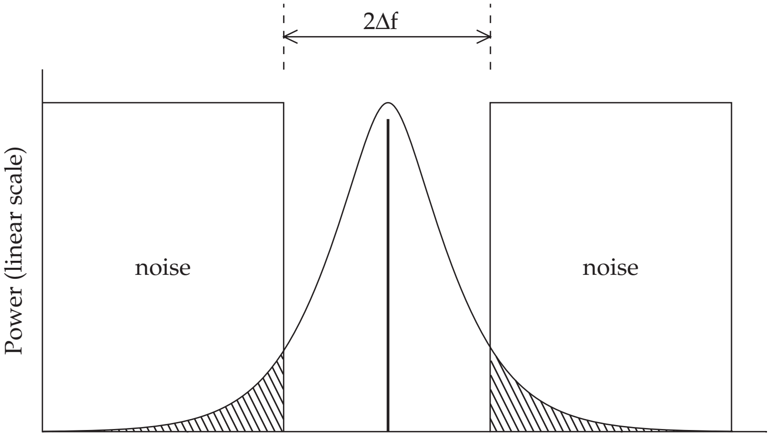


Figure 13.2 Schematic illustration of the technique used by Patterson (1976) to determine the shape of the auditory filter. The threshold of the sinusoidal signal (indicated by the bold vertical line) is measured as a function of the width of a spectral notch in the noise masker. The amount of noise passing through the auditory filter centered at the signal frequency is proportional to the shaded areas.

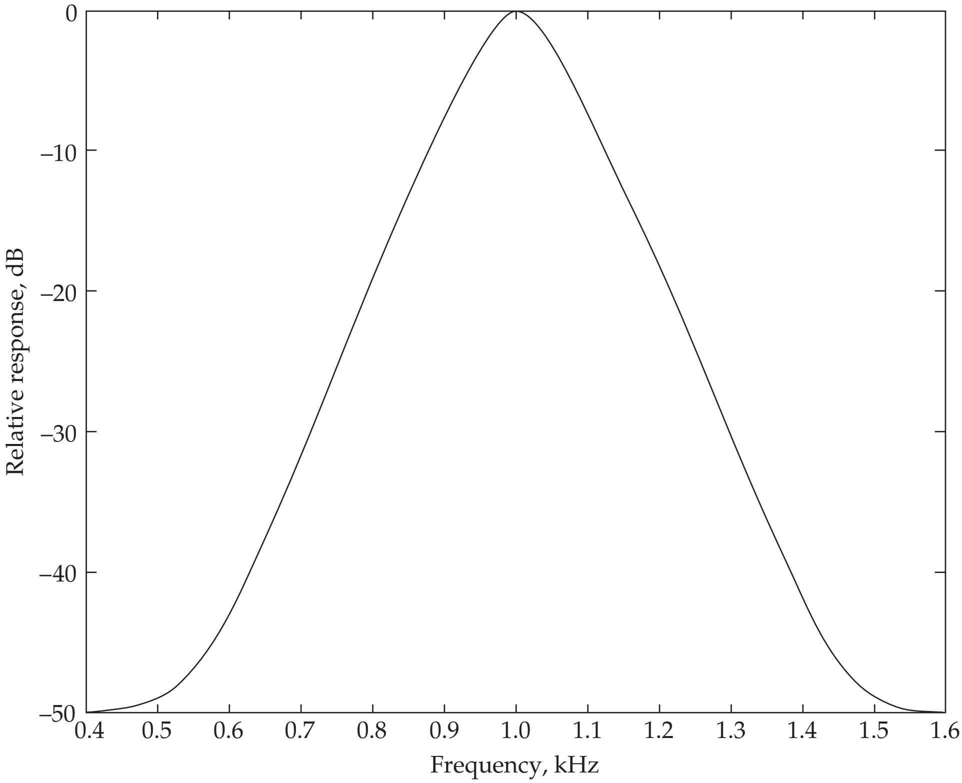


Figure 13.3 A typical auditory filter shape determined using the notched-noise method. The filter is centered at 1 kHz. The relative response of the filter (in dB) is plotted as a function of frequency.

As the width of the spectral notch is increased, less and less noise passes through the auditory filter. Thus the threshold of the signal drops. The amount of noise passing through the auditory filter is proportional to the area under the filter in the frequency range covered by the noise. This is shown as the shaded areas in Figure 13.2. Assuming that threshold corresponds to a constant signal-to-masker ratio at the output of the filter, the change in signal threshold with notch width indicates how the area under the filter varies with Δf. The area under a function between certain limits is obtained by integrating the value of the function over those limits. Hence by differentiating the function relating threshold to Δf, the relative response of the filter at that value of Δf is obtained. In other words, the relative response of the filter for a given deviation, Δf, from the center frequency is equal to the slope of the function relating signal threshold to notch width, at that value of Δf.

A typical auditory filter derived using this method is shown in Figure 13.3. It has a rounded top and quite steep skirts. The sharpness of the filter is often specified as the bandwidth of the filter at which the response has fallen by a factor of two in power, i.e., by 3 dB. The 3 dB bandwidths of the auditory filters derived using the notched-noise method are typically between 10 and 15 percent of the center frequency. An alternative measure is the equivalent rectangular bandwidth (ERB), which is the bandwidth of a rectangular filter which has the same peak transmission as the filter of interest and which passes the same total power for a white noise input. The ERB of the auditory filter is a little larger than the 3 dB bandwidth. In what follows, the mean ERB of the auditory filter determined using young listeners with normal hearing and using a moderate noise level is denoted $\mathrm{ERB_N}$ (where the subscript N denotes normal hearing). An equation describing the value of $\mathrm{ERB_N}$ as a function of center frequency, F (in Hz), is (Glasberg & Moore, 1990):

$$\mathrm{ERB_N} = 24.7(0.00437F + 1) \qquad (1)$$

Sometimes it is useful to plot psychoacoustical data on a frequency scale related to $\mathrm{ERB_N}$, called the $\mathrm{ERB_N}$-number scale. For example, the value of $\mathrm{ERB_N}$ for a center frequency of 1 kHz is about 132 Hz, so an increase in frequency from 934 to 1,066 Hz represents a step of one $\mathrm{ERB_N}$-number. A formula relating $\mathrm{ERB_N}$-number to frequency is (Glasberg & Moore, 1990):

$$\mathrm{ERB_N}\text{-number} = 21.4 \log_{10}(0.00437F + 1), \qquad (2)$$

where F is frequency in Hz. Each one-$\mathrm{ERB_N}$ step on the $\mathrm{ERB_N}$-number scale corresponds approximately to a constant distance (0.9 mm) along the basilar membrane (Moore, 1986). The $\mathrm{ERB_N}$-number scale is conceptually similar to the Bark scale (Zwicker & Terhardt, 1980), which has been widely used by speech researchers, although it differs somewhat in numerical values. The $\mathrm{ERB_N}$-number scale appears to give a more accurate representation of auditory frequency selectivity at low frequencies than the Bark scale or the Mel scale (a scale of subjective pitch that has also been used by speech researchers; Moore & Sek, 1995).

The notched-noise method has been extended to include conditions where the spectral notch in the noise is placed asymmetrically about the signal frequency. This allows the measurement of any asymmetry in the auditory filter, but the analysis of the results is more difficult, and has to take off-frequency listening into account (Patterson & Nimmo-Smith, 1980). It is beyond the scope of this chapter to give details of the method of analysis; the interested reader is referred to Patterson and Moore (1986), Moore and Glasberg (1987), Glasberg and Moore (1990, 2000), Rosen et al. (1998) and Unoki et al. (2006). The results show that the auditory filter is reasonably symmetric at moderate sound levels, but becomes increasingly asymmetric at high levels, the low-frequency side becoming shallower than the high-frequency side. The filter shapes derived using the notched-noise method are quite similar to inverted PTCs (Glasberg et al., 1984), except that PTCs are slightly sharper around their tips, probably as a result of off-frequency listening and beat detection.

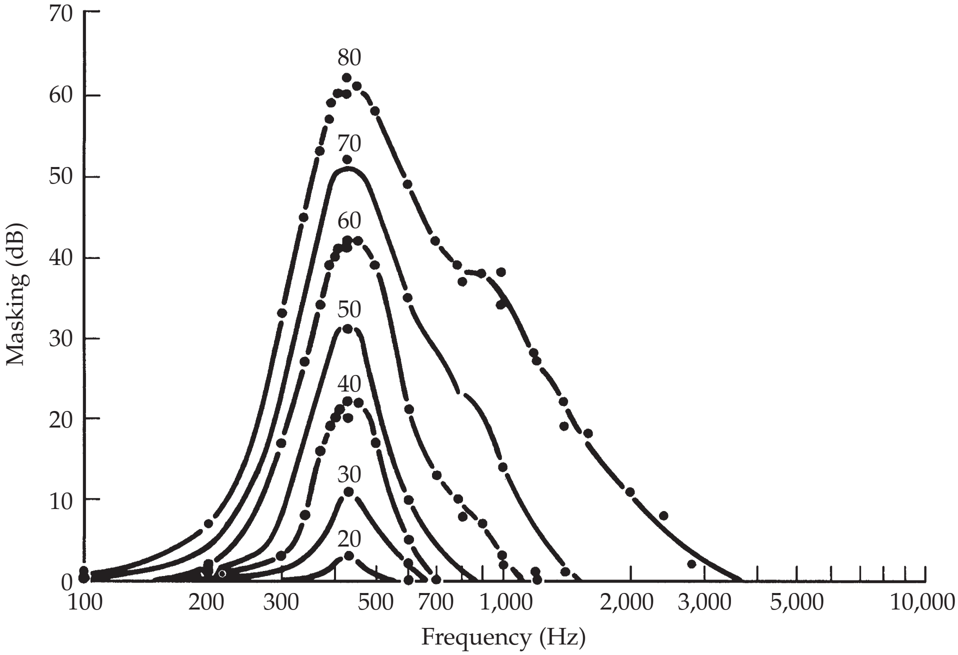


Figure 13.4 Masking patterns for a narrow-band noise masker centered at 410 Hz. Each curve shows the elevation in threshold of a pure-tone signal as a function of signal frequency. The overall noise level in dB SPL for each curve is indicated in the figure. (Data from Egan & Hake, 1950)

2.4 Masking patterns and excitation patterns

In the masking experiments described so far, the frequency of the signal was held constant, while the masker was varied. These experiments are most appropriate for estimating the shape of the auditory filter at a given center frequency. However, in many experiments the masker was held constant in both level and frequency and the signal threshold was measured as a function of the signal frequency. The resulting functions are called masking patterns or masked audiograms.

Masking patterns show steep slopes on the low-frequency side (when the signal frequency is below that of the masker), of between 55 and 240 dB/octave. The slopes on the high-frequency side are less steep and depend on the level of the masker. Figure 13.4 shows a typical set of results, obtained using a narrowband noise masker centered at 410 Hz, with the overall masker level varying from 20 to 80 dB SPL in 10 dB steps (data from Egan & Hake, 1950). Notice that on the high-frequency side the curve is shallower at the highest level. Around the tip of the masking pattern, the growth of masking is approximately linear; a 10 dB increase in masker level leads to roughly a 10 dB increase in the signal threshold. However, for signal frequencies well above the masker frequency, in the range from about 1,300 to 2,000 Hz, when the level of the masker is increased by 10 dB (e.g., from 70 to 80 dB SPL), the masked threshold increases by more than 10 dB; the amount of masking grows nonlinearly on the high-frequency side. This has been called the "upward spread of masking."

The masking patterns do not reflect the use of a single auditory filter. Rather, for each signal frequency the listener uses a filter centered close to the signal frequency. Thus the auditory filter is shifted as the signal frequency is altered. One way of interpreting the masking pattern is as a crude indicator of the excitation pattern of the masker (Zwicker & Fastl, 1999). The excitation pattern is a representation of the effective amount of excitation produced by a stimulus as a function of characteristic frequency (CF) on the basilar membrane (the CF is the frequency to which a given place on the basilar membrane is most sensitive), and is plotted as effective level (in dB) against CF. In the case of a masking sound, the excitation pattern can be thought of as representing the relative amount of vibration produced by the masker at different places along the basilar membrane. The signal is detected when the excitation it produces is some constant proportion of the excitation produced by the masker at places with CFs close to the signal frequency. Thus the threshold of the signal as a function of frequency is proportional to the masker excitation level. The masking pattern should be parallel to the excitation pattern of the masker, but shifted vertically by a small amount. In practice, the situation is not so straightforward, since the shape of the masking pattern is influenced by factors such as off-frequency listening, the detection of beats and combination tones (Moore et al., 1998; Alcántara et al., 2000) and by the physiological process of suppression (Delgutte, 1990).

Moore and Glasberg (1983b) have described a way of deriving the shapes of excitation patterns using the concept of the auditory filter. They suggested that the excitation pattern of a given sound can be thought of as the output of the auditory filters plotted as a function of their center frequency. To calculate the excitation pattern of a sound, it is necessary to calculate the output of each auditory filter in response to that sound, and to plot the output as a function of the filter center frequency. The characteristics of the auditory filters are determined using the notched-noise method described earlier. Figure 13.5 shows excitation patterns calculated in this way for 1,000 Hz sinusoids with various levels. The patterns are similar in form to the masking patterns shown in Figure 13.4. Software for calculating excitation patterns can be downloaded from http://hearing.psychol.cam.ac.uk/Demos/demos.html.

It should be noted that excitation patterns calculated as described above do not take into account the physiological process of suppression, whereby the response to a given frequency component can be suppressed or reduced by a strong neighboring frequency component (Sachs & Kiang, 1968). For speech sounds having spectra with strong peaks and valleys, such as vowels, suppression may have the effect of increasing the peak-to-valley ratio of the excitation pattern (Moore & Glasberg, 1983a). Also, the calculated excitation patterns are based on the power-spectrum model of masking, and do not take into account the effects of the relative phases of the components in complex sounds. However, it seems likely

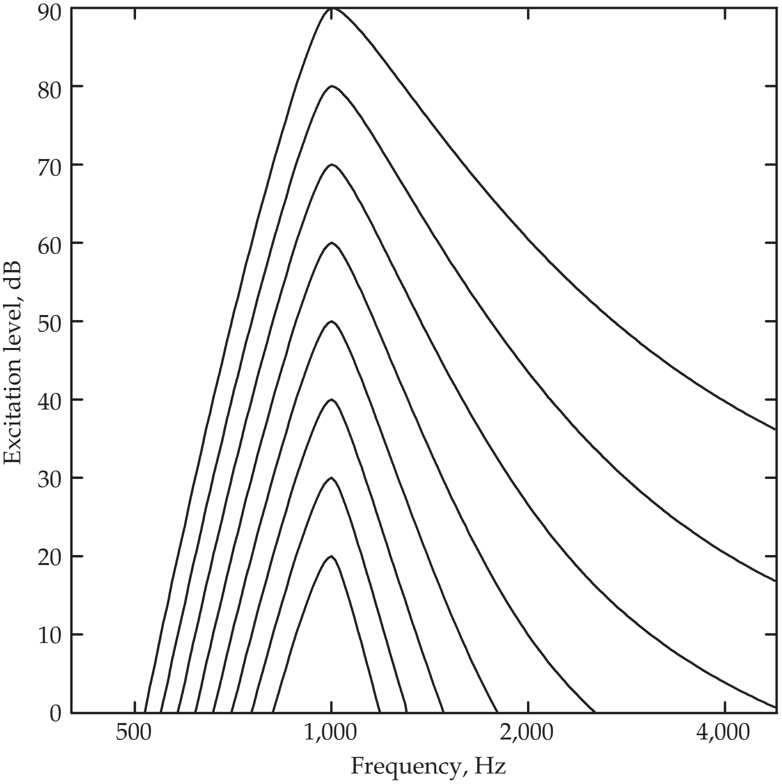


Figure 13.5 Excitation patterns for a 1,000 Hz sinusoid at levels ranging from 20 to 90 dB SPL in 10 dB steps.

that excitation patterns provide a reasonable estimate of the extent to which the spectral features of complex sounds are represented in the auditory system.

2.5 *Excitation pattern of a vowel sound*

The top panel of Figure 13.6 shows the spectrum of a synthetic vowel, /ɪ/ as in "bit," plotted on a linear frequency scale; this is the way that vowel spectra are often plotted. Each point represents the level of one harmonic in the complex sound (the fundamental frequency was 125 Hz). The middle panel shows the same spectrum plotted on an ERB$_N$-number scale; this gets somewhat closer to an auditory representation. The bottom panel shows the excitation pattern for the vowel, plotted on an ERB$_N$-number scale; this is closer still to an auditory representation. Several aspects of the excitation pattern are noteworthy. Firstly, the lowest few peaks in the excitation pattern do not correspond to formant frequencies, but rather to individual lower harmonics; these harmonics are resolved

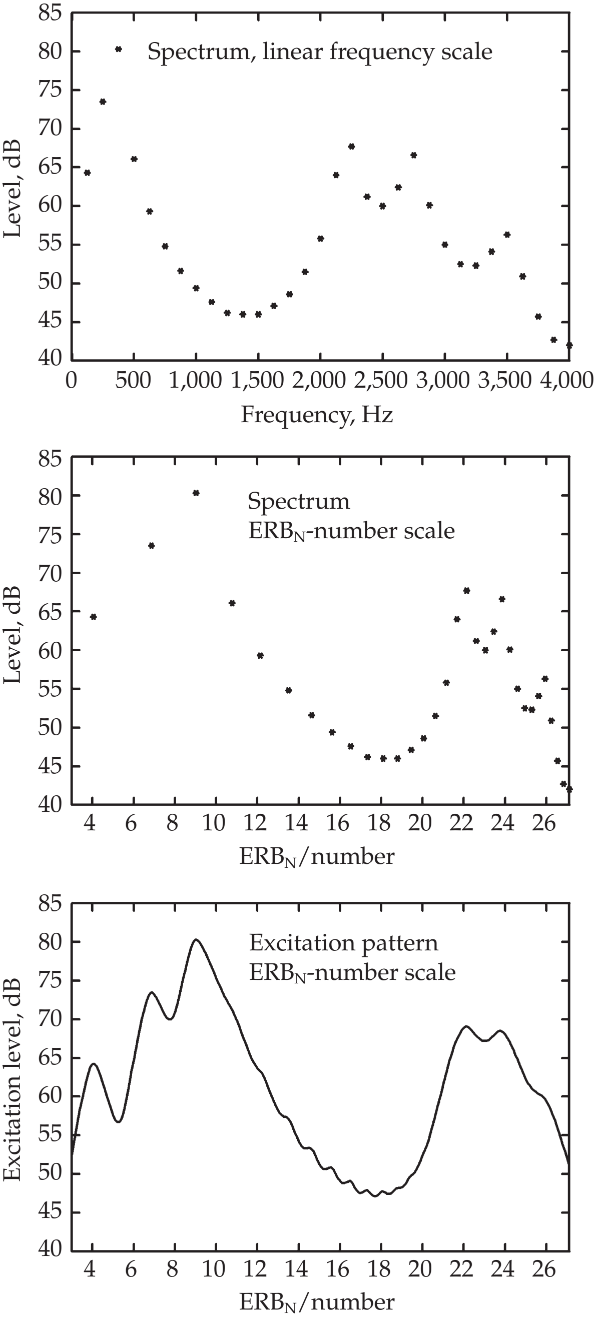


Figure 13.6 Top: the spectrum of a synthetic vowel /ɪ/ plotted on a linear frequency scale. Middle: the same spectrum plotted on an ERB$_N$-number scale. Bottom: the excitation pattern for the vowel plotted on an ERB$_N$-number scale.

in the peripheral auditory system, and can be heard out as separate tones under certain conditions (Plomp, 1964a; Moore & Ohgushi, 1993). Hence the center frequency of the first formant is not directly represented in the excitation pattern; if the frequency of the first formant is relevant for vowel identification, then it must be inferred from the relative levels of the peaks corresponding to the individual lower harmonics.

A second noteworthy aspect of the excitation pattern is that, for this specific vowel, the second, third, and fourth formants, which are clearly separately visible in the original spectrum, are not well resolved. Rather, they form a single prominence in the excitation pattern, with only minor ripples corresponding to the individual formants. Assuming that the excitation pattern does give a reasonable indication of the internal representation of the vowel, the perception of this vowel probably depends more on the overall prominence than on the frequencies of the individual formants. For other vowels, the higher formants often lead to separate peaks in the excitation pattern; see Figure 13.8.

3 Across-Channel Processes in Masking

The discrimination and identification of complex sounds, including speech, requires comparison of the outputs of different auditory filters. This section reviews data on across-channel processes in auditory masking, and their relevance for speech perception.

3.1 Co-modulation masking release

Hall et al. (1984) were among the first to demonstrate that across-filter comparisons could enhance the detection of a sinusoidal signal in a fluctuating noise masker. The crucial feature for achieving this enhancement was that the fluctuations should be correlated across different frequency bands. One of their experiments was similar to a classic experiment of Fletcher (1940). The threshold for detecting a 1,000 Hz, 400 ms sinusoidal signal was measured as a function of the bandwidth of a noise masker, keeping the spectrum level constant (the spectrum level is the level in a 1 Hz wide band, for example a band extending from 1,000 to 1,001 Hz). The masker was centered at 1,000 Hz. They used two types of masker. One was a random noise; this has irregular fluctuations in amplitude, and the fluctuations are independent in different frequency regions. The other was a random noise which was modulated in amplitude at an irregular, low rate; a noise lowpass filtered at 50 Hz was used as a modulator. The modulation resulted in fluctuations in the amplitude of the noise which were the same in different frequency regions. This across-frequency correlation was called "co-modulation" by Hall et al. (1984). Figure 13.7 shows the results of this experiment.

For the random noise (denoted by R), the signal threshold increases as the masker bandwidth increases up to about 100–200 Hz, and then remains constant, a result similar to that of Fletcher (1940). The value of $\mathrm{ERB_N}$ at this center frequency

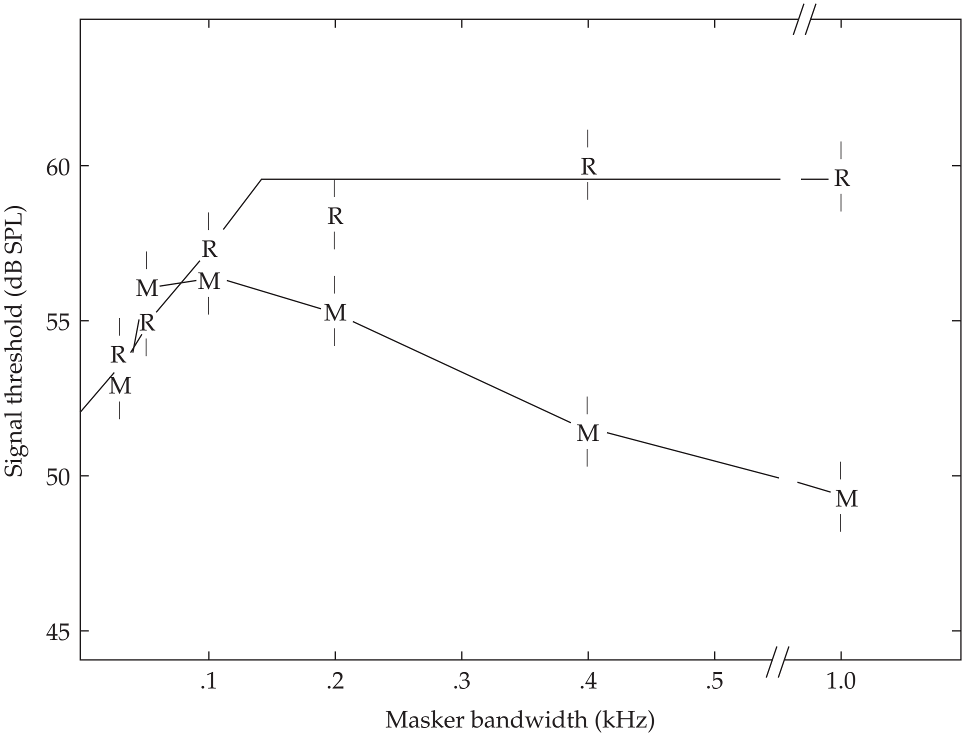


Figure 13.7 The points labeled "R" are thresholds for detecting a 1 kHz signal centered in a band of random noise, plotted as a function of the bandwidth of the noise. The points labeled "M" are the thresholds obtained when the noise was amplitude modulated at an irregular, low rate. (From Hall et al., 1984, by permission of the authors and the *Journal of the Acoustical Society of America*)

is about 130 Hz. Hence, for noise bandwidths up to 130 Hz, increasing the bandwidth results in more noise passing through the filter. However, increasing the bandwidth beyond 130 Hz does not substantially increase the noise power passing through the filter, so threshold does not increase. The pattern for the modulated noise (denoted by M) is quite different. For noise bandwidths greater than 100 Hz, the signal threshold decreases as the bandwidth increases. This suggests that subjects can compare the outputs of different auditory filters to enhance signal detection (see, however, Verhey et al. 1999). The fact that the decrease in threshold with increasing bandwidth only occurs with the modulated noise indicates that fluctuations in the masker are critical and that the fluctuations need to be correlated across frequency bands. Hence, this phenomenon has been called "co-modulation masking release" (CMR).

It seems likely that across-filter comparisons of temporal envelopes are a general feature of auditory pattern analysis, which may play an important role in extracting signals from noisy backgrounds, or separating competing sources of sound.

As pointed out by Hall et al. (1984, p. 56): "Many real-life auditory stimuli have intensity peaks and valleys as a function of time in which intensity trajectories are highly correlated across frequency. This is true of speech, of interfering noise such as 'cafeteria' noise, and of many other kinds of environmental stimuli." However, the importance of CMR for speech perception remains controversial. Some studies have suggested that it plays only a very minor role in the detection and identification of speech sounds in modulated background noise (Grose & Hall, 1992; Festen, 1993), although common modulation of target speech and background speech can lead to reduced intelligibility (Stone & Moore, 2004). For synthetic speech in which the cues are impoverished compared to normal speech (sinewave speech; see Remez et al., 1981), comodulation of the speech (amplitude modulation by a sinusoid) can markedly improve the intelligibility of the speech, both in quiet (Carrell & Opie, 1992) and in background noise (Carrell, 1993). The amplitude modulation may help because it leads to perceptual fusion of the components of the sinewave speech, so as to form an auditory object.

3.2 Profile analysis

Green and his colleagues (Green, 1988) have carried out a series of experiments demonstrating that, even for stimuli without distinct envelope fluctuations, subjects are able to compare the outputs of different auditory filters to enhance the detection of a signal. They investigated the ability to detect an increment in the level of one component in a complex sound relative to the level of the other components; the other components are called the "background." Usually the complex sound has been composed of a series of equal-amplitude sinusoidal components, uniformly spaced on a logarithmic frequency scale. To prevent subjects from performing the task by monitoring the magnitude of the output of the single auditory filter centered at the frequency of the incremented component, the overall level of the whole stimulus was varied randomly from one stimulus to the next, over a relatively large range (typically about 40 dB). This makes the magnitude of the output of any single filter an unreliable cue to the presence of the signal.

Subjects were able to detect changes in the relative level of the signal of only 1–2 dB. Such small thresholds could not be obtained by monitoring the magnitude of the output of a single auditory filter. Green and his colleagues have argued that subjects performed the task by detecting a change in the shape or profile of the spectrum of the sound; hence the name "profile analysis." In other words, subjects can compare the outputs of different auditory filters, and can detect when the output of one changes relative to that of others, even when the overall level is varied. This is equivalent to detecting changes in the shape of the excitation pattern.

Speech researchers will not find the phenomenon of profile analysis surprising. It has been known for many years that one of the main factors determining the timbre or quality of a sound is its spectral shape (see section 4). Our everyday experience tells us that we can recognize and distinguish familiar sounds, such

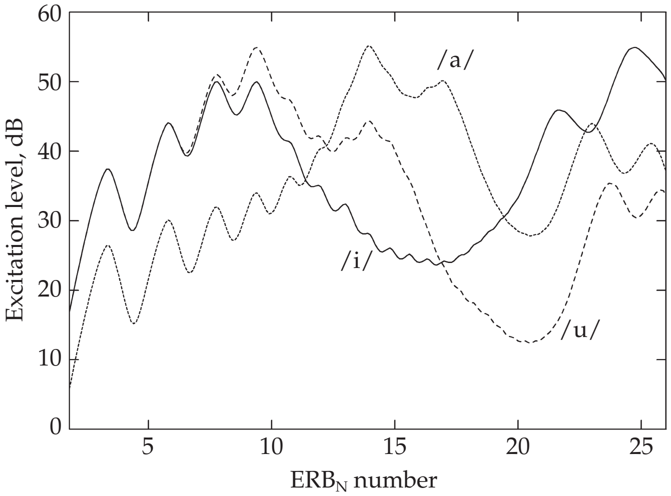


Figure 13.8 Excitation patterns for three vowels, /i/, /a/, and /u/, plotted on an ERB$_N$-number scale.

as the different vowels, regardless of the levels of those sounds. When we do this, we are distinguishing different spectral shapes in the face of variations in overall level. This is functionally the same as profile analysis. The experiments on profile analysis can be regarded as a way of quantifying the limits of our ability to distinguish changes in spectral shape. In this context, it is noteworthy that the differences in spectral shape between different vowels result in differences in the excitation patterns evoked by those sounds which are generally far larger than the smallest detectable changes as measured in profile analysis experiments. This is illustrated in Figure 13.8, which shows excitation patterns for three vowels, /i/, /a/, and /u/, plotted on an ERB$_N$-number scale. Each vowel had an overall level of about 58 dB SPL. It can be seen that the differences in the shapes of the excitation patterns are considerable.

3.3 Modulation discrimination interference

In some situations, the detection or discrimination of a signal is impaired by the presence of frequency components remote from the signal frequency. Usually, this happens when the task is either to detect modulation of the signal, or to detect a change in depth of modulation of the signal. Yost and Sheft (1989) showed that the threshold for detecting sinusoidal amplitude modulation (AM) of a sinusoidal carrier was increased in the presence of another carrier, amplitude modulated at the same rate, even when the second carrier was remote in frequency from the first. They called this modulation detection interference (MDI). They showed that MDI did not occur if the second carrier was unmodulated.

Moore et al. (1991) determined how thresholds for detecting an increase in modulation depth (sinusoidal AM or frequency modulation) of a 1,000 Hz carrier frequency (the target) were affected by modulation of carriers (interference) with frequencies of 230 Hz and 3,300 Hz. They found that modulation increment thresholds were increased (worsened) when the remote carriers were modulated. This MDI effect was greatest when the target and interference were modulated at similar rates, but the effect was broadly tuned for modulation rate. When the target and interfering sounds were both modulated at 10 Hz, there was no significant effect of the relative phase of modulation of the target and interfering sounds. A lack of effect of relative phase has also been found by other researchers (Moore, 1992; Hall et al., 1995).

The explanation for MDI remains unclear. Yost and Sheft (1989) suggested that MDI might be a consequence of perceptual grouping; the common AM of the target and interfering sounds might make them fuse perceptually, making it difficult to "hear out" the modulation of the target sound. However, certain aspects of the results on MDI are difficult to reconcile with an explanation in terms of perceptual grouping (Moore & Shailer, 1992). One would expect that widely spaced frequency components would only be grouped perceptually if their modulation pattern was very similar. Grouping would not be expected, for example, if the components were modulated out of phase or at different rates, but, in fact, it is possible to obtain large amounts of MDI under these conditions.

An alternative explanation for MDI is that it reflects the operation of "channels" specialized for detecting and analyzing modulation (Kay & Mathews, 1972; Dau et al., 1997a; 1997b). Yost et al. (1989) suggested that MDI might arise in the following way. The stimulus is first processed by an array of auditory filters. The envelope at the output of each filter is extracted. When modulation is present, channels are excited that are tuned for modulation rate. All filters responding with the same modulation rate excite the same channel, regardless of the filter center frequency. Thus, modulation at one center frequency can adversely affect the detection and discrimination of modulation at other center frequencies.

The purpose of the hypothetical modulation channels remains unclear. Since physiological evidence suggests that such channels exist in animals (Schreiner & Urbas, 1986; Langner & Schreiner, 1988), we can assume that they did not evolve for the purpose of speech perception. Nevertheless, it is possible, even likely, that speech analysis makes use of the modulation channels. There is evidence that amplitude modulation patterns in speech are important for speech recognition (Steeneken & Houtgast, 1980; Drullman et al., 1994a; Shannon et al., 1995). Thus, anything that adversely affects the detection and discrimination of the modulation patterns would be expected to impair intelligibility. One way of describing MDI is: modulation in one frequency region may make it more difficult to detect and discriminate modulation in another frequency region. Thus, it may be the case that MDI makes speech recognition more difficult in situations where there is a background sound that is modulated, such as one or more people talking (Brungart et al., 2005).

4 Timbre Perception

Timbre is usually defined as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" (ANSI, 1994). The distribution of energy over frequency is one of the major determinants of timbre. However, timbre depends upon more than just the frequency spectrum of the sound; fluctuations over time can play an important role, as discussed below.

4.1 *Role of spectral shape for steady sounds*

Timbre is multidimensional; there is no single scale along which the timbres of different sounds can be compared or ordered. Thus, a way is needed of describing the spectrum of a sound which takes into account this multidimensional aspect, and which can be related to the subjective timbre. For steady sounds, a crude first approach is to look at the overall distribution of spectral energy. The "brightness" or "sharpness" (von Bismarck, 1974) of sounds seems to be related to the spectral centroid. However, a much more quantitative approach has been described by Plomp and his colleagues (Plomp, 1970, 1976). They showed that the perceptual differences between different steady sounds, such as vowels, were closely related to the differences in the spectra of the sounds, when the spectra were specified as the levels in 18 frequency bands, each ⅓-octave wide. A bandwidth of ⅓ octave is slightly greater than the $\mathrm{ERB_N}$ of the auditory filter over most of the audible frequency range. Thus, timbre is related to the relative level produced at the output of each auditory filter. Put another way, the timbre of a steady sound is related to the excitation pattern of that sound.

It is likely that the number of dimensions required to characterize the timbre of steady sounds is limited by the number of $\mathrm{ERB_N}$s required to cover the audible frequency range. This would give a maximum of about 37 dimensions. For a restricted class of sounds, such as vowels, a much smaller number of dimensions may be involved. It appears to be generally true, both for speech and nonspeech sounds, that the timbres of steady tones are determined primarily by their magnitude spectra, although the relative phases of the components also play a small role (Plomp & Steeneken, 1969; Patterson, 1987).

4.2 *Other factors affecting timbre*

Differences in spectral shape are not always sufficient to allow the absolute identification of an "auditory object," such as a musical instrument or a speech sound. One reason for this is that the magnitude and phase spectrum of the sound may be markedly altered by the transmission path and room reflections (Watkins, 1991). In practice, the recognition of a particular timbre, and hence of an "auditory object," may depend upon several other factors. Schouten (1968) has suggested that these include: (1) whether the sound is periodic, having a tonal quality for repetition rates between about 20 and 20,000 per second, or irregular, and having a noise-like character; (2) whether the waveform envelope is constant, or fluctuates as a function of time, and in the latter case what the fluctuations are like; (3) whether any other aspect of the sound (e.g., spectrum or periodicity) is changing as a function of time; (4) what the preceding and following sounds are like.

A powerful demonstration of the last factor may be obtained by listening to a stimulus with a particular spectral structure and then switching rapidly to a stimulus with a flat spectrum, such as white noise. A white noise heard in isolation may be described as "colorless"; it has no pitch and has a neutral timbre. However, when a white noise follows immediately after a stimulus with spectral structure, the noise sounds "colored." The coloration corresponds to the inverse of the spectrum of the preceding sound. For example, if the preceding sound is a noise with a spectral notch, the white noise has a pitch-like quality, with a pitch value corresponding to the center frequency of the notch (Zwicker, 1964). It sounds like a noise with a small spectral peak. A harmonic complex tone with a flat spectrum may be heard as having a vowel-like quality if it is preceded by a harmonic complex having a spectrum which is the inverse of that of a vowel (Summerfield et al., 1987).

The cause of this effect is not clear. Three types of explanation have been advanced, based on adaptation in the auditory periphery, perceptual grouping, and comparison of spectral shapes of the preceding and test sounds (Summerfield & Assmann, 1987). All may play a role to some extent, depending on the exact properties of the stimuli. Whatever the underlying mechanism, it seems clear that the auditory system is especially sensitive to *changes* in spectral patterns over time (Kluender et al., 2003). This may be of value for communication in situations where the spectral shapes of sounds are (statically) altered by room reverberation or by a transmission channel with a nonflat frequency response.

4.3 Perceptual compensation for spectral distortion

Perceptual compensation for the effects of a nonflat frequency response has been studied extensively by Watkins and co-workers (Watkins, 1991; Watkins & Makin, 1996a, 1996b). In one series of experiments (Watkins & Makin, 1996b), they investigated how the identification of vowel test sounds was affected by filtering of preceding and following sounds. All sounds were edited and processed from natural speech spoken with a British accent. Listeners identified words from continua between /ɪtʃ/ and /ɛtʃ/ (itch and etch), /æpt/ and /ɒpt/ (apt and opt), or /sləʊ/ and /fləʊ/ (slow and flow). The parts of the stimuli other than the vowels (e.g., the /tʃ/ or the /pt/) were filtered with complex frequency responses corresponding to the difference of spectral envelopes from the endpoint test sounds (the vowels). An example of a "difference filter" is shown in the bottom panel of Figure 13.9. The shift in the phoneme boundary of the vowels was used to measure perceptual compensation for the effects of the spectral distortion of the consonants. When the words were presented without a precursor phrase, the results indicated perceptual compensation. Thus, information from the consonants

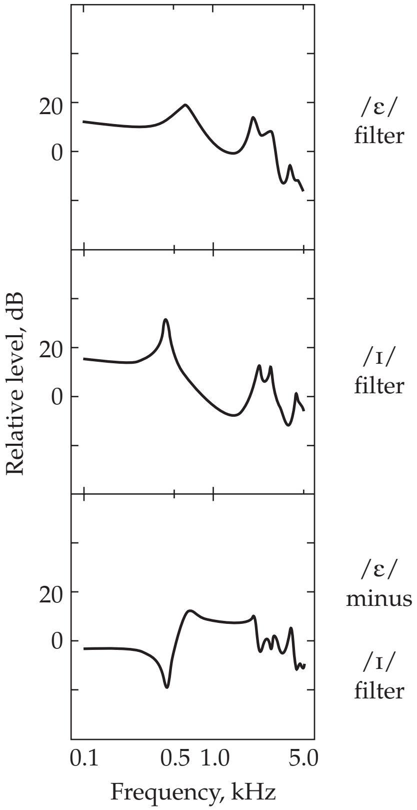


Figure 13.9 Illustration of the filters used by Watkins and Makin (1996a). The top and middle panels show "filters" corresponding to the spectral envelopes of the vowels /ɛ/ and /ɪ/. The bottom panel shows the filter corresponding to the difference between the spectral envelopes of the /ɛ/ and /ɪ/.

modified the perception of the preceding or following vowels. When a precursor phrase "the next word is" was used, and was filtered in the same way as the consonant, the effects were larger. However, the large effects were somewhat reduced when the precursor phrase was filtered but the following consonant was not. This clearly indicates a role for sounds following a vowel. The effects of following sounds found in these experiments clearly indicate that factors other than adaptation play a role. Presumably, these effects reflect relatively central perceptual compensation mechanisms.

Overall, the results described in this section indicate that the perceived timbre of brief segments of sounds can be strongly influenced by sounds that precede and follow those segments. In some cases, the observed effects appear to reflect relatively central perceptual compensation processes.

5 The Perception of Pitch

Pitch is usually defined as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from low to high" (ANSI, 1994). In other words, variations in pitch give rise to a sense of melody. For speech sounds, variations in voice pitch over time convey intonation information, indicating whether an utterance is a question or a statement, and helping to identify stressed words. Voice pitch can also convey information about the sex, age, and emotional state of the speaker (Rosen & Fourcin, 1986). In some languages ("tone" languages), pitch and variations in pitch distinguish different lexical items.

Pitch is related to the repetition rate of the waveform of a sound; for a pure tone this corresponds to the frequency and for a periodic complex tone to the fundamental frequency, f_0. There are, however, exceptions to this simple rule. Since voiced speech sounds are complex tones, this section will focus on the perception of pitch for complex tones.

5.1 *The phenomenon of the missing fundamental*

Although the pitch of a complex tone usually corresponds to its f_0, the component with frequency equal to f_0 does not have to be present for the pitch to be heard. Consider, as an example, a sound consisting of short impulses (clicks) occurring 200 times per second. This sound has a low pitch, which is very close to the pitch of a 200 Hz sinusoid, and a sharp timbre. It contains harmonics with frequencies 200, 400, 600, 800 . . . etc. Hz. However, if the sound is filtered so as to remove the 200 Hz component, the pitch does not alter; the only result is a slight change in the timbre of the note. Indeed, all except a small group of mid-frequency harmonics can be eliminated, and the low pitch still remains, although the timbre becomes markedly different.

Schouten (1970) called the low pitch associated with a group of high harmonics the "residue." He pointed out that the residue is distinguishable, subjectively, from a fundamental component which is physically presented or from a fundamental which may be generated (at high sound pressure levels) by nonlinear distortion in the ear. Thus the perception of a residue pitch does not require activity at the point on the basilar membrane which would respond maximally to a pure tone of similar pitch. Several other names have been used to describe residue pitch, including "periodicity pitch," "virtual pitch," and "low pitch." This chapter will use the term low pitch. Even when the fundamental component of a complex tone is present, the low pitch of the tone is usually determined by harmonics other than the fundamental. Thus the perception of a low pitch should not be regarded as unusual. Rather, low pitches are normally heard when listening to complex tones, including speech. For example, when listening over the telephone, the fundamental component for male speakers is usually inaudible, but the pitch of the voice can still be easily heard.

5.2 The principle of dominance

Ritsma (1967) carried out an experiment to determine which components in a complex sound are most important in determining its pitch. He presented complex tones in which the frequencies of a small group of harmonics were multiples of an f_0 which was slightly higher or lower than the f_0 of the remainder. The subject's pitch judgments were used to determine whether the pitch of the complex as a whole was affected by the shift in the group of harmonics. Ritsma (1967, p. 197) found that: "For fundamental frequencies in the range 100 Hz to 400 Hz, and for sensation levels up to at least 50 dB above threshold of the entire signal, the frequency band consisting of the third, fourth and fifth harmonics tends to dominate the pitch sensation as long as its amplitude exceeds a minimum absolute level of about 10 dB above threshold."

This finding has been broadly confirmed in other ways (Plomp, 1967), although the data of Moore et al. (1984, 1985) show that there are large individual differences in which harmonics are dominant, and for some subjects the first two harmonics play an important role. Other data also show that the dominant region is not fixed in terms of harmonic number, but depends somewhat on absolute frequency (Plomp, 1967; Patterson & Wightman, 1976). For high f_0s (above about 1,000 Hz), the fundamental is usually the dominant component, while for very low f_0s, around 50 Hz, harmonics above the fifth may be dominant (Moore & Glasberg, 1988; Moore & Peters, 1992). Finally, the dominant region shifts somewhat towards higher harmonics with decreasing duration (Gockel et al., 2005). For speech sounds, the dominant harmonics usually lie around the frequency of the first formant.

5.3 Discrimination of the pitch of complex tones

When the f_0 of a periodic complex tone changes, all of the components change in frequency by the same ratio, and a change in low pitch is heard. The ability to detect such changes is better than the ability to detect changes in a sinusoid at f_0 (Flanagan & Saslow, 1958) and can be better than the ability to detect changes in the frequency of any of the sinusoidal components in the complex tone (Moore et al., 1984). This indicates that information from the different harmonics is combined or integrated in the determination of low pitch. This can lead to very fine discrimination; changes in f_0 of about 0.2 percent can often be detected for f_0s in the range 100–400 Hz.

The discrimination of f_0 is usually best when low harmonics are present (Hoekstra & Ritsma, 1977; Moore & Glasberg, 1988; Shackleton & Carlyon, 1994). Somewhat less good discrimination (typically 1–4 percent) is possible when only high harmonics are present (Houtsma & Smurzynski, 1990). f_0 discrimination can be impaired (typically by about a factor of two) when the two sounds to be discriminated also differ in timbre (Moore & Glasberg, 1990); this can be the situation with speech sounds, where changes in f_0 are usually accompanied by changes in timbre.

In speech, intonation is typically conveyed by differences in the pattern of f_0 change over time. When the stimuli are dynamically varying, the ability to detect f_0 changes is markedly poorer than when the stimuli are steady. Klatt (1973) measured thresholds for detecting differences in f_0 for an unchanging vowel (i.e., one with static formant frequencies) with a flat f_0 contour, and also for a series of linear glides in f_0 around an f_0 of 120 Hz. For the flat contour, the threshold was about 0.3 Hz. When both contours were falling at the same rate (30 Hz over the 250 ms duration of the stimulus), the threshold increased markedly to 2 Hz. When the steady vowel was replaced by the sound /ya/, whose formants change over time, thresholds increased further by 25–65 percent.

Generally, the f_0 changes that are linguistically relevant for conveying stress and intonation are much larger than the limits of f_0 discrimination measured psychophysically using steady stimuli. This is another reflection of the fact that information in speech is conveyed using robust cues that do not severely tax the discrimination abilities of the auditory system.

It should be noted that in natural speech the period (corresponding to the time between successive closures of the vocal folds) varies randomly from one period to the next (Fourcin & Abberton, 1977). This jitter conveys information about the emotional state of the talker, and is required for a natural voice quality to be perceived. Human listeners can detect jitter of 1–2 percent (Pollack, 1968; Kortekaas & Kohlrausch, 1999). Large amounts of jitter are associated with voice pathologies, such as hoarseness (Yumoto et al., 1982).

5.4 Perception of pitch in speech

Data on the perception of f_0 contours in a relatively natural speech context were presented by Pierrehumbert (1979). She started with a natural nonsense utterance "ma-MA-ma-ma-MA-ma," in which the prosodic pattern was based on the sentence "The baker made bagels." The stressed syllables (MA) were associated with peaks in the f_0 contour. She then modified the f_0 of the second peak, over a range varying from below to above the f_0 of the first peak. Subjects listened to the modified utterances, and were required to indicate whether the first or second peak was higher in pitch. The results reflected what she called "normalization for expected declination"; when the two stressed syllables sounded equal in pitch, the second was actually lower in f_0. For first peak values of 121 and 151 Hz, the second peak had to be shifted over a range of about 20 Hz to change judgments from 75 percent "second peak lower" to 75 percent "second peak higher." This indicates markedly poorer discriminability than found for steady stimuli. Similarly, 't Hart (1981) found that about a 19 percent difference was necessary for successive pitch movements in the same direction to be reliably heard as different in extent.

Hermes and van Gestel (1991) studied the perception of the excursion size of prominence-lending f_0 movements in utterances resynthesized in different f_0 registers. The task of the subjects was to adjust the excursion size in a comparison stimulus in such a way that it lent equal prominence to the corresponding syllable in a fixed test stimulus. The comparison stimulus and the test stimulus had f_0s running parallel on either a logarithmic frequency scale, an ERB_N-number scale, or a linear frequency scale. They found that stimuli were matched in such a way that the average excursion sizes in different registers were equal when the ERB_N-number scale was used. Put another way, the perceived prominence of f_0 movements is related to the size of those movements expressed on an ERB_N-number scale.

6 Temporal Analysis

Time is a very important dimension in hearing, since almost all sounds change over time. For speech, much of the information appears to be carried in the changes themselves, rather than in the parts of the sounds which are relatively stable (Kluender et al., 2003). In characterizing temporal analysis, it is essential to take account of the filtering that takes place in the peripheral auditory system. Temporal analysis can be considered as resulting from two main processes: analysis of the time pattern occurring within each frequency channel; and comparison of the time patterns across channels. This chapter focuses on the first of these.

A major difficulty in measuring the temporal resolution of the auditory system is that changes in the time pattern of a sound are generally associated with changes in its magnitude spectrum – the distribution of energy over frequency. Thus, the detection of a change in time pattern can sometimes depend not on temporal resolution per se, but on the detection of the spectral change. Sometimes, the detection of spectral changes can lead to what appears to be extraordinarily fine temporal resolution. For example, a single click can be distinguished from a pair of clicks when the gap between the two clicks in a pair is only a few tens of microseconds, an ability that depends upon spectral changes at very high frequencies (Leshowitz, 1971). Although spectrally based detection of temporal changes can occur for speech sounds, this chapter focuses on experimental situations which avoid the confounding effects of spectral cues.

There have been two general approaches to avoiding the use of cues based on spectral changes. One is to use signals whose magnitude spectrum is not changed when the time pattern is altered. For example, the magnitude spectrum of white noise remains flat if a gap is introduced into the noise. The second approach uses stimuli whose spectra are altered by the change in time pattern, but extra background sounds are added to mask the spectral changes. Both of these approaches will be considered.

6.1 Within-channel temporal analysis using broadband sounds

The experiments described next all use broadband sounds whose long-term magnitude spectrum is unaltered by the temporal manipulation being performed. For example, interruption or amplitude modulation of a white noise does not change its long-term magnitude spectrum, and time-reversal of any sound also does not change its long-term magnitude spectrum.

The threshold for detecting a gap in a broadband noise provides a simple and convenient measure of temporal resolution. The gap threshold is typically 2–3 ms (Plomp, 1964b). The threshold increases at very low sound levels, when the level of the noise approaches the absolute threshold, but is relatively invariant with level for moderate to high levels.

Ronken (1970) used as stimuli pairs of clicks differing in amplitude. One click, labeled A, had an amplitude greater than that of the other click, labeled B. Typically the amplitude of A was twice that of B. Subjects were required to distinguish click pairs differing in the order of A and B: either AB or BA. The ability to do this was measured as a function of the time interval or gap between A and B. Ronken found that subjects could distinguish the click pairs for gaps down to 2–3 ms. Thus the limit to temporal resolution found in this task is similar to that found for the detection of a gap in broadband noise. It should be noted that, in this task, subjects do not hear the individual clicks within a click pair. Rather, each click pair is heard as a single sound with its own characteristic quality. For example, the two click pairs AB and BA might sound like "tick" and "tock."

The experiments described above each give a single value to describe temporal resolution. A more general approach is to measure the threshold for detecting changes in the amplitude of a sound as a function of the rapidity of the changes. In the simplest case, white noise is sinusoidally amplitude modulated, and the threshold for detecting the modulation is determined as a function of modulation rate. The function relating threshold to modulation rate is known as a temporal modulation transfer function (TMTF; Viemeister, 1979). An example of a TMTF is shown in Figure 13.10 (data from Bacon & Viemeister, 1985). The thresholds are expressed as $20\log m$, where m is the modulation index ($m = 0$ corresponds to no modulation and $m = 1$ corresponds to 100 percent modulation). For low modulation rates, performance is limited by the amplitude resolution of the ear, rather than by temporal resolution. Thus, the threshold is independent of modulation rate for rates up to about 50 Hz. As the rate increases beyond 50 Hz, temporal resolution starts to have an effect; performance worsens, and for rates above about 1,000 Hz the modulation is hard to detect at all. Thus, sensitivity to modulation becomes progressively lower as the rate of modulation increases. The shapes of TMTFs do not vary much with overall sound level, but the ability to detect the modulation does worsen at low sound levels. Over the range of modulation rates important for speech perception, below about 50 Hz (Steeneken & Houtgast, 1980; Drullman et al., 1994a, 1994b), the sensitivity to modulation is rather good.

6.2 *Within-channel temporal analysis using narrowband sounds*

Experiments using broadband sounds provide no information regarding the question of whether the temporal resolution of the auditory system varies with center

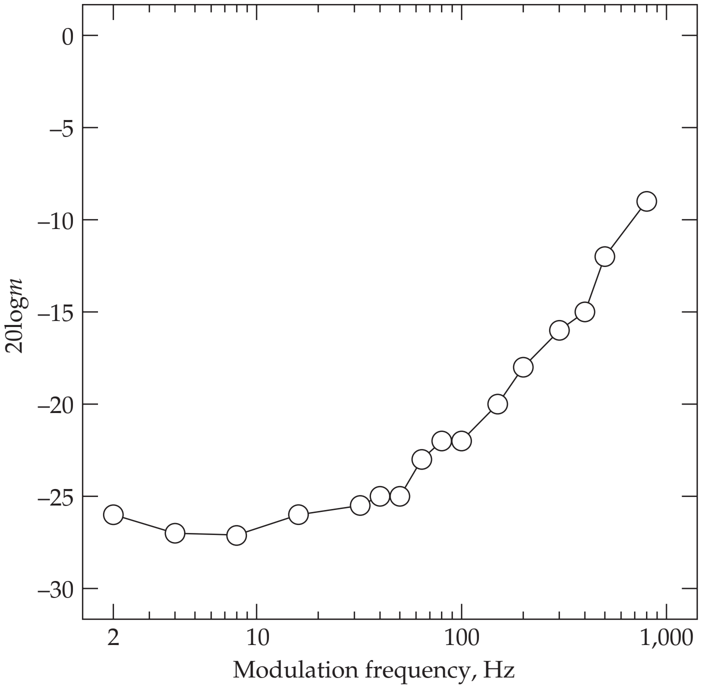


Figure 13.10 A temporal modulation transfer function (TMTF). A broadband white noise was sinusoidally amplitude modulated, and the threshold amount of modulation required for detection is plotted as a function of modulation rate. The amount of modulation is specified as 20log*m*, where *m* is the modulation index. The higher the sensitivity to modulation, the more negative is this quantity. (Data from Bacon & Viemeister, 1985)

frequency. This issue can be examined by using narrowband stimuli that excite a limited number of auditory filters.

Green (1973) used stimuli where each stimulus consisted of a brief pulse of a sinusoid in which the level of the first half of the pulse was 10 dB different from that of the second half. Subjects were required to distinguish two signals, differing in whether the half with the high level was first or second. Green measured performance as a function of the total duration of the stimuli. The threshold was similar for center frequencies of 2 and 4 kHz, and was between 1 and 2 ms. However, the threshold was slightly higher for a center frequency of 1 kHz, being between 2 and 4 ms.

Performance in this task was actually a nonmonotonic function of duration. Performance was good for durations in the range 2–6 ms, worsened for durations around 16 ms, and then improved again as the duration was increased beyond 16 ms. For the very short durations, subjects listened for a difference in quality between the two sounds – rather like the "tick" and "tock" described earlier for Ronken's stimuli. At durations around 16 ms, the tonal quality of the bursts became more prominent, and the quality differences were harder to hear. At much longer durations the soft and loud segments could be separately heard, in a distinct order. It appears, therefore, that performance in this task was determined by two separate mechanisms, one based on timbre differences associated with the difference in time pattern, and the other based on the perception of a distinct succession of auditory events.

Several researchers have measured thresholds for detecting gaps in narrowband sounds, either noises (Fitzgibbons, 1983; Shailer & Moore, 1983; Buus & Florentine, 1985; Eddins et al., 1992) or sinusoids (Shailer & Moore, 1987; Moore et al., 1993). When a temporal gap is introduced into a narrowband sound, the spectrum of the sound is altered. Energy "splatter" occurs outside the nominal frequency range of the sound. To prevent the splatter being detected, the sounds are presented in a background sound, usually a noise, designed to mask the splatter.

Gap thresholds for noise bands decrease with increasing bandwidth but show little effect of center frequency when the bandwidth is held constant. For noises of moderate bandwidth (a few hundred Hz), the gap threshold is typically about 10 ms. Gap thresholds for narrowband noises tend to decrease with increasing sound level for levels up to about 30 dB above absolute threshold, but remain roughly constant after that.

Shailer and Moore (1987) showed that the detectability of a gap in a sinewave was strongly affected by the phase at which the sinusoid was turned off and on to produce the gap (Shailer & Moore, 1987). Only the simplest case is considered here, called "preserved phase" by Shailer and Moore (1987). In this case the sinusoid was turned off at a positive-going zero crossing (i.e., as the waveform was about to change from negative to positive values) and it started (at the end of the gap) at the phase it would have had if it had continued without interruption. Thus, for the preserved-phase condition it was as if the gap had been "cut out" from a continuous sinusoid. For this condition, the detectability of the gap increased monotonically with increasing gap duration.

Shailer and Moore (1987) found that the threshold for detecting a gap in a sinewave was roughly constant at about 5 ms for center frequencies of 400, 1,000, and 2,000 Hz. Moore et al. (1993) found that gap thresholds were almost constant at 6–8 ms over the frequency range 400–2,000 Hz, but increased somewhat at 200 Hz, and increased markedly, to about 18 ms, at 100 Hz. Individual variability also increased markedly at 100 Hz.

Overall, the results of experiments using narrowband stimuli indicate that temporal resolution does not vary markedly with center frequency, except perhaps for a worsening at very low frequencies (200 Hz and below). Gap thresholds for narrowband stimuli are typically higher than those for broadband noise. However, for moderate noise bandwidths, gap thresholds are typically around 10 ms or less. The smallest detectable gap is usually markedly larger than temporal gaps that are relevant for speech perception (for example, "sa" and "sta" may be distinguished by a temporal gap lasting several tens of milliseconds).

6.3 *Modeling temporal resolution*

Most models of temporal resolution are based on the idea that there is a process at levels of the auditory system higher than the auditory nerve which is "sluggish" in some way, thereby limiting temporal resolution. The models assume that the internal representation of stimuli is "smoothed" over time, so that rapid temporal changes are reduced in magnitude but slower ones are preserved. Although this smoothing process almost certainly operates on neural activity, the most widely used models are based on smoothing a simple transformation of the stimulus, rather than its neural representation.

Most models include an initial stage of bandpass filtering, reflecting the action of the auditory filters. Each filter is followed by a nonlinear device. This nonlinear device is meant to reflect the operation of several processes that occur in the peripheral auditory system such as amplitude compression on the basilar membrane and neural transduction, whose effects resemble half-wave rectification. The output of the nonlinear device is fed to a "smoothing" device, which can be implemented either as a lowpass filter (Viemeister, 1979) or (equivalently) as a sliding temporal integrator (Moore et al., 1988; Plack & Moore, 1990). The device determines a kind of weighted average of the output of the compressive nonlinearity over a certain time interval or "window." This weighting function is sometimes called the "shape" of the temporal window. The window is assumed to slide in time, so that the output of the temporal integrator is a weighted running average of the input. This has the effect of smoothing rapid fluctuations while preserving slower ones. When a sound is turned on abruptly, the output of the temporal integrator takes some time to build up. Similarly, when a sound is turned off, the output of the integrator takes some time to decay. The shape of the window is assumed to be asymmetric in time, such that the build up of its output in response to the onset of a sound is more rapid than the decay of its output in response to the cessation of a sound. The output of the sliding temporal integrator is fed to a decision device. The decision device may use different "rules" depending on the task required. For example, if the task is to detect a brief temporal gap in a signal, the decision device might look for a "dip" in the output of the temporal integrator. If the task is to detect amplitude modulation of a sound, the device might assess the amount of modulation at the output of the sliding temporal integrator (Viemeister, 1979).

7 Calculation of the Internal Representation of Sounds

I describe next a method of calculating the internal representation of sounds, including speech, based on processes that are known to occur in the auditory system and taking into account the frequency and temporal resolution of the auditory system. The spectrogram is often regarded as a crude representation of the spectro-temporal analysis that takes place in the auditory system, although this representation is inaccurate in several ways (Moore, 2003a). I outline below a model that probably gives a better representation, although it is still oversimplified in several respects. The model is based on the assumption that there are certain "fixed" processes in the peripheral auditory system, which can be modeled as a series of stages including:

1 Fixed filters representing transfer of sound through the outer ear (Shaw, 1974) and middle ear (Aibara et al., 2001). The overall transfer function through the outer and middle ear for a frontally incident sound in free field has been estimated by Glasberg and Moore (2002) and is shown in their figure 1. The effect of this transfer function is that low frequencies (below 500 Hz) and high frequencies (above 5,000 Hz) are attenuated relative to middle frequencies.
2 An array of bandpass filters (the auditory filters).
3 Each auditory filter is followed by nonlinear processes reflecting the compression that occurs on the basilar membrane (Oxenham & Moore, 1994; Ruggero et al., 1997). The compression is weak for very low sound levels (below about 30 dB SPL), and perhaps for very high levels (above 90 dB SPL), but it has a strong influence for mid-range sound levels. Half-wave or full-wave rectification may also be introduced, to mimic the transformation from basilar-membrane vibration to neural activity effected by the inner hair cells.
4 An array of devices (sliding temporal integrators) that "smooth" the output of each nonlinearity. As described earlier, the smoothing is assumed to reflect a relatively central process, occurring after the auditory nerve.

In some models, the filtering and the compressive nonlinearity are combined in a single nonlinear filter bank (Irino & Patterson, 2001; Lopez-Poveda & Meddis, 2001; Zhang et al., 2001). Also, the transformation from basilar-membrane vibration to neural activity can be simulated more accurately using a hair-cell model (Sumner et al., 2002). However, the basic features of the internal representation can be represented reasonably well using models of the type defined by stages (1)–(4) above. The internal representation of a given stimulus can be thought of as a three-dimensional array with center frequency as one axis (corresponding to the array of auditory filters with different center frequencies), and time and magnitude as the other axes (corresponding to the output of each temporal integrator plotted as a function of time). The resulting pattern can be called a spectro-temporal excitation pattern (STEP) (Moore, 1996). An example is shown in Figure 13.11, adapted from Moore (2003c). The figure shows the calculated STEP of the word *tips*. In this figure, the frequency scale has been transformed to an ERB_N-number scale, as described earlier. The corresponding frequency is also shown.

It should be noted that the STEP does not represent information that is potentially available in the temporal "fine structure" at the output of each auditory filter. The role of this fine structure in speech perception is uncertain, and some have suggested that it plays little role (Shannon et al., 1995). However, temporal

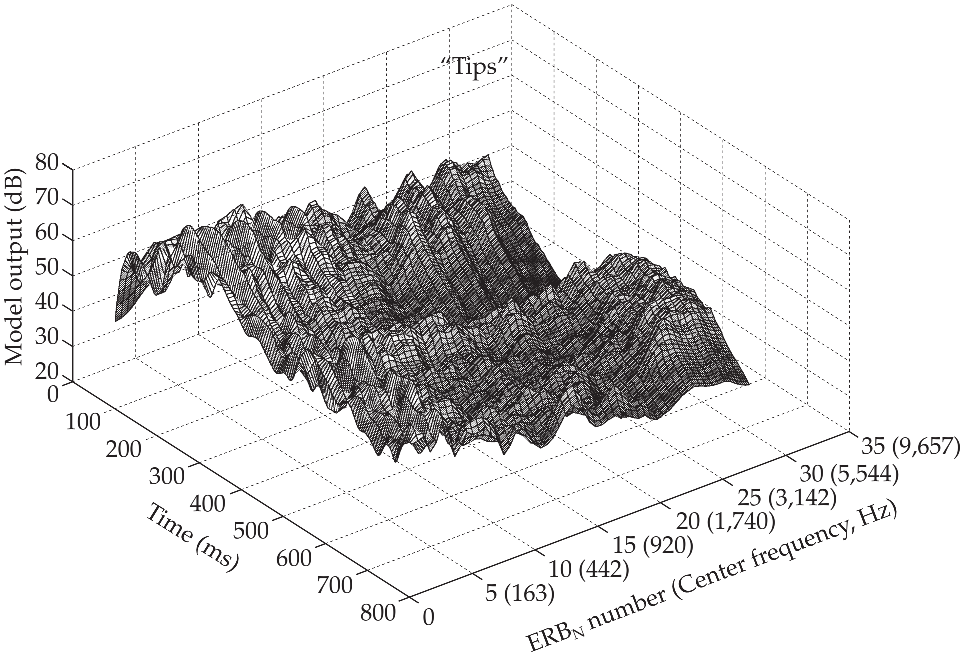


Figure 13.11 Spectro-temporal excitation pattern (STEP) of the word *tips*. The figure was produced by Prof. C. J. Plack. (Adapted from Moore, 2003c)

fine structure may play a role in the perception of pitch (Moore et al., 2006), in the separation of simultaneous sounds (Moore, 2003a) and especially in the process of understanding one talker in the presence of another talker (Lorenzi et al., 2006).

8 Concluding Remarks

This chapter has reviewed several aspects of auditory perception that are relevant to the perception of speech. These aspects include frequency selectivity, timbre perception, the perception of pitch, and temporal analysis. A recurring theme has been the finding that the basic discrimination abilities of the auditory system, measured using simple nonspeech stimuli, are very good when considered relative to the acoustic differences that distinguish speech sounds. This partially accounts for the robust nature of speech perception. Indeed, it is remarkable that speech remains reasonably intelligible even under conditions of extreme distortion, such as infinite peak clipping (Licklider & Pollack, 1948), time reversal of segments of speech (Saberi & Perrott, 1999), representation of speech by three or four sinewaves tracking the formant frequencies (Remez et al., 1981), or representation of speech by a few amplitude-modulated noise bands (Shannon et al., 1995). However, this robustness applies to speech presented in quiet. Speech is often heard under much less ideal conditions. For example, reverberation, background noise, and competing talkers may be present. Under these conditions, many of the cues in speech become less discriminable, and some cues may be completely inaudible. Speech perception then becomes much less robust, especially if the functioning of the auditory system is impaired (Moore, 2003b).

NOTE

Parts of this chapter are reproduced in the following: Moore, B. C. J. (2008) Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society B*, 363, 947–63. I would like to thank Tom Baer and Chris Darwin for helpful comments on an earlier version of this chapter.

REFERENCES

Aibara, R., Welsh, J. T., Puria, S., & Goode, R. L. (2001) Human middle-ear sound transfer function and cochlear input impedance. *Hearing Research*, 152, 100–9.

Alcántara, J. I., Moore, B. C. J., & Vickers, D. A. (2000) The relative role of beats and combination tones in determining the shapes of masking patterns at 2 kHz, I: Normal-hearing listeners. *Hearing Research*, 148, 63–73.

ANSI (1994) *ANSI S1.1-1994. American National Standard Acoustical Terminology*. New York: American National Standards Institute.

Bacon, S. P. & Viemeister, N. F. (1985) Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects. *Audiology*, 24, 117–34.

Bismarck, G. von (1974) Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30, 159–72.

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., & Kidd, G., Jr. (2005) Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task. *Journal of the Acoustical Society of America*, 117, 292–304.

Buus, S. & Florentine, M. (1985) Gap detection in normal and impaired listeners: The effect of level and frequency. In A. Michelsen (ed.), *Time Resolution in Auditory Systems* (pp. 159–79). New York: Springer.

Carrell, T. (1993) The effect of amplitude comodulation on extracting sentences from noise: Evidence from a variety of contexts. *Journal of the Acoustical Society of America*, 93, 2327.

Carrell, T. D. & Opie, J. M. (1992) The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception and Psychophysics*, 52, 437–45.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997a) Modeling auditory processing of amplitude modulation, I: Detection and masking with narrowband carriers. *Journal of the Acoustical Society of America*, 102, 2892–905.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997b) Modeling auditory processing of amplitude modulation, II: Spectral and temporal integration. *Journal of the Acoustical Society of America*, 102, 2906–19.

Delgutte, B. (1990) Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. *Journal of the Acoustical Society of America*, 87, 791–809.

Drullman, R., Festen, J. M., & Plomp, R. (1994a) Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95, 2670–80.

Drullman, R., Festen, J. M., & Plomp, R. (1994b) Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95, 1053–64.

Eddins, D. A., Hall, J. W., & Grose, J. H. (1992) Detection of temporal gaps as a function of frequency region and absolute noise bandwidth. *Journal of the Acoustical Society of America*, 91, 1069–77.

Egan, J. P. & Hake, H. W. (1950) On the masking pattern of a simple auditory stimulus. *Journal of the Acoustical Society of America*, 22, 622–30.

Festen, J. M. (1993) Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice. *Journal of the Acoustical Society of America*, 94, 1295–300.

Fitzgibbons, P. J. (1983) Temporal gap detection in noise as a function of frequency, bandwidth and level. *Journal of the Acoustical Society of America*, 74, 67–72.

Flanagan, J. L. & Saslow, M. G. (1958) Pitch discrimination for synthetic vowels. *Journal of the Acoustical Society of America*, 30, 435–42.

Fletcher, H. (1940) Auditory patterns. *Reviews of Modern Physics*, 12, 47–65.

Fourcin, A. J. & Abberton, E. (1977) Laryngograph studies of vocal-fold vibration. *Phonetica*, 34, 313–15.

Glasberg, B. R. & Moore, B. C. J. (1990) Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47, 103–38.

Glasberg, B. R. & Moore, B. C. J. (2000) Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *Journal of the Acoustical Society of America*, 108, 2318–28.

Glasberg, B. R. & Moore, B. C. J. (2002) A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50, 331–42.

Glasberg, B. R., Moore, B. C. J., Patterson, R. D., & Nimmo-Smith, I. (1984) Dynamic range and asymmetry of the auditory filter. *Journal of the Acoustical Society of America*, 76, 419–27.

Gockel, H., Carlyon, R. P., & Plack, C. J. (2005) Dominance region for pitch: Effects of duration and dichotic presentation. *Journal of the Acoustical Society of America*, 117, 1326–36.

Green, D. M. (1973) Temporal acuity as a function of frequency. *Journal of the Acoustical Society of America*, 54, 373–79.

Green, D. M. (1988) *Profile Analysis*. Oxford: Oxford University Press.

Grose, J. H. & Hall, J. W. (1992) Comodulation masking release for speech stimuli. *Journal of the Acoustical Society of America*, 91, 1042–50.

Hall, J. W., Grose, J. H., & Mendoza, L. (1995) Across-channel processes in masking. In B. C. J. Moore (ed.), *Hearing* (pp. 243–66). San Diego: Academic Press.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984) Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, 76, 50–6.

Helmholtz, H. L. F. (1863) *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik* [On the sensations of tone as a physiological basis for the theory of music]. Braunschweig: F. Vieweg.

Hermes, D. J. & Gestel, J. C. van (1991) The frequency scale of speech intonation. *Journal of the Acoustical Society of America*, 90, 97–102.

Hoekstra, A. & Ritsma, R. J. (1977) Perceptive hearing loss and frequency selectivity. In E. F. Evans & J. P. Wilson (eds.), *Psychophysics and Physiology of Hearing* (pp. 263–71). London: Academic Press.

Houtsma, A. J. M. & Smurzynski, J. (1990) Pitch identification and discrimination for complex tones with many harmonics. *Journal of the Acoustical Society of America*, 87, 304–10.

Irino, T. & Patterson, R. D. (2001) A compressive gammachirp auditory filter for both physiological and psychophysical data. *Journal of the Acoustical Society of America*, 109, 2008–22.

Johnson-Davies, D. & Patterson, R. D. (1979) Psychophysical tuning curves: Restricting the listening band to the signal region. *Journal of the Acoustical Society of America*, 65, 765–70.

Kay, R. H. & Mathews, D. R. (1972) On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *Journal of Physiology*, 225, 657–77.

Klatt, D. H. (1973) Discrimination of fundamental frequency contours in speech: Implications for models of pitch perception. *Journal of the Acoustical Society of America*, 53, 8–16.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003) Sensitivity to change in perception of speech. *Speech Communication*, 41, 59–69.

Kluk, K. & Moore, B. C. J. (2004) Factors affecting psychophysical tuning curves for normally hearing subjects. *Hearing Research*, 194, 118–34.

Kohlrausch, A., Fassel, R., & Dau, T. (2000) The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *Journal of the Acoustical Society of America*, 108, 723–34.

Kortekaas, R. W. & Kohlrausch, A. (1999) Psychoacoustical evaluation of PSOLA, II: Double-formant stimuli and the role of vocal perturbation. *Journal of the Acoustical Society of America*, 105, 522–35.

Langner, G. & Schreiner, C. E. (1988) Periodicity coding in the inferior colliculus of the cat, I: Neuronal mechanisms. *Journal of Neurophysiology*, 60, 1799–822.

Leshowitz, B. (1971) Measurement of the two-click threshold. *Journal of the Acoustical Society of America*, 49, 426–66.

Licklider, J. C. & Pollack, I. (1948) Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech. *Journal of the Acoustical Society of America*, 20, 42–52.

Lopez-Poveda, E. A. & Meddis, R. (2001) A human nonlinear cochlear filterbank. *Journal of the Acoustical Society of America*, 110, 3107–18.

Lorenzi, C., Gilbert, G., Carn, C., Garnier, S., & Moore, B. C. J. (2006) Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences USA*, 103, 18866–9 Published online: www.pnas.org/cgi/doi/10.1073/pnas.0607364103.

Moore, B. C. J. (1986) Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. *Scandinavian Audiology*, Supplement 25, 139–52.

Moore, B. C. J. (1992) Across-channel processes in auditory masking. *Journal of the Acoustical Society of Japan (E)*, 13, 25–37.

Moore, B. C. J. (1996) Masking in the human auditory system. In N. Gilchrist & C. Grewin (eds.), *Collected Papers on Digital Audio Bit-Rate Reduction* (pp. 9–19). New York: Audio Engineering Society.

Moore, B. C. J. (2003a) *An Introduction to the Psychology of Hearing*, 5th edn. San Diego: Academic Press.

Moore, B. C. J. (2003b) Speech processing for the hearing-impaired: Successes, failures, and implications for speech mechanisms. *Speech Communication*, 41, 81–91.

Moore, B. C. J. (2003c) Temporal integration and context effects in hearing. *Journal of Phonetics*, 31, 563–74.

Moore, B. C. J., Alcántara, J. I., & Dau, T. (1998) Masking patterns for sinusoidal and narrowband noise maskers. *Journal of the Acoustical Society of America*, 104, 1023–38.

Moore, B. C. J. & Glasberg, B. R. (1983a) Masking patterns of synthetic vowels in simultaneous and forward masking. *Journal of the Acoustical Society of America*, 73, 906–17.

Moore, B. C. J. & Glasberg, B. R. (1983b) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74, 750–3.

Moore, B. C. J. & Glasberg, B. R. (1987) Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns. *Hearing Research*, 28, 209–25.

Moore, B. C. J. & Glasberg, B. R. (1988) Effects of the relative phase of the components on the pitch discrimination of complex tones by subjects with unilateral cochlear impairments. In H. Duifhuis, H. Wit, & J. Horst (eds.), *Basic Issues in Hearing* (pp. 421–30). London: Academic Press.

Moore, B. C. J. & Glasberg, B. R. (1990) Frequency discrimination of complex tones with overlapping and non-overlapping harmonics. *Journal of the Acoustical Society of America*, 87, 2163–77.

Moore, B. C. J., Glasberg, B. R., Flanagan, H. J., & Adams, J. (2006) Frequency discrimination of complex tones: Assessing the role of component resolvability and temporal fine structure. *Journal of the Acoustical Society of America*, 119, 480–90.

Moore, B. C. J., Glasberg, B. R., Gaunt, T., & Child, T. (1991) Across-channel masking of changes in modulation depth for amplitude- and frequency-modulated signals. *Quarterly Journal of Experimental Psychology*, 43A, 327–47.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1985) Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, 77, 1853–60.

Moore, B. C. J., Glasberg, B. R., Plack, C. J., & Biswas, A. K. (1988) The shape of the ear's temporal window. *Journal of the Acoustical Society of America*, 83, 1102–16.

Moore, B. C. J., Glasberg, B. R., & Shailer, M. J. (1984) Frequency and intensity difference limens for harmonics within complex tones. *Journal of the Acoustical Society of America*, 75, 550–61.

Moore, B. C. J. & Jorasz, U. (1992) Detection of changes in modulation depth of a target sound in the presence of other modulated sounds. *Journal of the Acoustical Society of America*, 91, 1051–61.

Moore, B. C. J. & Ohgushi, K. (1993) Audibility of partials in inharmonic complex tones. *Journal of the Acoustical Society of America*, 93, 452–61.

Moore, B. C. J. & Peters, R. W. (1992) Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity. *Journal of the Acoustical Society of America*, 91, 2881–93.

Moore, B. C. J., Peters, R. W., & Glasberg, B. R. (1993) Detection of temporal gaps in sinusoids: Effects of frequency and level. *Journal of the Acoustical Society of America*, 93, 1563–70.

Moore, B. C. J. & Sek, A. (1995) Auditory filtering and the critical bandwidth at low frequencies. In G. A. Manley, G. M. Klump, C. Köppl, H. Fastl, & H. Oeckinghaus (eds.), *Advances in Hearing Research* (pp. 425–36). Singapore: World Scientific.

Moore, B. C. J. & Shailer, M. J. (1992) Modulation discrimination interference and auditory grouping. *Philosophical Transactions of the Royal Society of London, B*, 336, 339–46.

O'Loughlin, B. J. & Moore, B. C. J. (1981a) Improving psychoacoustical tuning curves. *Hearing Research*, 5, 343–6.

O'Loughlin, B. J. & Moore, B. C. J. (1981b) Off-frequency listening: Effects on psychoacoustical tuning curves obtained in simultaneous and forward masking. *Journal of the Acoustical Society of America*, 69, 1119–25.

Oxenham, A. J. & Moore, B. C. J. (1994) Modeling the additivity of nonsimultaneous masking. *Hearing Research*, 80, 105–18.

Patterson, R. D. (1976) Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59, 640–54.

Patterson, R. D. (1987) A pulse ribbon model of monaural phase perception. *Journal of the Acoustical Society of America*, 82, 1560–86.

Patterson, R. D. & Moore, B. C. J. (1986) Auditory filters and excitation patterns as representations of frequency resolution. In B. C. J. Moore (ed.), *Frequency Selectivity in Hearing* (pp. 123–77). London: Academic Press.

Patterson, R. D. & Nimmo-Smith, I. (1980) Off-frequency listening and auditory filter asymmetry. *Journal of the Acoustical Society of America*, 67, 229–45.

Patterson, R. D. & Wightman, F. L. (1976) Residue pitch as a function of component spacing. *Journal of the Acoustical Society of America*, 59, 1450–9.

Pierrehumbert, J. (1979) The perception of fundamental frequency declination. *Journal of the Acoustical Society of America*, 66, 363–8.

Plack, C. J. & Moore, B. C. J. (1990) Temporal window shape as a function of frequency and level. *Journal of the Acoustical Society of America*, 87, 2178–87.

Plomp, R. (1964a) The ear as a frequency analyzer. *Journal of the Acoustical Society of America*, 36, 1628–36.

Plomp, R. (1964b) The rate of decay of auditory sensation. *Journal of the Acoustical Society of America*, 36, 277–82.

Plomp, R. (1967) Pitch of complex tones. *Journal of the Acoustical Society of America*, 41, 1526–33.

Plomp, R. (1970) Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (eds.), *Frequency Analysis and Periodicity Detection in Hearing* (pp. 397–414). Leiden, The Netherlands: Sijthoff.

Plomp, R. (1976) *Aspects of Tone Sensation*. London: Academic Press.

Plomp, R. & Steeneken, H. J. M. (1969) Effect of phase on the timbre of complex tones. *Journal of the Acoustical Society of America*, 46, 409–21.

Pollack, I. (1968) Periodicity discrimination for auditory pulse trains. *Journal of the Acoustical Society of America*, 43, 1113–19.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981) Speech perception without traditional speech cues. *Science*, 212, 947–50.

Ritsma, R. J. (1967) Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America*, 42, 191–8.

Ronken, D. (1970) Monaural detection of a phase difference between clicks. *Journal of the Acoustical Society of America*, 47, 1091–9.

Rosen, S., Baker, R. J., & Darling, A. (1998) Auditory filter nonlinearity at 2 kHz in normal hearing listeners. *Journal of the Acoustical Society of America*, 103, 2539–50.

Rosen, S. & Fourcin, A. (1986) Frequency selectivity and the perception of speech. In B. C. J. Moore (ed.), *Frequency Selectivity in Hearing* (pp. 373–487). London: Academic Press.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., & Robles, L. (1997) Basilar-membrane responses to tones at the base of the chinchilla cochlea. *Journal of the Acoustical Society of America*, 101, 2151–63.

Saberi, K. & Perrott, D. R. (1999) Cognitive restoration of reversed speech. *Nature*, 398, 760.

Sachs, M. B. & Kiang, N. Y. S. (1968) Two-tone inhibition in auditory nerve fibers. *Journal of the Acoustical Society of America*, 43, 1120–8.

Schouten, J. F. (1968) The perception of timbre. *6th International Conference on Acoustics*, 1, GP-6-2.

Schouten, J. F. (1970) The residue revisited. In R. Plomp & G. F. Smoorenburg (eds.), *Frequency Analysis and Periodicity Detection in Hearing* (pp. 41–54). Leiden, The Netherlands: Sijthoff.

Schreiner, C. E. & Urbas, J. V. (1986) Representation of amplitude modulation in the auditory cortex of the cat, I: The anterior auditory field (AAF). *Hearing Research*, 21, 227–41.

Shackleton, T. M. & Carlyon, R. P. (1994) The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *Journal of the Acoustical Society of America*, 95, 3529–40.

Shailer, M. J. & Moore, B. C. J. (1983) Gap detection as a function of frequency, bandwidth and level. *Journal of the Acoustical Society of America*, 74, 467–73.

Shailer, M. J. & Moore, B. C. J. (1987) Gap detection and the auditory filter: Phase effects using sinusoidal stimuli. *Journal of the Acoustical Society of America*, 81, 1110–17.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science*, 270, 303–4.

Shaw, E. A. G. (1974) Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *Journal of the Acoustical Society of America*, 56, 1848–61.

Steeneken, H. J. M. & Houtgast, T. (1980) A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, 69, 318–26.

Stone, M. A. & Moore, B. C. J. (2004) Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task. *Journal of the Acoustical Society of America*, 116, 2311–23.

Summerfield, A. Q. & Assmann, P. (1987) Auditory enhancement in speech perception. In M. E. H. Schouten (ed.), *The Psychophysics of Speech Perception* (pp. 140–50). Dordrecht: Martinus Nijhoff.

Summerfield, A. Q., Sidwell, A. S., & Nelson, T. (1987) Auditory enhancement of changes in spectral amplitude. *Journal of the Acoustical Society of America*, 81, 700–8.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., & Meddis, R. (2002) A revised model of the inner-hair cell and auditory-nerve complex. *Journal of the Acoustical Society of America*, 111, 2178–88.

't Hart, J. (1981) Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*, 69, 811–21.

Unoki, M., Irino, T., Glasberg, B. R., Moore, B. C. J., & Patterson, R. D. (2006) Comparison of the roex and gammachirp filters as representations of the auditory filter. *Journal of the Acoustical Society of America*, 120, 1474–92.

Verhey, J. L., Dau, T., & Kollmeier, B. (1999) Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model. *Journal of the Acoustical Society of America*, 106, 2733–45.

Viemeister, N. F. (1979) Temporal modulation transfer functions based on modulation thresholds. *Journal of the Acoustical Society of America*, 66, 1364–80.

Vogten, L. L. (1978) Low-level pure-tone masking: A comparison of "tuning curves" obtained with simultaneous and forward masking. *Journal of the Acoustical Society of America*, 63, 1520–7.

Watkins, A. J. (1991) Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, 90, 2942–55.

Watkins, A. J. & Makin, S. J. (1996a) Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, 99, 3749–57.

Watkins, A. J. & Makin, S. J. (1996b) Some effects of filtered contexts on the perception of vowels and fricatives. *Journal of the Acoustical Society of America*, 99, 588–94.

Yost, W. A. & Sheft, S. (1989) Across-critical-band processing of amplitude-modulated tones. *Journal of the Acoustical Society of America*, 85, 848–57.

Yost, W. A., Sheft, S., & Opie, J. (1989) Modulation interference in detection and discrimination of amplitude modulation. *Journal of the Acoustical Society of America*, 86, 2138–47.

Yumoto, E., Gould, W. J., & Baer, T. (1982) Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, 71, 1544–9.

Zhang, X., Heinz, M. G., Bruce, I. C., & Carney, L. H. (2001) A phenomenological model for the responses of auditory-nerve fibers, I: Nonlinear tuning with compression and suppression. *Journal of the Acoustical Society of America*, 109, 648–70.

Zwicker, E. (1964) "Negative afterimage" in hearing. *Journal of the Acoustical Society of America*, 36, 2413–15.

Zwicker, E. & Fastl, H. (1999). *Psychoacoustics: Facts and Models*, 2nd edn. Berlin: Springer.

Zwicker, E. & Terhardt, E. (1980) Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68, 1523–5.

14 Cognitive Processes in Speech Perception

JAMES M. MCQUEEN AND
ANNE CUTLER

1 Introduction

The recognition of spoken language involves the extraction of acoustic-phonetic information from the speech signal, and the mapping of this information onto cognitive representations. To develop accurate psycholinguistic models of this process, we need to know what information is extracted from the signal, and when and how it is integrated with stored knowledge.

The central knowledge store for speech perception is the mental lexicon, that is, our stored representations of words. The utterances that we hear may be new to us, but they are made up of known words; by recognizing the words and parsing their sequence we are able to understand what has been said. Word recognition, we argue, is therefore at the heart of speech perception.

The cognitive processes we will discuss, therefore, concern the relationship between lexical processing (the recognition of words) and prelexical processing (getting from the acoustic-phonetic input to the representations of words). Models of spoken-word recognition have been distinguished principally by their characterization of this relationship, in particular their proposals regarding the directionality of flow of information between the prelexical and lexical levels. The role of the lexicon in prelexical speech processing is discussed in section 2.

In every language, the vocabulary contains tens or hundreds of thousands of words. Since all these words are made up from just a few phonemes (across languages, the average phoneme repertoire size is around 30; Maddieson, 1984), it is necessarily the case that words resemble one another. Successful speech recognition thus entails discriminating the phonemic contrasts that distinguish a word from the other words that resemble it – for instance, deciding that we have heard *word*, and not *bird*, *ward*, or *work*. The processing of segmental information in word recognition is discussed in section 3.

One of the most salient and important facts about speech perception is our ability to understand speech from talkers we have never heard before, and to perceive the same phoneme despite acoustically very different realizations (e.g., by a child's voice versus an adult male's). This ability has been characterized as a process of "normalization," in which an underlying commonality is extracted from the surface variation. The relevant research is discussed in section 3.2 (and see also Johnson, 2005, for a review of models of the normalization process). The role played by abstract underlying representations of the kind assumed in such accounts has, however, been called into question in recent years because of accumulated evidence both that individual speech perception episodes contribute to the knowledge that is stored about speech, and that they can influence later processing (see, e.g., Johnson & Mullennix, 1997). Speech perception theory is currently in an exciting state of transition to a new generation of models in which a role for abstract representations is combined with a role for veridical representations of speech episodes or exemplars. The evidence which shows that both types of representation are involved in speech perception is also summarized in section 3.2.

An account of the cognitive process of speech perception would, finally, be incomplete if it considered only segmental and lexical information, for the suprasegmental dimensions in which the speech signal varies also contribute significantly to listeners' processing decisions. This is documented in section 4.

The recognition of speech is one of humankind's most useful and significant achievements; underlying it is cognitive processing of enormous complexity but also admirable efficiency. Modeling this process has occupied psycholinguists and speech scientists for decades, and, as our concluding summary in section 5 demonstrates, the search for the ultimately accurate model is not over yet.

2 Lexical Information

We examine the role of lexical information in speech processing by contrasting interactive models with autonomous models. Interactive models hold that lexical information influences prelexical processing. We will focus on one particular model of this class, TRACE (McClelland & Elman, 1986; McClelland, 1991; see left panel of Figure 14.1). This interactive-activation model has three levels of processing containing, respectively, featural representations, phonemes, and words. Units within a level compete with each other via lateral inhibitory connections. Units at lower levels activate the units at higher levels with which they are consistent via facilitatory connections. Thus, during word recognition, activation of a feature node leads to activation of consistent phoneme nodes, which in turn activate word nodes. Importantly, higher-level units also facilitate lower-level units. Activated word units boost the activation of their constituent phonemes: this top-down facilitation instantiates the claim that lexical information influences prelexical processing.

We will contrast the TRACE model with an autonomous model which holds that lexical information is not involved in prelexical processing. The Merge model (Norris et al., 2000; see right panel of Figure 14.1) also has three levels of processing: a prelexical level, a lexical level, and a level at which explicit decisions are made about speech sounds. Units within each of the latter two levels inhibit each other. Prelexical units facilitate lexical and decision-level units with which they

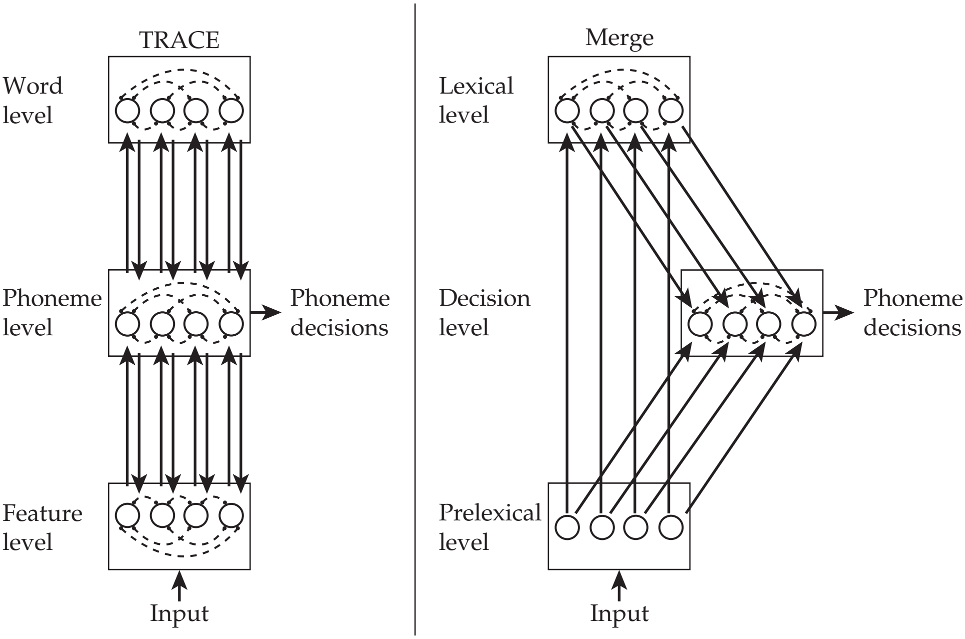


Figure 14.1 Sketches of the processing architecture of, on the left, the interactive TRACE model (McClelland & Elman, 1986), and, on the right, the autonomous Merge model (Norris et al., 2000). Excitatory connections between nodes at different levels are shown with solid lines and arrows; bidirectional inhibitory connections between nodes within levels are shown with dashed lines and closed circles. Not all connections are shown.

are consistent, and lexical units also facilitate decision-level units. There are only these feedforward connections in Merge. Critically, there is no feedback from the lexical to the prelexical level, so the lexicon cannot influence prelexical processing. Merge is linked to the Shortlist model of spoken-word recognition (Norris, 1994; Norris & McQueen, 2008), which is also based on the assumption that there is no lexical feedback. The prelexical and lexical levels in Merge (i.e., those involved in word recognition) are interchangeable with those two levels in Shortlist, while the decision units (and the feedforward connections to them) are not part of the word-recognition process: they are used only when listeners have to make explicit phonetic decisions. Below, we will describe how the TRACE and Merge models, as instances of interactive and autonomous theories, account for lexical involvement in various phonetic tasks.

2.1 Lexical effects

2.1.1 Monitoring Phoneme monitoring is sensitive to phonetic factors. Foss and Gernsbacher (1983) found an effect of vowel length: the longer the vowel, the longer the reaction time (RT) to the preceding target consonant. Another factor is the phonological similarity of the target phonemes to preceding phonemes (Newman & Dell, 1978; Dell & Newman, 1980). Detection of target phonemes in sentences is slower when the word preceding the target-bearing word begins with a phoneme closely related to the target. Several studies, however, have failed to find lexical effects. Foss et al. (1980) found that monitoring was no faster for words than for nonwords, and that the frequency of occurrence of the target-bearing word did not influence RT. Segui et al. (1981) also failed to find an RT advantage for word responses over nonword responses, and Segui and Frauenfelder (1986) found no frequency effect when subjects were required to monitor only for word-initial phonemes ("standard" phoneme monitoring).

These results support the claim that phoneme monitoring is based on prelexical processing which is open to the influence of phonetic information but not lexical information. But there are some studies which have demonstrated lexical effects. Segui and Frauenfelder (1986), for instance, did obtain a word-frequency effect when subjects were required to monitor not just for word-initial targets, but for targets which could appear anywhere in the words ("generalized" phoneme monitoring). Rubin et al. (1976) also found a word/nonword effect: subjects were faster to detect e.g. /b/ in *bat* than in *bal*. Cutler et al. (1987) examined word/nonword effects in a series of experiments. Lexical effects were found to come and go. Responses to targets in words were faster than those to targets in nonwords only when task monotony was reduced.

Lexical effects thus appear to be present only in some phoneme-monitoring experiments. Stemberger et al. (1985) took this variability as support for interactive models like TRACE. Lexical influences were taken to result from top-down facilitation from word nodes increasing the level of activation of target phoneme nodes, thus speeding responses to targets in words relative to nonwords. Where there were no lexical effects, it was assumed that responses were being made from the phoneme-node level, with lexical feedback switched off through some kind of attentional process. But the presence of lexical effects, and their variability, can equally well be explained by autonomous models (Cutler et al., 1987; Norris et al., 2000). In Merge, lexical effects in phoneme monitoring are due to the feed-forward influence of the lexical level on decision nodes, and their absence, just as in the TRACE account, is assumed to reflect the fact that the lexical influence, due for example to attentional factors, has been switched off.

Both models can therefore account for the lexical effect, and its variability, in phoneme monitoring. In another task, rhyme monitoring, where subjects detect words and nonwords which rhyme with a prespecified cue, responses are faster to words than to nonwords, and responses are faster to high- than to low-frequency rhyming words (McQueen, 1993). Again both models can explain these lexical effects.

2.1.2 Phonemic restoration If the medial /s/ of the word *legislatures* is replaced with a cough, listeners report hearing a cough and the complete *legislatures*, with the absent phoneme perceptually restored (Warren, 1970). Low-level factors influence the effect: if the replacing noise is acoustically similar to the removed phoneme, the illusion is more likely to occur (Warren & Obusek, 1971; Samuel, 1981a, 1981b); and there is more restoration for fricatives and stops (which are more noise-like) than for liquids, vowels and nasals (Samuel, 1981a, 1981b).

Samuel (1981a) found that several lexical factors influenced the extent of the illusion: there was more restoration for longer than for shorter words; there was a more reliable illusion in words than in phonologically legal nonwords; and presenting an intact version of the target word before the target word also increased restoration. Samuel (1987) found further that there was more restoration for items with several possible restorations (e.g., **egion*: *legion* or *region*) than for items with a unique restoration (e.g., **esion*: *lesion*). He also found that there was more phonemic restoration in words which become unique early, moving left to right through the word (e.g., *boysenb*rry*) than in words which became unique late (e.g., *indel*ble*). Samuel (1996) showed that, as with lexical involvement in phoneme monitoring, lexical effects in phonemic restoration are variable; to use his words, they are "real but fragile." Samuel explained these results in terms of lexical feedback.

An autonomous account of these data, however, is once again also possible. If the illusion is due to attention being focused on lexical information, then the lexical effects can be explained without recourse to top-down connections. In Merge's terms, lexical influences in restoration occur when listeners are using the connections from the lexical level to the decision level. Just as with the monitoring tasks, the evidence for lexical involvement in phoneme restoration reviewed so far does not allow us to distinguish between the two models.

2.1.3 Phonetic categorization In the phonetic categorization task, with a continuum of sounds from /d/ to /t/ in the contexts *deep–teep* and *deach–teach*, for example, a lexical effect would be shown by an increased proportion of /d/ responses in the ambiguous region of the continuum when the voiced endpoint formed a word (*deep*), and an increased proportion of /t/ responses when the unvoiced endpoint formed a word (*teach*). This effect was originally demonstrated by Ganong (1980), and replicated by Fox (1984). In TRACE, this effect is once again accounted for by top-down connections. In Merge, the effect once again reflects the integration of prelexical and lexical information at the decision level.

Connine and Clifton (1987) found both a lexical shift and an RT advantage for word responses relative to nonword responses in the boundary region. They further showed that the lexical effect was not due to postperceptual bias: it was not equivalent to an effect obtained using monetary reward to bias subjects' responses. Lexical effects have also been reported by Burton et al. (1989), who found that the categorization of a word-initial continuum depended on the acoustic-phonetic quality of the continuum, and by Miller and Dexter (1988), who showed that lexical involvement in categorization (as in phoneme monitoring and the phonemic restoration illusion) is not mandatory.

McQueen (1991) and Pitt and Samuel (1993) have found lexical effects for phonemes in word-final position (e.g., for an /ʃ/–/s/ continuum in contexts such as *fish–fiss* and *kish–kiss*). McQueen (1991) also replicated Burton et al.'s (1989) finding that lexical effects in the categorization task only appear when the materials are of poor acoustic quality. Pitt and Samuel (1993), however, have shown that poor stimulus quality is not a necessary condition for a lexical effect: lexical shifts were obtained with high-quality materials in both word-initial and word-final categorization. Critically, however, the basic lexical effect in this task is consistent with both types of model.

2.2 Test cases

Both models can account for lexical effects in several tasks. Are there any test cases which might allow us to distinguish between the models? Can we establish whether or not lexical information is used in prelexical processing? Several attempts have been made to contrast divergent predictions of the TRACE and Merge models.

2.2.1 Inhibitory lexical effects in phoneme monitoring Frauenfelder et al. (1990) presented evidence from the phoneme-monitoring task which challenges the interactive position. TRACE predicts that activation of a lexical candidate will both boost the activation of its constituent phonemes by top-down facilitation and inhibit the activation of nonconstituent phonemes because of phoneme-to-phoneme inhibition. As this study showed, there are strong facilitatory effects on the detection of targets (such as /p/ in *olympiade*), which occur after the word becomes unique, relative to matched nonwords (e.g., *arimpiako*). In TRACE terms, this could be due to top-down facilitation of /p/ from the word node. If this were the case, detection of /t/ in *vocabutaire* should be inhibited relative to detection of /t/ in a matched nonword such as *socabutaire*, because of top-down facilitation of /l/ from the activated *vocabulaire* node followed by inhibition of other phoneme nodes by the /l/ node. No such inhibition was found.

Mirman et al. (2005) have recently shown, however, that lexically induced delays in phoneme monitoring do occur, but only if the target phoneme and the lexically consistent phoneme are phonetically similar (e.g., /t/ detection in *arsenit* was delayed because /t/ is similar to the lexically consistent /k/ in the base word *arsenic*, but /t/ detection in *abolit* was not delayed, presumably due to greater dissimilarity between /t/ and the /ʃ/ of *abolish*). Mirman et al. present TRACE simulations showing that this kind of lexical inhibition only arises in the model if there is some bottom-up support for the lexically consistent sound (i.e., when it is acoustically similar to the target). As Norris et al. (2000) argue, Merge predicts that there should be lexical inhibition in phoneme monitoring when there is perceptual support for two competing phonemes. Thus, while the absence of lexical inhibition in phoneme monitoring was for many years a problem for the interactive account, more recent research has shown that this is not in fact a test case that distinguishes between interactive and autonomous models.

2.2.2 The time course of lexical effects in categorization McQueen (1991) showed that the lexical effect in categorization of word-final ambiguous fricatives, in contexts such as *fish–fiss* and *kish–kiss*, was larger for faster responses. This finding was replicated by Pitt and Samuel (1993) and by McQueen et al. (2003). Previous research (Fox, 1984; Miller & Dexter, 1988; Pitt & Samuel, 1993) had shown that lexical effects in categorization of ambiguous word-initial sounds (e.g., /d/ and /t/ in *deep–teep* vs. *deach–teach*) build up over time, but McQueen et al. (2003) also showed that, at least under some circumstances, lexical effects in word-initial categorization can also decrease over time. In TRACE, the lexical feedback loop becomes stronger over time, and bottom-up evidence thus tends to be overwritten by lexical knowledge. TRACE thus wrongly predicts that lexical effects should build up gradually over time, for word-initial (McClelland & Elman, 1986) and word-final categorization (McClelland, 1987). In Merge, in contrast, there is no feedback loop, and lexical knowledge does not overwrite bottom-up evidence. So lexical effects can reach an asymptote, and indeed die away over time (if lexical input to the decision nodes is switched off, the bottom-up evidence, as represented at the prelexical level, can re-assert itself at the decision level). These time-course analyses thus favor autonomous accounts.

2.2.3 Compensation for coarticulation One type of result appears to support interactive models. Mann and Repp (1981) showed that stops midway between /t/ and /k/ were more often categorized as /k/ after /s/, but as /t/ after /ʃ/. The perceptual system thus appears to compensate for fricative–stop coarticulation. Elman and McClelland (1988) replicated this effect for ambiguous word-initial stops following fricative-final words such as *Christmas* and *foolish*, and, most importantly, they showed that the effect occurred when the word-final fricatives were replaced with an ambiguous fricative. When the final /s/ in *Christmas* was replaced with an ambiguous sound /?/, midway between /s/ and /ʃ/, there were again more /k/ responses to the ambiguous stops. With *fooli?*, there were more /t/ responses.

Elman and McClelland (1988) claimed that this effect was strong evidence in favor of interactive models like TRACE. Lexical information appears to be influencing a compensation process that can be assumed to be operating prelexically. This seems to be direct evidence against the autonomous assumption that there is no lexical feedback. But Pitt and McQueen (1998) argued that there was an alternative account of these data. In English, /s/ is more likely than /ʃ/ after schwa (as in *Christmas*), and /ʃ/ is more likely than /s/ after /ɪ/ (as in *foolish*). Pitt and McQueen then showed, first, that if vowel-fricative transitional probabilities were controlled, there was no lexical effect with ambiguous fricatives in stop categorization, and second, that if vowel–fricative transitional probabilities were manipulated in nonwords, those probability biases on ambiguous fricative identification did lead to a consequent shift in stop categorization. Pitt and McQueen thus argued that if the prelexical level were sensitive to transitional probabilities, the Elman and McClelland results would be consistent with an autonomous model. Studies in the next round of this debate have claimed to show that lexical influences in fricative–stop compensation for coarticulation can be found when vowel–fricative transitional probabilities are controlled (Magnuson et al., 2003; Samuel & Pitt, 2003). The Magnuson et al. result, however, appears to be due to a bias induced by their practice trials (McQueen et al., 2009), and Samuel and Pitt's findings may reflect longer-range transitional probabilities than those concerning the vowel–fricative sequence alone (McQueen, 2003). This issue is not yet resolved (McClelland et al., 2006; McQueen, Norris, et al., 2006), and appears to depend on increasingly subtle experimental manipulations. There has to date been no completely convincing demonstration of lexical involvement in compensation for coarticulation.

Another aspect of Pitt and McQueen's (1998) data, replicated by McQueen et al. (2009), is problematic for interactive models. Listeners were asked to identify the fricatives (as /s/ or /ʃ/) as well as the stops (as /t/ or /k/). A lexical effect was found in the fricative judgments (just as in the Ganong, 1980, and McQueen, 1991, studies, listeners judged more ambiguous sounds to be lexically consistent than to be lexically inconsistent). Yet in the very same trials there was no lexical effect in the stop judgments. If the lexical effect on the fricatives were due to feedback, as in the TRACE account, then that feedback ought to have had an effect on the prelexical compensation for coarticulation mechanism, and there thus ought also to have been a lexical effect on the stops. This dissociation challenges TRACE. It is consistent with Merge, however, since lexical effects on fricative decisions reflect the influence of the lexicon on the decision level and thus not on the prelexical level where the compensation mechanism is assumed to operate.

2.2.4 Selective adaptation Samuel (1997, 2001) used a logic similar to Elman and McClelland (1988), and again tested for lexical influences on a prelexical process (selective adaptation rather than compensation for coarticulation). In selective adaptation, judgments about speech sounds change through repeated exposure to one sound (e.g., after hearing /da/ repeatedly, listeners report more stimuli on a /da–ta/ continuum to be /ta/; Eimas & Corbit, 1973). The locus of this adaptation effect appears to be prelexical (Samuel & Kat, 1996). Samuel (1997) used sounds replaced with noise as adaptors (capitalizing on the phoneme restoration illusion), and Samuel (2001) used ambiguous sounds as adaptors (capitalizing on the Ganong, 1980, effect). In both cases these ambiguous adaptors appeared in lexical contexts. Adaptation effects were found which were similar to those that would be observed with unambiguous lexically consistent sounds. In line with interactive models such as TRACE, lexical knowledge thus appeared to be influencing the prelexical adaptation process. As we discuss below, however, these results appear to be consistent with the autonomous view that the lexicon does not influence on-line prelexical processing.

2.3 *On-line feedback versus feedback for learning*

As our review of these test cases reveals, there is no clear winning theory. Some experiments favor the interactive view embodied in TRACE, some favor the autonomous view in Merge, and some test cases have proven not to distinguish between the models. Norris et al. (2000) argued, however, that even if the data are not definitive, there are important theoretical arguments to consider. They pointed out that feedback, as instantiated in TRACE, cannot be of any benefit to word recognition, and can be harmful to phoneme recognition. The best a word-recognition system can do is recognize the words that are most consistent with the input. Thus, if processing is optimal, feedback cannot improve on the decisions made at the lexical level (all it does is copy the decisions made at the lexical level onto the prelexical level). Feedback can help with phoneme recognition (e.g., when a sound is ambiguous), but can potentially create phonemic hallucinations, where lexical knowledge overwrites perceptual evidence. Norris et al. thus argued that, in the absence of clear experimental data in support of interactive models, autonomous models such as Merge should be preferred simply because there is no good reason to postulate feedback connections.

There is one critical exception to this argument. Feedback for perceptual learning would be of benefit in speech perception. If listeners could adjust their prelexical representations over time, using lexical knowledge, then they could learn how to interpret a talker speaking in an unusual way (e.g., a talker with a speech impediment, or someone with a regional or foreign accent). If, for example, a talker with a lisp produces an ambiguous /s/ sound in words where an /s/ (and not an /f/) is expected (e.g., at the end of *platypu-*), then listeners could use this lexical knowledge to adjust their category boundary between /f/ and /s/, facilitating subsequent recognition of speech by that talker. Norris et al. (2003) found support for this prediction in a Dutch perceptual-learning experiment. After exposure to only 20 /s/-final words ending with an ambiguous /fs/ sound (e.g., *radij?*, based on *radijs*, 'radish'), listeners' category boundaries shifted to include more /f/-like sounds in the /s/ category. Another group of listeners heard the same sound in 20 /f/-final lexical contexts (e.g., *olij?*, based on *olijf*, 'olive'), and learned to include more sounds in their /f/ category. Control conditions showed that lexical knowledge was required for this kind of perceptual learning. Subsequent research with this lexical retuning paradigm has shown that the learning can be, but need not be, talker-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006, 2007), and that it is stable over at least 12 hours (Eisner & McQueen, 2006).

Norris et al. (2003) argued, therefore, that there is lexical feedback in perceptual learning. The prelexical level appears to be flexible enough to be able to adjust its operation over time, using stored lexical knowledge, as it encounters different talkers. This kind of feedback is logically distinct from the kind of feedback instantiated in TRACE, where lexical knowledge modulates, on-line, the perceptual process. Thus, while the feedback mechanism in TRACE may offer an explanation for both on-line and learning effects (McClelland et al., 2006), there is no necessity that both types of effect be explained by the same mechanism. A perceptual learning mechanism using lexical feedback could therefore be added to Merge, without there being any effects of that feedback loop on on-line perception (Norris et al., 2003). Furthermore, this kind of model could explain the Samuel (1997, 2001) results. The selective adaptation paradigm requires repeated exposure to stimuli, and, as Norris et al. argued, may thus involve the same kind of perceptual retuning as they observed (see also McQueen, Norris, et al., 2006; Vroomen et al., 2007).

In our chapter in the first edition of this book we concluded that the lexicon does not influence prelexical processing. This conclusion has stood the test of time, at least with respect to on-line processing. It now appears, however, that there is lexical involvement in prelexical perceptual learning. Perhaps most importantly, while on-line feedback is of no benefit to word recognition, retuning of perception over time does benefit speech perception. There is thus only feedback of the type that helps listeners.

3 Segmental Information

We now consider two key questions about the role of segmental information in the cognitive processes underlying speech perception. First, we ask whether, during word recognition, segmental information is processed serially, or in a cascaded fashion. Our focus is on word recognition, since, as we have already argued, it is only through recognizing words that the listener can understand spoken messages. It is entirely uncontroversial that segmental information must play a central role in spoken-word recognition. The primary way in which listeners determine whether they have heard *word* or *bird* is through identifying that the first sound is a /w/ and not a /b/. It is also uncontroversial that a wide variety of acoustic cues are used in segment perception (see Raphael, 2005). The question we must ask, then, is how these cues modulate word recognition. One possibility is that they do so in a serial manner: First, segments are identified on the basis of these cues (at the prelexical stage of processing), and then, second, words are recognized. Alternatively, acoustic-phonetic information may flow in cascade through the recognition system, such that it influences lexical-level processes without any prior definitive categorization of the input into segmental categories. We consider these two alternatives in section 3.1.

But there is an even more fundamental question about the uptake of segmental information. So far, we have assumed that there is a prelexical stage of processing, where the speech input is recoded (in either a serial or cascaded fashion) into linguistically abstract segmental categories prior to and for lexical access. But there may be no such prelexical abstraction process. We consider arguments for and against abstraction in section 3.2.

3.1 Cascaded processing of segmental information

Multiple lexical hypotheses are activated during the word-recognition process (Swinney, 1981; Marslen-Wilson, 1987, 1990; Zwitserlood, 1989; Shillcock, 1990; Gow & Gordon, 1995; Tabossi et al., 1995; Vroomen & de Gelder, 1997; Allopenna et al., 1998). We can thus ask whether segmental information is passed serially or in cascade to the lexical level, by measuring whether lexical activation changes as a function of subsegmental differences in the input. According to a serial model, subsegmental ambiguities should, if possible, be resolved prelexically and thus should not affect lexical activation levels. But in a cascaded model subsegmental differences should be passed on to the lexical level and influence the degree of activation of different lexical hypotheses.

Andruski et al. (1994) reduced the Voice Onset Time (VOT) of the initial stop of, for example, *king*. In English, VOT is a major acoustic-phonetic cue to the distinction between voiceless stops (e.g., /k/, with long VOTs) and voiced stops (e.g., /g/, with short VOTs). In a cross-modal priming task, responses to semantically related targets (e.g., *queen*) were faster after *king* than after an unrelated word. This priming effect became smaller as VOT was reduced (i.e., as the /k/ became more like a /g/). This suggests that subsegmental detail influences lexical activation (the degree of activation of *king* was reduced as the /k/ was shortened), in keeping with the cascaded model. Further evidence is provided by McMurray et al. (2002) and Utman et al. (2000).

Other evidence in favor of cascaded models comes from research examining the effects of mispronunciations. The phonetic similarity between an intended word (e.g., *cabinet*; Connine et al., 1997) and a mispronounced nonword (e.g., *gabinet* vs. *mabinet* vs. *shuffinet*) influences how disruptive the mispronunciation is to lexical access. The greater the similarity between the mismatching sound and the intended sound, the more strongly the intended word appears to be activated (see also Connine et al., 1993; Marslen-Wilson et al., 1996; Ernestus & Mak, 2004). In the domain of research on continuous speech processes (such as place assimilation in English, Gow, 2002; liaison in French, Spinelli et al., 2003; and /t/ reduction in Dutch, Mitterer & Ernestus, 2006) it appears that fine-grained phonetic detail (e.g., the duration or spectral structure of segments) modulates the degree of lexical activation, again as predicted by the cascaded account.

A considerable body of evidence suggests further that, once words consistent with the input speech have been activated, they compete with each other (Goldinger et al., 1989, 1992; Cluff & Luce, 1990; Slowiaczek & Hamburger, 1992; McQueen et al., 1994; Norris et al., 1995; Vroomen & de Gelder, 1995; Vitevitch & Luce, 1998, 1999; Luce & Large, 2001; Gaskell & Marslen-Wilson, 2002). Accessed words compete with each other until one word dominates the others; this one word can then be recognized. This competition process is instantiated, in different ways, in current models of spoken-word recognition: the Neighborhood Activation Model (Luce & Pisoni, 1998), TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994; Norris & McQueen, 2008) and the Distributed Cohort Model (DCM; Gaskell & Marslen-Wilson, 1997).

The lexical competition process provides a litmus test about how segmental information flows through the speech recognition system. If fine-grained phonetic detail modulates the competition process, then it must have been passed forward to the lexical level. Several of the effects just reviewed have been shown to interact with lexical factors. Van Alphen and McQueen (2006) have shown, for example, that the influence of VOT variability on lexical activation in Dutch depends on the lexical competitor environment (i.e., whether the voiced and voiceless interpretations of the stops are both words, as in, e.g., the English pair *bear–pear*, or both nonwords, as in English *blem–plem*, or one word and one nonword, as in *blue–plue* and *brince–prince*), and Marslen-Wilson et al. (1996) also found that segmental mismatch effects were modulated by lexical factors.

Clear demonstrations of this interaction between subsegmental and lexical information come from a series of experiments with cross-spliced stimuli (Streeter & Nigro, 1979; Whalen, 1984, 1991; Marslen-Wilson & Warren, 1994; McQueen et al., 1999; Dahan et al., 2001). Cross-splicing the initial consonant and vowel of *jog* with the final consonant of *job*, for example, produces a stimulus which sounds like *job*, but which contains (in the vocalic portion) acoustic evidence for a /g/. These phonetically mismatching stimuli are more difficult to process, but the extent to which they interfere depends on whether the entire sequence is a word or nonword (e.g., *job* vs. *shob*) and on whether its components derive from words (e.g., *jog*) or nonwords (e.g., *jod*). It thus seems very clear that there is cascade of segmental information to the lexical level.

3.2 Segmental abstraction in speech perception

There are two ways in which segmental information might cascade to the lexicon. One possibility is that phonetic segments are extracted explicitly, in some prelexical level of representation, with a classification of the speech signal into linguistically abstract "units of perception" (McNeill & Lindig, 1973; Healy & Cutting, 1976). Units which have been postulated include acoustic-phonetic features (Eimas & Corbit, 1973; Marslen-Wilson, 1987; Marslen-Wilson & Warren, 1994; Stevens, 2002), phonemes (Foss & Blank, 1980; McClelland & Elman, 1986; Norris, 1994; Norris et al., 2000), context-sensitive allophones (Wickelgren, 1969), syllables (Cole & Scott, 1974; Mehler, 1981), and articulatory gestures (Liberman & Mattingly, 1985). Any of these units could operate in cascade, passing information continuously forward to the lexicon. Alternatively, segmental information could be used implicitly, in a direct and continuous mapping of the signal onto the lexicon with no explicit intermediate classification into prelexical units. Klatt (1979, 1989) suggested a template-matching process, where spectral information, as analyzed by the peripheral auditory system, is mapped directly onto a lexicon of spectral templates of diphone sequences. The models of Goldinger (1996, 1998), Johnson (1997), Pierrehumbert (2002), and Hawkins (2003), though differing in many respects, share the assumption that the lexicon consists of episodic memory traces of particular tokens of words, stored with all their acoustic detail (e.g., including talker- and situation-specific details).

The class of models with abstract prelexical representations provide a ready solution to the invariance problem. It is well known that the acoustic cues to segments are far from invariant. They vary greatly depending on a large number of factors, including: coarticulation (the realization of segments depends upon both preceding and following phonological context; Fowler, 1984; see Farnetani & Recasens, this volume); speech rate (e.g., temporal cues such as VOT change depending on speed of articulation, requiring rate-dependent processing; Miller, 1981; Gordon, 1988); and variation between speakers due to differences in sex, age, and dialect (see Ní Chasaide & Gobl, this volume). Some authors have argued that this variation is dealt with by the extraction of acoustic cues which are invariant (Stevens & Blumstein, 1981); others that the variation is lawful, and can be exploited by the listener (Elman & McClelland, 1986). In either case, however, it is clear that the perceptual system must be able to deal with this variability through some kind of normalization process. The same physical signal must be interpretable as different segments, and different signals must be interpretable as the same segment (Repp & Liberman, 1987). If normalization takes place prelexically, prior to lexical access, then only abstract phonological knowledge would need to be stored in the mental lexicon.

In a study of speech rate normalization, for example, Miller et al. (1984) demonstrated that subjects labeled more ambiguous consonants, midway between /b/ and /p/ and embedded in the continuum *bath–path*, in a contextually congruent manner (i.e., more *bath* responses in a bathing context), but only when subjects were explicitly told to attend to the sentence context. These effects were absent in a speeded response condition, which focused the subjects' attention on the target words. Speaking rate was also varied, resulting in shifts in the category boundary between /b/ and /p/, but the task-demand manipulation did not influence this rate-dependent boundary placement. Miller and Dexter (1988) also used the phonetic categorization task to examine effects of lexical status and speaking rate. They found that under speeded response conditions, there was no tendency to label ambiguous initial consonants in a lexically consistent manner (e.g., as /b/ in a *beef–peef* continuum and as /p/ in a *beece–peace* continuum). Listeners could not ignore the rate manipulation, however: even under speeded-response instructions they based their decision on the early portion of the syllable, treating it as if it was physically short (the /b/–/p/ boundary shifted to a smaller VOT for fast responses). These studies neatly demonstrate that rate normalization (unlike the use of lexical knowledge) is a mandatory feature of speech processing. The analysis of acoustic information specifying speech rate appears to be essential for accurate lexical access (Miller, 1987). Such results thus support the view that rate normalization is a prelexical process that necessarily modulates word recognition.

Other, more recent evidence on the need for prelexical abstraction comes from the lexical retuning paradigm reviewed earlier. Norris et al. (2003) argued that adjustments to prelexical phonetic categories would be of benefit to speech perception because, once an adjustment had been made, it could be applied to the recognition of any words containing the adjusted sound. McQueen, Cutler, et al. (2006) tested whether there was indeed generalization of learning to the processing of words that were spoken by the trained talker but that had not been heard during the training phase. Instead of using the phonetic categorization test task (as in, e.g., the Norris et al. study described in section 2.3), they used a cross-modal identity-priming task. The experiment was again in Dutch, and the test phase contained minimal pairs such as *doof–doos*, 'deaf–box'. In a training phase identical to the Norris et al. study (i.e., with no minimal-pair words), listeners were encouraged to learn that an ambiguous /fs/ sound, "[?]", was either /f/ or (for a second group) /s/. In the subsequent test phase, the pattern of priming effects indicated that the listeners in the first group tended to hear [do?] as *doof*, while listeners in the second group tended to hear it as *doos*. This demonstration of generalization of learning to previously unheard words confirms that the locus of the learning effect is prelexical. It also underlines the major benefit of prelexical abstraction: Once something about how a talker realizes a speech segment has been learned, and that knowledge is stored prelexically, then it can automatically be used to assist in the recognition of all words containing that segment that that talker might produce. Further evidence of lexical generalization of perceptual learning has come from experiments examining adjustments to vocoded speech (Davis et al., 2005) and to an artificial dialect (Maye et al., 2008).

Findings from subliminal priming studies (Kouider & Dupoux, 2005), identification tasks (Nearey, 1990), phonological priming (Radeau et al., 1995; Slowiaczek et al., 2000), and studies on phoneme-sequence learning (Onishi et al. 2002) also support prelexical abstraction. Yet more evidence comes from the second-language literature. Listeners have considerable difficulty learning new phonemic categories (Logan et al., 1991; Strange, 1995), and are influenced by the phonemic categories of their native language while listening to a nonnative language (Best, 1994; Pallier et al., 2001; Weber and Cutler, 2004; Cutler et al., 2006). These findings suggest that, once abstract prelexical categories have been acquired, they necessarily influence speech recognition, even in a second language.

There is therefore considerable evidence for prelexical abstraction. According to an extreme version of the abstractionist position, details about speaking rate, talker, etc., could be discarded prior to lexical access. But there is also considerable evidence that episodic details of words (e.g., how individual talkers produced specific words) are preserved in long-term memory, as measured, for example, in recognition memory experiments (Martin et al., 1989; Goldinger et al., 1991; Palmeri et al., 1993; Church & Schacter, 1994; Goldinger, 1996; Luce & Lyons, 1998): Words are recognized better as having already occurred in the experiment if they are repeated by the same talker. There are also effects of talker-specific detail when participants have to repeat words that they hear (Goldinger, 1998): Repetitions tend to become more like the way the input talker produces them. All of these results show that talker-specific detail cannot be thrown away during word recognition, and thus that extreme abstractionist models are not tenable. But extreme episodic models – in which all acoustic-phonetic detail, including talker-specific attributes, is passed on to the lexicon without normalization – are equally untenable. Such models have the disadvantage that they (unlike abstractionist models) have no ready solution to the invariance problem, and they cannot account for the experimental evidence on abstraction.

What appears to be required, therefore, is a hybrid model in which there is prelexical abstraction, but in which episodic details are not thrown away. On this view, talker-specific features in the speech signal (and other situational details) may be stored in nonlinguistic long-term memory (i.e., not in the mental lexicon), just like other episodic memories (e.g., for faces, colors, or odors), but may also be used in the word-recognition process. That is, just as there appears to be prelexical normalization for speaking rate, there may also be prelexical talker normalization. Indeed, there is evidence that prelexical processing involves adjustments based on talker variability. As already noted, for example, perceptual learning about unusual segments can be talker-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007). Mullennix et al. (1989) showed that listeners could identify words more easily in lists spoken by a single talker than when the same word-lists were spoken by 15 different talkers, and that this effect was more marked when the speech signal was physically degraded. Nygaard et al. (1994; see also Nygaard & Pisoni, 1998) found, in addition, that familiarity with voices (after extensive training in associating those novel voices with names) made it easier to recognize new words spoken by those voices. Adjustment to talker differences thus appears to occur at a prelexical level, just like rate normalization.

Mullennix and Pisoni (1990) have further shown that talker normalization, again like rate normalization, is mandatory. Subjects could not ignore voice variability when categorizing unambiguous initial phonemes in lists of words spoken by one or several talkers, nor could they ignore variability in those initial consonants when categorizing the words as being spoken by either a male or female speaker. Asymmetries in this interference suggested that extraction of phonetic and speaker information are independent but closely related processes: phonetic decisions appear to be at least partially contingent upon the process of talker normalization, and vice versa.

Segmental information is thus extracted prelexically, recoded into linguistically abstract representations, and used in word recognition. The evidence summarized in section 3.1 suggests that this process operates in cascade. The evidence just reviewed suggests further that this abstraction process is not destructive: acoustic-phonetic details about rate, voice, and so on are used by the normalization process, but are then stored rather than discarded.

4 Suprasegmental Information

As Lehiste (1970) pointed out, it is difficult to carve out a domain of speech research dealing just with suprasegmentals. All speech is realized in time with every subcomponent having a measurable duration, fundamental frequency (f_0) and amplitude. Segment identification thus depends on computations involving f_0, duration (for instance, of vocal tract closure), or amplitude (for instance, of frication in a given frequency range; see Johnson, 2005; Raphael, 2005, for reviews). In quantity languages, there are contrasts between long and short versions of the same segment (Estonian vowels have three levels of duration, for example). Nevertheless, the durational, amplitude, and f_0 patterns of speech also encode structural information at higher levels, and listeners exploit this information in the process of recognizing words (section 4.1), parsing prosodic structure (section 4.2), and segmenting the continuous speech stream (section 4.3).

4.1 Suprasegmental cues to lexical identity

Just as durational contrasts at the segment level allow listeners to distinguish between words, so do durational contrasts at the syllable level; for instance, the greater length of a syllable in isolation than in a polysyllabic sequence (*speed* by itself is longer than *speed* in *speedy* or *speediness*; Lehiste, 1971) allows listeners to know whether they are hearing *ham* or *hamster*, *dock* or *doctor* (Davis et al., 2002; Salverda et al., 2003). This usefully allows listeners to avoid accidentally recognizing words which are only spuriously present in the speech signal, embedded in longer words (such as *ham* in *hamster*). However, there are also ways in which word pairs which are segmentally identical may contrast suprasegmentally.

In languages with lexical tone, such as Mandarin, Vietnamese, or Yoruba, syllables are realized with a f_0 pattern which is phonemically contrastive. Word recognition in such languages depends on processing this f_0 information; to all intents and purposes the tones function at the segmental level, so that a given vowel with a rise–fall tone and the same vowel with a level tone may be regarded as different segments. The available experimental evidence on spoken-word recognition in tone languages indeed supports a parallelism between segmental processing and the processing of tonal information.

For instance, lexical information can affect tone categorization in just the same way as it affects segment categorization. In the segment categorization study of Ganong (1980), listeners' category boundaries between /t/ and /d/ shifted to produce more /d/ responses preceding *-eep* but more /t/ responses preceding *-each*; similarly, in a tone categorization study by Fox and Unkefer (1985), listeners' category boundaries between two tones of Mandarin Chinese shifted as a function of which endpoint tone produced a real word given the syllable the tone was produced on. (This was of course only true when the listeners were Mandarin speakers; English listeners showed no such shift.) Lexical priming studies in Cantonese also suggest that the role of a syllable's tone in word recognition is analogous to the role of the vowel (Yip, 2001; Lee, 2007). The f_0 cues to tone are realized over vocalic segments, but the consequence of this is that vowel identity is apprehended more rapidly than the tone information encoded in the same portion of the speech signal (Cutler & Chen, 1997; Ye & Connine, 1999); thus listeners can detect the difference between two CV syllables with the same onset and the same tone but a different vowel more rapidly than they can detect the difference between two CV syllables with the same onset and the same vowel but a different tone. Suprasegmental contrasts between lexical items in other languages pattern across syllable sequences. In pitch accent languages such as Japanese, words exhibit one of a small number of permissible patterns across syllables (thus the accent pattern of *Toyota* is HLL and of *Mitsubishi* LHLL; in each case the syllable labeled H is accented). Although pitch accent patterns are defined across words, however, their realization in the f_0 contour is easily apprehensible for listeners. Listeners can tell from which of two words differing in accentual structure a given syllable has been extracted (e.g, *ka* from *baka* HL vs. *gaka* LH; Cutler & Otake, 1999), incorrect accent patterns delay word identification (Minematsu & Hirose, 1995), and accent patterns are used to distinguish between competing word candidates in spoken-word recognition, so that, for example, listeners hearing *na-* from *nagasa* HLL know that it cannot be the beginning of *nagashi* LHH (Cutler & Otake, 1999; Sekiguchi & Nakajima, 1999).

Like pitch accent, word stress patterns in lexical stress languages are realized across polysyllabic sequences. There is now an extensive literature on the realization and perception of lexical stress, which has recently been analyzed in detail by Cutler (2005). Listeners can use lexical stress patterns in word recognition, too, so that Dutch listeners can accurately tell from which of two words differing in stress a given syllable has been extracted (e.g, *voor* from initially stressed *voornaam* 'first name' vs. finally stressed *voornaam* 'respectable'; Cutler & Donselaar, 2001), incorrect stress in Dutch words inhibits word recognition (van Leyden & van Heuven, 1996), and Dutch listeners can use stress to distinguish between competing word candidates in spoken-word recognition, so that hearing *domi-* from initially stressed *dominee* 'minister' is evidence that it cannot be the beginning of finally stressed *dominant* 'dominant' (Donselaar et al., 2005). The same results are observed in Spanish: *espi-* from finally stressed *espiral* 'spiral' is perceived as evidence against *espiritu* 'spirit' with stress on the second syllable (Soto-Faraco et al., 2001).

However, the same experiment in English, testing for example *admi-* from initially stressed *admiral* versus *admiration* with stress on the third syllable, produces a much weaker effect (Cooper et al., 2002). There is also considerable evidence that mis-stressing effects in English are quite weak unless vowel quality is also changed (Bond & Small, 1983; Cutler & Clifton, 1984; Small et al., 1988; Slowiaczek, 1990), and cross-splicing English vowels with different stress patterns likewise produces unacceptable results only if vowel quality is changed (Fear et al., 1995). In Fear et al.'s study, listeners heard tokens of, say, *autumn*, which has primary stress on the initial vowel, and *audition*, which has an unstressed but unreduced vowel, with the initial vowels exchanged; they rated these tokens as insignificantly different from the original, unspliced, tokens.

The difference in the patterns of results across lexical stress languages can be ascribed to the relative usefulness of stress in distinguishing between words. It is true that studies of English vocabulary structure show that stress pattern information could be of use in word recognition; thus a partial phonetic transcription which includes stress pattern information applies to a smaller candidate set of words than one which does not (Aull, 1984; Waibel, 1988), and an automatic recognition algorithm operating at this level of phonetic specification performs significantly better with stress pattern information than without (Port et al., 1988). But the equivalent effects in Spanish and Dutch are very much larger (Cutler & Pasveer, 2006). This is because of the widespread reduction of vowels in unstressed syllables in English. Cognate examples of the cross-language asymmetry abound. For example, stressed word-initial *com-* in English could be the beginning of *comedy*, or of *comma*, *compliment*, etc., but *comedy*'s morphological relative *comedian* has the reduced vowel schwa in its initial syllable. In Spanish *comedia* 'comedy' (stress on the second syllable) and *comico* 'comedian' (initial stress) have the same vowel in the first syllables; in Dutch, *komedie* 'comedy' (stress on the second syllable) and *komediant* 'comedian' (stress on the fourth syllable) have the same vowels in their first and in their second syllables. English listeners can instantly distinguish between *comedy* and *comedian* as word candidates on the basis of the vowel in the first syllable, whereas Spanish or Dutch listeners can only achieve such early discrimination by paying attention to the syllable's stress.

Thus listeners are usually able to distinguish between words in English using segmental information alone; suprasegmental information can be ignored with relatively little penalty. In Dutch, which has much less vowel reduction, and in Spanish, which does not allow vowel reduction at all, the penalty for ignoring suprasegmental information would be much more severe. This explains the different strength of the effects in the word recognition experiments. It explains why Dutch listeners outperform native English listeners at telling from which of two words differing in stress a given English syllable has been extracted (e.g, *mus-* from *music* versus *museum*; Cooper et al., 2002). The lesson from these studies is that quite small cross-language differences in how a given level of structure is realized (e.g., the likelihood of unstressed syllables manifesting vowel reduction) can radically affect the value of the structural information for rapid distinguishing between words, and thus determine the degree to which listeners make use of acoustic cues to that structure in speech.

4.2 Prosodic structure

The suprasegmental patterns of speech encode the prosodic structure in utterances. Even though prosodic and syntactic structure are independently determined (Shattuck-Hufnagel & Turk, 1996), it has long been known that listeners derive syntactic boundaries, and discourse boundaries too, from phrase-final lengthening and from the f_0 contour (see Cutler et al., 1997, for a review). More recently it has also become clear that the fine structure of segments is influenced by prosodic boundary placement, such that, for example, segments at the onset of a prosodic constituent are strengthened (Keating et al., 2003).

Listeners make use of such effects to parse speech into words. Cho et al. (2007) examined potentially ambiguous sequences such as *bus tickets*. There are competing English words in which there is no boundary after the /s/ (*bust*, *busty*), and their question was: could prosodic strengthening assist listeners in recognizing the word *bus* in this potentially ambiguous context? They compared utterances such as *When you get on the bus, tickets should be shown to the driver* (in which prosodic strengthening should enhance the /t/) versus *John bought several bus tickets for his family* (no strengthening). They found that the word *bus* was indeed more easily recognized in the phrase *bus tickets* taken from the former context than in the same phrase taken from the latter context. Likewise, Christophe et al. (2004) found that French word sequences such as *chat grinchaux* ('grumpy cat') were likely to be confused with the accidentally embedded *chagrin* if they formed part of a single phrase, but not if a phrase boundary occurred after *chat*.

Listeners are capable of using differences in the duration of segments within words to distinguish between alternative candidates (for instance, the difference in duration of syllable-final /l/ in Italian *silvestre* 'sylvan' versus syllable-initial /l/ in *silencio* 'silence' is available even without the following /v/ or /ɛ/ which disambiguates; Tabossi et al., 2000). Listeners are also capable of using the same sort of differences to distinguish between alternative phrases (for instance, the

/s/ duration distinguishes Dutch *een spot* 'a spotlight' from *eens pot* 'jar once'; Shatzman & McQueen, 2006a). The differences which are used vary across languages. Thus, in both English and Dutch, listeners make use of prosodic strengthening of /t/ to infer the onset of a prosodic constituent, but the way the strengthening is realized is exactly opposite in these languages (Cho & McQueen, 2005). In English, a strengthened /t/ has a longer VOT (which enhances the contrast with short-VOT /d/). In Dutch, strengthened /t/ has a shorter VOT (because the strengthening consists in a longer closure, enhancing the contrast with /d/ which in Dutch is prevoiced).

Even the durational differences which allow listeners to distinguish between stand-alone words and accidentally embedded words such as *hamster* versus *ham* embedded in *hamster* (Davis et al., 2002; Salverda et al., 2003; see also Shatzman & McQueen, 2006b) are modulated by prosodic structure. Thus *taxi* (/taksi/) is embedded in both the sequence *pak de tak sinaasappels* 'grab the branch of oranges' and *pak de tak citroenen* 'grab the branch of lemons'; but in the first, the syllable /si/ after *tak* is stressed, while in the second it is unstressed just as is the second syllable of *taxi*. Salverda et al. made two versions of a sentence containing, for example, *taxi*, by cross-splicing the *tak* syllable either from the above *sinaasappels* context or from the *citroenen* context. They found that the competing shorter word *tak* was more available in the former case. Moreover, they found that what primarily influenced the relative availability of the words was the duration of the initial syllable in the ambiguous portion; *ham*, *tak*, etc. are longer than the same syllables in *hamster*, *taxi*, etc., but the isolated words are also longer when they are followed by a stressed rather than an unstressed syllable. By manipulating the duration of this syllable, Salverda et al. found that they could influence what listeners considered as the most likely word at that moment.

4.3 Rhythmic structure and the recognition of continuous speech

Speech is a continuous signal without consistent demarcation of the words which make it up. Listeners must extract the component words from each speech signal in order to understand the speaker's message. One of the ways they do this is by exploiting the relationship between the rhythmic structure of speech and word-boundary location.

In English, and similar languages such as Dutch, rhythmic structure is stress-based, and segmentation of continuous speech can be usefully based on an assumption that strong syllables are most likely to be word-initial. Evidence that English- and Dutch-speaking listeners do actually operate with such an assumption comes partly from studies of word boundary misperceptions, in which listeners most commonly err by assuming strong syllables to be word-initial and weak syllables to be noninitial (Cutler & Butterfield, 1992; Vroomen et al., 1996). Further evidence is to be found in studies with the word-spotting task, in which real words embedded in nonsense bisyllables are harder to detect if detecting them requires processing segments from two consecutive strong syllables, i.e., across the canonical point of speech segmentation (Cutler & Norris, 1988; McQueen et al., 1994; Norris et al., 1995; Vroomen & de Gelder, 1995; Vroomen et al., 1996). Syllable strength is here encoded in terms of vowel quality; recall that Fear et al.'s (1995) study described in section 4.1 showed that this is the most important feature of syllable strength in English. Cutler and Norris (1988), for example, compared detection of the word *mint* in *mintayf* and *mintef*; both bisyllables had initial stress, but they differed in the vowel which occurred in the second syllable. The embedded word *mint* was much harder to detect when the second vowel was strong, as in *mintayf*. The explanation is that the strong syllable *-tayf* triggered speech segmentation, so that the /t/ was momentarily considered to be the beginning of a new word rather than the end of *mint*; recovering from this interfering segmentation delayed recognition of *mint*.

In other languages with different rhythmic patterns, segmentation procedures also exploit rhythmic structure: syllabic rhythm in French (Cutler et al., 1986, 1992; Kolinsky et al., 1995) and Korean (Kim et al., 2008), moraic rhythm in Japanese (Otake et al., 1993; Cutler & Otake, 1994) and Telugu (Murty et al., 2007). In fact, rhythm allows a single, universally valid description of otherwise very different segmentation procedures used across languages (see Cutler, 1994, for further detail of this proposal).

Rhythmic structure allows listeners to predict accentual patterning as well. The initial phonemes of nonsense words are detected more rapidly when sentence rhythm predicts that the syllables containing the target will be accented (Shields et al., 1974). Pitt and Samuel (1990) presented acoustically constant versions of disyllabic minimal stress pairs at the ends of auditory lists in which all the disyllabic items had the same stress pattern; detection of a phoneme in these words was again faster when the syllable containing the target phoneme was predicted to be stressed, suggesting that listeners used the predictive information to attend selectively to stressed syllables. Likewise, listeners direct attention to words bearing sentence accent; detection of the initial phoneme of an acoustically constant word token is faster when the word occurs in a prosodic context consistent with sentence accent falling at that point than when it occurs in a context consistent with lack of accent (Cutler, 1976; Cutler & Darwin, 1981). Listeners can derive sufficient information to perform this attentional focus when f_0 variation has been removed (Cutler & Darwin, 1981), although when dimensions of prosodic information conflict such that, for example, timing predicts accent where f_0 predicts no accent, listeners refrain from deriving predictive information from prosody at all (Cutler, 1987).

5 Conclusions

Knowledge about the cognitive processes in speech perception has advanced considerably since we wrote our chapter in the first edition of this book. This development can be seen most clearly, perhaps, in the way the questions that are asked about this field have changed. In the 1990s, binary questions were being asked: Does lexical knowledge control prelexical decisions, or not? Is prelexical processing serial or is it cascaded? Is speech perception episodic or abstractionist? The questions we now have to ask are more nuanced. With respect to the issue of feedback of lexical knowledge, for example, recent research has shown that we have to distinguish between feedback in on-line processing and feedback in perceptual learning. A simple yes/no answer to whether there is feedback or not is no longer appropriate. Furthermore, given that it has been established that phonetic information cascades continuously to the lexical level, we now have to ask more specific questions about the nature of the information that is passed forward to lexical processing. Might its role depend on its informational value, as our review of cross-linguistic differences on uptake of lexical stress information suggests? Finally, we now have to consider how speech recognition can be episodic and abstractionist at the same time. All of these developments are indicative of a very active field of enquiry; they show that real progress is being made in our understanding of the cognitive aspects of speech perception.

The field is flourishing in another way. Even though the questions have become more refined, there are just as many being asked, and some that have still not been answered. For instance, we do not yet know what the "units of perception" are (McNeill & Lindig, 1973; Healy & Cutting, 1976). It seems clear that there is a prelexical level of processing that mediates between auditory (i.e., not speech-specific) processing and lexical processing, that the prelexical level involves abstraction and normalization, and that it operates in cascade. Furthermore, while it appears that this level is impervious to the immediate influence of lexical feedback, lexical knowledge can be used to retune prelexical processing over time. But we do not yet know what the unit(s) of representation are at the prelexical level. An important issue for future research will be to specify whether linguistic abstraction of segmental information prior to lexical access involves, for example, featural, allophonic, phonemic, or syllabic representations.

Critically, however, the way in which suprasegmental information is extracted prelexically will also have to be specified, and an account will have to be developed for the way in which segmental and suprasegmental information is integrated in modulating the word-recognition process. One possibility is that there are indeed two processing channels, one extracting segmental material (e.g., a mechanism computing the current sequence of phones), and one extracting suprasegmental material (e.g., a device building the prosodic structure of the current utterance). These two prelexical channels could operate independently, but could still both influence lexical processing. Another possibility is that there is a single processing channel which constructs an integrated multidimensional structure consisting of larger and smaller elements.

The current weight of evidence favors an autonomous account of on-line processing. If convincing data for on-line lexical–prelexical feedback were to be found, however, then it would be necessary to establish the cognitive function that such feedback serves in normal listening. Since on-line feedback appears to be of no benefit to word recognition, one possibility is that on-line effects are an epiphenomenon of the need for feedback in perceptual learning. If, however, the conclusion in favor of autonomy in on-line processing continues to stand the test of time, then it will be necessary to develop a model of speech recognition in which lexical knowledge cannot modulate prelexical processing as it is happening, but can retune those processes over time.

An important constraint on the operation of the prelexical level, and thus possibly also on the nature of the representations constructed there, is that it must be flexible. The evidence we reviewed on perceptual learning in speech suggests that the way in which the speech signal is mapped onto the lexicon can be retuned after very brief exposure, and that that retuning can be specific to the speech of a single talker. It will be important to ascertain what the limits are on this kind of flexibility. For example, might there also be retuning of suprasegmental representations, and are other sources of knowledge (i.e., other than the lexicon) used to supervise perceptual learning?

Perhaps the greatest current challenge for cognitive modeling of speech perception, however, is how to include abstractionist and episodic components in the same model. Recent research suggests that episodic detail (e.g., how individual talkers produce specific speech sounds) is used to modulate the prelexical level. But this contribution of episodic knowledge to the flexibility of the prelexical processor is unlikely to be the whole story. Recent research also suggests that details of encounters with specific tokens of words are stored in long-term memory. The question, then, is how those long-term memories relate to abstract linguistic processing: do they exist only at the prelexical level, or also at the lexical level, or do they reside in a more general episodic memory store (i.e., not in the mental lexicon)?

It is important to note that this debate does not concern talker-specific segmental details alone. It also concerns suprasegmental details. For example, tokens of words differ in acoustic-phonetic detail because of their position in the prosodic hierarchy. Are all of these tokens stored, or is there abstraction of prosodic knowledge? This debate is also about the role of word frequency. One way in which the frequency of occurrence of a word can be coded is through storage of all encounters with that word, as in episodic models. But frequency can also be handled by models with abstract representations (either at the prelexical level or the lexical level, or both). Reconciling abstractionist and episodic accounts will thus entail specifying how multiple sources of information – about segments, suprasegmental structures, talker- and situation-specific details, and lexical frequency – are brought together as listeners hear spoken words. Experimentalists and computational modelers have plenty still to do.

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998) Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419–39.

Alphen, P. M. van & McQueen, J. M. (2006) The effect of voice onset time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 178–96.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994) The effect of subphonetic differences on lexical access. *Cognition*, 52, 163–87.

Aull, A. M. (1984) Lexical stress and its application in large vocabulary speech recognition. Masters dissertation, Massachusetts Institute of Technology.

Best, C. T. (1994) The emergence of language-specific phonemic influences in infant speech perception. In J. Goodman & H. C. Nusbaum (eds.), *Development of Speech Perception: The Transition from Speech Sounds to Spoken Words* (pp. 167–224) Cambridge, MA: MIT Press.

Bond, Z. & Small, L. H. (1983) Voicing, vowel, and stress mispronunciations in continuous speech. *Perception and Psychophysics*, 34, 470–4.

Burton, M. W., Baum, S. R., & Blumstein, S. E. (1989) Lexical effects on the phonetic categorization of speech: The role of acoustic structure. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 567–75.

Cho, T. & McQueen, J. M. (2005) Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, 33, 121–57.

Cho, T., McQueen, J. M., & Cox, E. A. (2007) Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, 35, 210–43.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004) Phonological phrase boundaries constrain lexical access, I: Adult data. *Journal of Memory and Language*, 51, 523–47.

Church, B. A. & Schacter, D. L. (1994) Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 521–33.

Cluff, M. S. & Luce, P. A. (1990) Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 551–63.

Cole, R. A. & Scott, B. (1974) Toward a theory of speech perception. *Psychological Review*, 81, 348–74.

Connine, C. M., Blasko, D. G., & Titone, D. (1993) Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32, 193–210.

Connine, C. M. & Clifton, C. (1987) Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 291–9.

Connine, C. M., Titone, D., Deelman, T., & Blasko, D. (1997) Similarity mapping in spoken word recognition. *Journal of Memory and Language*, 37, 463–80.

Cooper, N., Cutler, A., & Wales, R. (2002) Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, 45, 207–28.

Cutler, A. (1976) Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, 20, 55–60.

Cutler, A. (1987) Components of prosodic effects in speech recognition. *Proceedings of the 11th International Congress of Phonetic Sciences*, Tallinn, Estonia, 1, 84–7.

Cutler, A. (1994) Segmentation problems, rhythmic solutions. *Lingua*, 92, 81–104.

Cutler, A. (2005) Lexical stress. In D. B. Pisoni & R. E. Remez (eds.), *The Handbook of Speech Perception* (pp. 264–89). Oxford: Blackwell.

Cutler, A. & Butterfield, S. (1992) Rhythmic cues to speech segmentation:

Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218–36.

Cutler, A. & Chen, H.-C. (1997) Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics*, 59, 165–79.

Cutler, A. & Clifton, C. (1984) The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (eds.), *Attention and Performance, 10: Control of Language Processes* (pp. 183–96). Hillsdale, NJ: Lawrence Erlbaum.

Cutler, A., Dahan, D., & Donselaar, W. van (1997) Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40, 141–201.

Cutler, A. & Darwin, C. J. (1981) Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, 29, 217–24.

Cutler, A. & Donselaar, W. van (2001) *Voornaam* is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech*, 44, 171–95.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986) The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385–400.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987) Phoneme identification and the lexicon. *Cognitive Psychology*, 19, 141–77.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992) The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, 24, 381–410.

Cutler, A. & Norris, D. (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–21.

Cutler, A. & Otake, T. (1994) Mora or phonemes? Further evidence for language-specific listening. *Journal of Memory and Language*, 33, 824–44.

Cutler, A. & Otake, T. (1999) Pitch accent in spoken-word recognition in Japanese. *Journal of the Acoustical Society of America*, 105, 1877–88.

Cutler, A. & Pasveer, D. (2006) Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. *Proceedings of Speech Prosody 2006*, Dresden, Germany, 237–400.

Cutler, A., Weber, A., & Otake, T. (2006) Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, 34, 269–84.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001) Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 16, 507– 34.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005) Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134, 222–41.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. (2002) Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 218–44.

Dell, G. S. & Newman, J. E. (1980) Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, 19, 608–23.

Donselaar, W. van, Koster, M., & Cutler, A. (2005) Exploring the role of lexical stress in lexical recognition. *Quarterly Journal of Experimental Psychology*, 58A, 251–73.

Eimas, P. D. & Corbit, J. D. (1973) Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4, 99–109.

Eisner, F. & McQueen, J. M. (2005) The specificity of perceptual learning in speech processing. *Perception and Psychophysics*, 67, 224–38.

Eisner, F. & McQueen, J. (2006) Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, 119, 1950–3.

Elman, J. L. & McClelland, J. L. (1986) Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (eds.), *Invariance and Variability of Speech Processes* (pp. 360–80). Hillsdale, NJ: Lawrence Erlbaum.

Elman, J. L. & McClelland, J. L. (1988) Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27, 143–65.

Ernestus, M. & Mak, W. M. (2004) Distinctive phonological features differ in relevance for both spoken and written word recognition. *Brain and Language*, 90, 378–92.

Fear, B. D., Cutler, A., & Butterfield, S. (1995) The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America*, 97, 1893–904.

Foss, D. J. & Blank, M. A. (1980) Identifying the speech codes. *Cognitive Psychology*, 12, 1–31.

Foss, D. J. & Gernsbacher, M. A. (1983) Cracking the dual code: Toward a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, 22, 609–32.

Foss, D. J., Harwood, D. A., & Blank, M. A. (1980) Deciphering decoding decisions: Data and devices. In R. A. Cole (ed.), *Perception and Production of Fluent Speech* (pp. 165–99). Hillsdale, NJ: Lawrence Erlbaum.

Fowler, C. A. (1984) Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, 36, 359–68.

Fox, R. A. (1984) Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 526–40.

Fox, R. A. & Unkefer, J. (1985) The effect of lexical status on the perception of tone. *Journal of Chinese Linguistics*, 13, 69–90.

Frauenfelder, U. H., Segui, J., & Dijkstra, T. (1990) Lexical effects in phonemic processing: Facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance*, 16, 77–91.

Ganong, W. F. (1980) Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 110–25.

Gaskell, M. & Marslen-Wilson, W. D. (1997) Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12, 613–56.

Gaskell, M. & Marslen-Wilson, W. D. (2002) Representation and competition in the perception of spoken words. *Cognitive Psychology*, 45, 220–66.

Goldinger, S. D. (1996) Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1166–83.

Goldinger, S. D. (1998) Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251–79.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989) Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28, 501–18.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992) Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1211–38.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991) On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 17, 152–62.

Gordon, P. C. (1988) Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception and Psychophysics*, 43, 137–46.

Gow, D. W., Jr. (2002) Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, 28, 163–179.

Gow, D. W., & Gordon, P. C. (1995) Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 344–59.

Hawkins, S. (2003) Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, 373–405.

Healy, A. F. & Cutting, J. E. (1976) Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior*, 15, 73–83.

Johnson, K. (1997) Speech perception without speaker normalization: An exemplar model. In K. A. Johnson & J. W. Mullennix (eds.), *Talker Variability in Speech Processing* (pp. 145–66). San Diego, CA: Academic Press.

Johnson, K. (2005) Speaker normalization in speech perception. In D. B. Pisoni & R. Remez (eds.), *The Handbook of Speech Perception* (pp. 363–89). Oxford: Blackwell.

Johnson, K. A. & Mullennix, J. W. (eds.) (1997) *Talker Variability in Speech Processing*. San Diego, CA: Academic Press.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. (2003) Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (eds.), *Laboratory Phonology 6* (pp. 143–61). Cambridge: Cambridge University Press.

Kim, J., Davis, C., & Cutler, A. (2008) Perceptual tests of rhythmic similarity, II: Syllable rhythm. *Language and Speech*, 51, 343–59.

Klatt, D. H. (1979) Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (ed.), *Perception and Production of Fluent Speech* (pp. 243–88). Hillsdale, NJ: Lawrence Erlbaum.

Klatt, D. H. (1989) Review of selected models of speech perception. In W. D. Marslen-Wilson (ed.), *Lexical Representation and Process* (pp. 169–226). Cambridge, MA: MIT Press.

Kolinsky, R., Morais, J., & Cluytens, M. (1995) Intermediate representations in spoken word recognition: Evidence from word illusions. *Journal of Memory and Language*, 34, 19–40.

Kouider, S. & Dupoux, E. (2005) Subliminal speech priming. *Psychological Science*, 16, 617–25.

Kraljic, T. & Samuel, A. G. (2005) Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51, 141–78.

Kraljic, T. & Samuel, A. G. (2006) Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, 13, 262–8.

Kraljic, T. & Samuel, A. G. (2007) Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56, 1–15.

Lee, C.-Y. (2007) Does horse activate mother? Processing lexical tone in form priming. *Language and Speech*, 50, 101–23.

Lehiste, I. (1970) *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. (1971) The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, 51, 2018–24.

Leyden, K. van & Heuven, V. J. van (1996) Lexical stress and spoken word recognition: Dutch vs. English. In C. Cremers & M. den Dikken (eds.), *Linguistics in the Netherlands 1996* (pp. 59–170). Amsterdam: John Benjamins.

Liberman, A. M. & Mattingly, I. G. (1985) The motor theory of speech perception revised. *Cognition*, 21, 1–36.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991) Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89, 874–86.

Luce, P. A. & Large, N. R. (2001) Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16, 565–81.

Luce, P. A. & Lyons, E. A. (1998) Specificity of memory representations for spoken words. *Memory and Cognition*, 26, 708–15.

Luce, P. A. & Pisoni, D. B. (1998) Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–36.

Maddieson, I. (1984) *Patterns of Sounds*. Cambridge: Cambridge University Press.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003) Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science*, 27, 801–5.

Mann, V. A. & Repp, B. H. (1981) Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, 69, 548–58.

Marslen-Wilson, W. D. (1987) Functional parallelism in spoken word-recognition. *Cognition*, 25, 71–102.

Marslen-Wilson, W. (1990) Activation, competition, and frequency in lexical access. In G. T. M. Altmann (ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 148–72). Cambridge, MA: MIT Press.

Marslen-Wilson, W., Moss, H. E., & Halen, S. van (1996) Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1376–92.

Marslen-Wilson, W. & Warren, P. (1994) Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, 101, 653–75.

Martin, C., Mullennix, J., Pisoni, D., & Summers, W. (1989) Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 676–84.

Maye, J., Aslin, R. N., & Tanenhaus, M. T. (2008) The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science*, 32, 543–62.

McClelland, J. L. (1987) The case for interactionism in language processing. In M. Coltheart (ed.), *Attention and Performance 12: The Psychology of Reading* (pp. 1–36). London: Lawrence Erlbaum.

McClelland, J. L. (1991) Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23, 1–44.

McClelland, J. L. & Elman, J. L. (1986) The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006) Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, 10, 363–9.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002) Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86, B33–B42.

McNeill, D. & Lindig, K. (1973) The perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior*, 12, 431–61.

McQueen, J. M. (1991) The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 433–43.

McQueen, J. M. (1993) Rhyme decisions to spoken words and nonwords. *Memory and Cognition*, 21, 210–22.

McQueen, J. M. (2003) The ghost of Christmas future: Didn't Scrooge learn to be good? Commentary on Magnuson, McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science*, 27, 795–9.

McQueen, J. M., Cutler, A., & Norris, D. (2003) Flow of information in the spoken word recognition system. *Speech Communication*, 41, 257–70.

McQueen, J. M., Cutler, A., & Norris, D. (2006) Phonological abstraction in the mental lexicon. *Cognitive Science*, 30, 1113–26.

McQueen, J. M., Jesse, A., & Norris, D. (2009) No lexical-prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language*, 61, 1–18.

McQueen, J. M., Norris, D., & Cutler, A. (1994) Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 621–38.

McQueen, J. M., Norris, D., & Cutler, A. (1999) Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 1363–89.

McQueen, J. M., Norris, D., & Cutler, A. (2006) Are there really interactive processes in speech perception? *Trends in Cognitive Science*, 10, 533.

Mehler, J. (1981) The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society of London*, 295, 333–52.

Miller, J. L. (1981) Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (eds.), *Perspectives on the Study of Speech* (pp. 39–74). Hillsdale, NJ: Lawrence Erlbaum.

Miller, J. L. (1987) Mandatory processing in speech perception: A case study. In J. L. Garfield (ed.), *Modularity in Knowledge Representation and Natural-Language Understanding* (pp. 309–22). Cambridge, MA: MIT Press.

Miller, J. L. & Dexter, E. R. (1988) Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 369–78.

Miller, J. L., Green, K., & Schermer, T. M. (1984) A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception and Psychophysics*, 36, 329–37.

Minematsu, N. & Hirose, K. (1995) Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese. *Journal of the Acoustical Society of Japan*, 16, 311–20.

Mirman, D., McClelland, J. M., & Holt, L. L. (2005) Computational and behavioral investigations of lexically induced delays in phoneme recognition. *Journal of Memory and Language*, 52, 424–43.

Mitterer, H. & Ernestus, M. (2006) Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch. *Journal of Phonetics*, 34, 73–103.

Mullennix, J. W. & Pisoni, D. B. (1990) Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, 47, 379–90.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989) Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, 85, 365–78.

Murty, L., Otake, T., & Cutler, A. (2007) Perceptual tests of rhythmic similarity, I: Mora rhythm. *Language and Speech*, 50, 77–99.

Nearey, T. M. (1990) The segment as a unit of speech perception. *Journal of Phonetics*, 18, 347–73.

Newman, J. E. & Dell, G. S. (1978) The phonological nature of phoneme monitoring: A critique of some ambiguity studies. *Journal of Verbal Learning and Verbal Behavior*, 17, 359–74.

Norris, D. (1994) Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52, 189–234.

Norris, D. & McQueen, J. M. (2008) Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115, 357–95.

Norris, D., McQueen, J. M., & Cutler, A. (1995) Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1209–28.

Norris, D., McQueen, J. M., & Cutler, A. (2000) Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003) Perceptual learning in speech. *Cognitive Psychology*, 47, 204–38.

Nygaard, L. C., & Pisoni, D. B. (1998) Talker-specific learning in speech perception. *Perception and Psychophysics*, 60, 355–76.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994) Speech perception as a talker-contingent process. *Psychological Science*, 5, 42–6.

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002) Learning phonotactic constraints from brief auditory exposure. *Cognition*, 83, B13–B23.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993) Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258–78.

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001) The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, 12, 445–9.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993) Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 309–28.

Pierrehumbert, J. (2002) Word-specific phonetics. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7* (pp. 101–40). Berlin: Mouton de Gruyter.

Pitt, M. A. & McQueen, J. M. (1998) Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39, 347–70.

Pitt, M. A. & Samuel, A. G. (1990) The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 564–73.

Pitt, M. A. & Samuel, A. G. (1993) An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 699–725.

Port, R. F., Reilly, W. T., & Maki, D. P. (1988) Use of syllable-scale timing to discriminate words. *Journal of the Acoustical Society of America*, 83, 265–73.

Radeau, M., Morais, J., & Segui, J. (1995) Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 1297–311.

Raphael, L. J. (2005) Acoustic cues to the perception of segmental phonemes. In D. Pisoni & R. E. Remez (eds.), *The Handbook of Speech Perception* (pp. 182–206). Oxford: Blackwell.

Repp, B. H. & Liberman, A. M. (1987) Phonetic category boundaries are flexible. In S. R. Harnad (ed.), *Categorical Perception* (pp. 89–112). Cambridge: Cambridge University Press.

Rubin, P., Turvey, M. T., & Gelder, P. van (1976) Initial phonemes are detected faster in spoken words than in non-words. *Perception and Psychophysics*, 19, 394–8.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003) The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90, 51–89.

Samuel, A. G. (1977) The effect of discrimination training on speech perception: Noncategorical perception. *Perception and Psychophysics*, 22, 321–30.

Samuel, A. G. (1981a) Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, 110, 474–94.

Samuel, A. G. (1981b) The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1124–31.

Samuel, A. G. (1987) Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language*, 26, 36–56.

Samuel, A. G. (1996) Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General*, 125, 28–51.

Samuel, A. G. (1997) Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, 32, 97–127.

Samuel, A. G. (2001) Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12, 348–51.

Samuel, A. G. & Kat, D. (1996) Early levels of analysis of speech. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 676–94.

Samuel, A. G. & Pitt, M. A. (2003) Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language*, 48, 416–34.

Segui, J. & Frauenfelder, U. (1986) The effect of lexical constraints upon speech perception. In F. Klix & H. Hagendorf (eds.), *Human Memory and Cognitive Capabilities: Mechanisms and Performances* (pp. 795–808). Amsterdam: North-Holland.

Segui, J., Frauenfelder, U., & Mehler, J. (1981) Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, 72, 471–7.

Sekiguchi, T. & Nakajima, Y. (1999) The use of lexical prosody for lexical access of the Japanese language. *Journal of Psycholinguistic Research*, 28, 439–54.

Shattuck-Hufnagel, S. & Turk, A. E. (1996) A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25, 193–247.

Shatzman, K. B. & McQueen, J. M. (2006a) Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics*, 68, 1–16.

Shatzman, K. B. & McQueen, J. M. (2006b) Prosodic knowledge affects the recognition of newly-acquired words. *Psychological Science*, 17, 372–7.

Shields, J. L., McHugh, A., & Martin, J. G. (1974) Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, 102, 250–5.

Shillcock, R. (1990) Lexical hypotheses in continuous speech. In G. T. M. Altmann (ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 24–49). Cambridge, MA: MIT Press.

Slowiaczek, L. M. (1990) Effects of lexical stress in auditory word recognition. *Language and Speech*, 33, 47–68.

Slowiaczek, L. M. & Hamburger, M. (1992) Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1239–50.

Slowiaczek, L. M., McQueen, J. M., Soltano, E. G., & Lynch, M. (2000) Phonological representations in prelexical speech processing: Evidence from form-based priming. *Journal of Memory and Language*, 43, 530–60.

Small, L. H., Simon, S. D., & Goldberg, J. S. (1988) Lexical stress and lexical access: Homographs versus nonhomographs. *Perception and Psychophysics*, 44, 272–80.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001) Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language*, 45, 412–32.

Spinelli, E., McQueen, J. M., & Cutler, A. (2003) Processing resyllabified words in French. *Journal of Memory and Language*, 48, 233–54.

Stemberger, J. P., Elman, J. L., & Haden, P. (1985) Interference between phonemes during monitoring: Evidence for an interactive activation model of speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 11, 475–89.

Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111, 1872–91.

Stevens, K. & Blumstein, S. (1981) The search for invariant acoustic correlates of phonetic features. In P. Eimas & J. L. Miller (eds.), *Perspectives on the Study of Speech* (pp. 1–38). Hillsdale, NJ: Lawrence Erlbaum.

Strange, W. (1995) *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research*. Timonium, MD: York Press.

Streeter, L. A. & Nigro, G. N. (1979) The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America*, 65, 1533–41.

Swinney, D. (1981) Lexical processing during sentence comprehension: Effects of higher-order constraints and implications for representation. In T. Myers, J. Laver, & J. Anderson (eds.), *The Cognitive Representation of Speech* (pp. 201–9). Amsterdam: North-Holland.

Tabossi, P., Burani, C., & Scott, D. (1995) Word identification in fluent speech. *Journal of Memory and Language*, 34, 440–67.

Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000) Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 758–75.

Utman, J. A., Blumstein, S. E., & Burton, M. W. (2000) Effects of subphonetic and syllable structure variation on word recognition. *Perception and Psychophysics*, 62, 1297–311.

Vitevitch, M. S. & Luce, P. A. (1998) When words compete: Levels of processing in perception of spoken words. *Psychological Science*, 9, 325–9.

Vitevitch, M. S. & Luce, P. A. (1999) Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40, 374–408.

Vroomen, J. & Gelder, B. de (1995) Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 98–108.

Vroomen, J. & Gelder, B. de (1997) Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 710–20.

Vroomen, J., Linden, S. van, Gelder, B. de, & Bertelson, P. (2007) Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45, 572–7.

Vroomen, J., Zon, M. van, & Gelder, B. de (1996) Cues to speech segmentation: Evidence from juncture misperceptions and word spotting. *Memory and Cognition*, 24, 744–55.

Waibel, A. (1988) *Prosody and Speech Recognition*. London: Pitman.

Warren, R. M. (1970) Perceptual restoration of missing speech sounds. *Science*, 167, 392–3.

Warren, R. M. & Obusek, C. J. (1971) Speech perception and phonemic restorations. *Perception and Psychophysics*, 9, 358–62.

Weber, A. & Cutler, A. (2004) Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50, 1–25.

Whalen, D. (1984) Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics*, 35, 49–64.

Whalen, D. (1991) Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics*, 50, 351–60.

Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76, 1–15.

Ye, Y. & Connine, C. M. (1999) Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes*, 14, 609–30.

Yip, M. C. (2001) Phonological priming in Cantonese spoken-word processing. *Psychologia: An International Journal of Psychology in the Orient*, 44, 223–9.

Zwitserlood, P. (1989) The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32, 25–64.

Part IV Linguistic Phonetics

15 The Prosody of Speech: Timing and Rhythm

JANET FLETCHER

1 Introduction

Speech is an activity that unfolds in time, and listeners are sensitive to fluctuations in the temporal flow of an utterance. Teasing apart the role(s) of time and timing in speech communication is not a trivial task, as will become clear when we look at different approaches to timing and prosody in this chapter. The nature of speech timing has been the subject of a great deal of experimental research and many aspects of speech timing have been addressed in other chapters of this volume. The primary goal of this chapter is to present an overview of the major findings in the experimental literature that relate to prosody and the durational patterns of spoken discourse. Even the term prosody has a couple of quite distinct but related meanings. Prosody or *prosodic* features are for many phoneticians and speech scientists synonymous with variations in *suprasegmental* parameters such as duration, intensity, and f_0 that contribute in various combinations to the production and perception of stress, rhythm and tempo, lexical tone, and intonation of an utterance. In other words it refers to those phonetic and phonological aspects of spoken language that cannot necessarily be reduced to individual consonants and vowels but generally extend across several segments or syllables. Traditionally *suprasegmental* timing is viewed as something separate from *segmental* timing. For example, speech rhythm has historically been viewed as a phenomenon that deals with the timing patterns associated with units above the level of the phonetic segment (see section 3). Likewise speech tempo and pausing are also classed as *suprasegmental* timing phenomena (see section 4). However it is quite impossible to exclude subsyllabic timing effects from any discussion of stress, or the lengthening in syllables that is often observed at the edge of larger stretches of speech. This leads us to the more abstract phonological use of the term prosody: as the abstract hierarchical phonological structure(s) of an utterance, and *prominence* relations within that structure. According to Beckman and Edwards (1992, p. 359), "we can think of prosody as the organizational structure that measures off chunks of speech into countable units of various sizes." Like others, they maintain that this view of prosody is built on a direct link between meter or metrics and the organization of spoken communication. This also raises another important question that has been addressed for many years in phonetics research: What do we actually measure in order to get access to prosody and timing in connected speech?

1.1 Speech timing research: What do we measure, and why?

Speech timing research has been undertaken since the development of instrumental techniques made measurement of "visible speech" possible. The duration of *intervals* or temporal patterning of *events*, corresponding to consonants, vowels, and syllable-like speech units are measured, or the articulatory behavior during these intervals or associated with these events. This is not a straightforward process because as Lindblom (1983, p. 495) explains: "the speech signal cannot be unambiguously segmented into temporally non-overlapping chunks corresponding to linear sequences of phonemes, syllables and words." Many speech scientists acknowledge this, but for want of an alternative method of speech signal interpretation, we mostly continue to use the conventional unit definitions when referring to the temporal structures of a language (see also Beckman, 1997; van Santen & Shih, 2000; and Kohler, 2003 for an insightful discussion of some of the issues). A linear segmentation of the speech signal can be used as a structure within which to view elements of variance and discreteness in a physical representation of speech. It is possible to measure intervals corresponding to linguistically significant constructs like phonemes, or syllables, but segment boundaries cannot always be easily identified (e.g., Peterson, 1955, cited in van Santen & Shih, 2000), and syllable definition based solely on phonetic criteria is not possible (e.g., Ladefoged, 1967). The duration of articulatory gestures or movements, and measures of intergestural coordination and timing associated with segments, syllables, or some other kind of interval are also routinely carried out using articulatory monitoring and tracking techniques (see Stone, this volume). Other units, like vowel-to-vowel onsets or the intervals between successive stressed-vowel onsets, are also measured to capture language-specific timing patterns that may or may not depend on linguistic constructs like syllables or stress feet (see, e.g., Pompino-Marschall, 1989; Scott, 1998; Barbosa, 2007). According to one view, it is questionable whether measurement of "intervals" corresponding to segments or syllables tells us everything about timing. Kohler (2003, p. 8) also suggests that these kinds of measurements are not about timing per se. They are direct measures of duration, whereas timing refers to the unfolding of articulatory or perceptual processes and strategies in time.

Whatever the pervasive ideology underlying the speech researcher's motivations, conventions for acoustic segmentation have been well defined for a number of years (see, e.g., Harrington, this volume). It is also possible to identify periods of relative stability or articulatory landmarks using articulatory monitoring techniques. The procedure remains more or less the same in most instrumental studies of speech timing and duration: the durations of consonant, vowel, or subsegmental components like transitions or regions of relative stability, or larger units like the mora or syllable-like speech intervals, are acoustically analyzed, or the spatiotemporal behavior of different articulators is monitored. The classic acoustic production and perception studies of vowel and consonant duration of the fifties and early sixties (e.g., Fischer-Jørgensen, 1954; Fry, 1955, 1958; Peterson & Lehiste, 1960; House, 1961; Delattre, 1962, among others) formed the groundwork for many of the experimental studies of segmental duration and prosodic timing factors like stress or vowel quantity that were to follow. Similarly, the early instrumental investigations of English speech rhythm by Classe (1939), and models of cross-linguistic rhythm proposed by Pike (1946) then Abercrombie (1964) gave rise to a generation of work on the phonetic correlates of linguistic rhythm (summarized in section 3).

Nooteboom (1972) rightly claimed that the principal problem with research on the temporal organization of speech up until the early seventies had been the tendency to try and isolate the various factors that influence segment duration. Researchers were primarily interested in discovering the intrinsic or inherent durations of vowel and consonant segments that were the source of phonological opposition between words. During the seventies and the eighties, there was increased awareness of the interactive nature of speech timing and the influence of multiple variables on segment duration. Many (although not all) of the pioneering studies of this time were directed towards modeling segment duration for text-to-speech synthesis and speech technology applications. The summaries of Klatt (1976) and van Santen (1992a) list a range of factors that influence the durational patterns of an utterance. These include vowel duration differences due to vowel height (i.e., close vowels tend to be shorter than open vowels), vowel lengthening before voiced obstruents, and durational contrasts between phonologically short and long vowels (i.e., vowel quantity contrasts). Syllable structure can also affect segment duration (e.g., segments tend to be shorter in complex versus simple syllables) and the length of a word can influence syllable duration (i.e., medial syllables tend to be shorter in polysyllabic words). All things being equal, segments or syllables at the beginning or end of a prosodic phrase or larger discourse segment are longer than medial syllables, speeding up and slowing down tempo tends to shorten and lengthen syllables respectively, and finally, speaking style in general can also influence speech segment duration.

Many of these factors will be examined in detail in the sections that follow (2–4) but it is useful to mention here some of the most influential pioneering studies that explored the role of factors like stress or accent, vowel quantity, word and phrase position, on segment or syllable duration These include Barnwell (1971), Lehiste (1970, 1972, 1973, 1975, 1977), Lehiste et al. (1976), Allen (1972a, 1972b, 1975), Huggins (1972, 1975), Haggard (1973), Oller (1973, 1979), Klatt (1975, 1976), Umeda (1975, 1977), Port et al. (1980), Port (1981), Cooper and Danly (1981), Crystal and House (1982, 1988a, 1988b, 1990) for English; Lindblom (1963), Elert (1964), Lindblom and Rapp (1973), Bruce (1977) for Swedish; Nooteboom (1972, 1973), Nooteboom and Slis (1972) for Dutch; Kohler (1983, 1986) for German; Delattre (1966, 1969) for comparisons of English, French, German, and Spanish; Fujisaki and Sudo (1971) for Japanese. Some earlier studies also modeled the perceptual significance of segment or syllable duration patterns. Klatt and Cooper (1975) found there is a "just noticeable difference" or JND of 25 ms for speech segment duration, suggesting that listeners are not sensitive to durational adjustments below this level. Klatt (1976, p. 1219) discussed this in terms of Weber's law and proposes that "only changes of about 20% or more may serve as primary perceptual cues" to signaling differences in segmental duration. Listeners may also be more sensitive to vowel duration adjustments compared to consonant adjustments (e.g., Huggins, 1972), and position on word and other factors can influence listeners' percepts, although there appears to be some consensus that within an interval duration range of 30–300 ms, JNDs vary from 10–40 ms (see Nooteboom, 1973; Klatt, 1976; Bochner et al., 1988; and Kato et al., 1998 for summaries of this research).

One of the most sophisticated and influential interactive segment duration models that emerged in the seventies is the rule system developed by Klatt (1979) for American English. This model encompasses a wide range of durational factors or "rules" that govern segment duration from the discourse level to the physiological (i.e., whereby articulatory constraints relating to the production of a consonant or vowel segment influence the level of shortening or *incompressibility* it can sustain to maintain its "segment" identity). Notions of "inherent" or "intrinsic" duration of consonant and vowels, that were evident in earlier work (e.g., Peterson & Lehiste, 1960; House, 1961) are also central to this model, although Klatt (1976) was also well aware of the acoustic, articulatory, and perceptual difficulties in defining the acoustic durational properties of individual segments. The principles of incompressibility or, conversely, expandability or *elasticity* in response to the many interacting factors that can influence segment or syllable duration have since become integral components of duration algorithms for speech synthesis applications (see also Campbell & Isard, 1991; van Santen & Shih, 2000 for a discussion of this), and Klatt's model of segment duration for American English has served as a template for segment duration algorithms in other languages, like French (e.g., O'Shaugnessy, 1984), and Swedish (Carlson & Granström, 1986). There are now a large number of complete segment and syllable duration models for various languages that have built on the pioneering work of the fifties, sixties, and seventies. Many duration models have also been developed for speech technology applications and use statistical methods, clustering analysis techniques, or neural networks to model segment or syllable durations based on large databases of connected speech. This is by no means an exhaustive list, but some representative studies include Coker et al. (1973), van Santen and Olive (1990), Campbell and Isard (1991), Riley (1992), van Santen (1992a, 1992b), Fletcher and McVeigh (1993), Greenberg et al. (2003) for English; Fujisaki et al. (1975), Carlson and Granström, (1986), Fant and Kruckenberg (1989), Takeda et al. (1989), Fant et al. (1991), Campbell (1992a, 1992b), Sagisaka (1992), Kaiki and Sagisaka (1992), for Japanese; O'Shaugnessy (1981, 1984), Bartkova (1991), Pasdeloup (1992), Zellner (1996) for French; Möbius and van Santen (1996) for German; Shih and Ao (1997), van Santen and Shih (2000) for Mandarin. For more detailed descriptions of these computational duration models, the reader should see relevant chapters in volumes edited by Bailly et al. (1992), van Santen et al. (1996), Sagisaka et al. (1997). The overview presented in Klatt (1987) is also recommended to the reader, and van Santen (1992b) and van Santen and Shih (2000) are also particularly insightful with regard to the relative virtues of models based on syllable-sized units or segment intervals.

As noted by Beckman (1997), many of the speech synthesis algorithms developed since the seventies can produce an accurate representation of individual segment or syllable duration for an utterance, but there are still timing effects that cannot be modeled via stretching or compressing whole segment or syllable intervals based on acoustic duration measurements alone. Not all types of lengthening effects (i.e., slowing down tempo, increasing stress or prosodic prominence, observed lengthening of a vowel due to voicing of a following obstruent) are produced in the same way in all components of a syllable, and there is now a large body of research devoted to examining the fine-grained phonetic detail that goes into constructing prosodic patterns of an utterance (see sections 2.1 and 4.3 below). The time course of a naturally produced utterance consists of more than the sum of its segmental parts. Kozhevnikov and Chistovich (1965) proposed that to understand speech timing, we need to think about relative temporal patterning within larger speech units, and not just about raw durations of segments and syllables. A large body of fine experimental research on the temporal organization of speech in the seventies was inspired by this work, often in the quest for evidence of underlying (but not necessarily surface) temporal invariance (see for example Nooteboom & Slis, 1969; Lindblom & Rapp, 1973; Kohler, 1986). Similarly, the early instrumental investigations of English speech rhythm by Classe (1939), and models of cross-linguistic rhythm proposed by Pike (1946), then Abercrombie (1964), gave rise to a generation of work on the phonetic correlates of linguistic rhythm. These influences can still be seen to a certain extent in work by "syllable-centricists" (e.g., Campbell & Isard, 1991; Greenberg et al., 2003), and is also evident in the renewed interest in developing new metrics to measure linguistic rhythm since the late nineties (see section 3).

The evolution of theories of speech production and speech perception research will also continue to influence what we measure in an attempt to try and uncover the temporal organization of speech, while taking into account the high level of inter-speaker and intra-speaker variability in speech. Moreover, many recent studies of speech communication, including speech timing studies, have been influenced by Lindblom's H&H ("Hyper-articulation and Hypo-articulation") theory (e.g., Lindblom, 1990), which is based on the premise that speakers modify their articulatory strategies in response to the communicative needs of a particular speech situation, and temporal variability has to be taken into account in the modeling process (see Harrington, this volume). Lindblom (1983) describes the abilities of speakers to modify their articulatory patterns in response to the particular communicative situation in natural speech in the following way:

> In normal speech the production system is rarely driven to its limits. Typically we speak at a comfortable volume or rate, and we use a degree of articulatory precision that seems "natural". On the other hand we are of course occasionally perfectly

> capable of hyperarticulating and (hypo-articulating) that is of adjusting the loudness, tempo, clarity etc. of our speech to the needs of the situation, thereby exploiting more of the full range of phonetic possibilities. (Lindblom, 1983, p. 219)

Phonological theory has also provided alternative frameworks to feature-/segment-driven research. Proponents of Firthian prosodic analysis (e.g., Coleman, 1992; Local & Ogden, 1996) have long questioned the usefulness of thinking about speech timing in terms of conventional successive segment-like units. The intersection of articulatory phonology (e.g., Browman & Goldstein 1990, 1993), and dynamical theories of speech production (e.g., Saltzman & Munhall, 1989) has also provided an alternative framework within which to investigate the spatio-temporal signatures of prosodic structure that can either go beyond, or can occur within, conventional linguistically defined unit boundaries. Nevertheless, in the sections that follow it will become apparent that a great deal of recent research on prosody and timing is still based on measurement of acoustic intervals corresponding to phoneme-like or syllable units.

1.2 Prosody as structure

The two uses of the term prosody outlined in section 1, as either *suprasegmentals* or abstract hierarchical phonological structure, are not completely unrelated because phonetic parameters like f_0, intensity and duration (the "classic" suprasegmental parameters according to most phonetics textbooks), and their perceptual correlates pitch, loudness, and length contribute to the signaling of different aspects of prosodic structure. While one of the many goals of phonetic science has been to try and quantify phonetic differences among languages by comparing vowel inventories, voice-onset-time (VOT) characteristics, and so forth, it is perhaps less straightforward to compare the phonetics of prosody on the basis of phonetic features like f_0 or duration without an understanding of prosodic structure and how these structures are realized in languages at all levels (see Beckman, 1992; and Jun, 2005 for further discussion of this). A great deal of research on speech timing and prosody since the nineties has been conducted within or around the framework of prosodic phonology, along with metrical phonology (Liberman & Prince, 1977; Selkirk, 1984; Nespor & Vogel, 1986; Hayes, 1995). While it is also true that the conventional view of prosody as a collection of suprasegmental *parameters* is still apparent in many areas of speech timing research, it is hard to dispute the influence of developments in phonological theory on a new generation of speech scientists. Many of the earlier influential studies of prosodic timing also assume a hierarchical structure in terms of speech timing patterns that are governed by syntax or rhythmic factors (e.g., Nooteboom, 1972, Lehiste, 1973, 1977; Lindblom & Rapp, 1973; Klatt, 1975; Kohler, 1986), although the view of *formal* prosodic structure as independent from syntax did not really emerge until the late seventies and eighties.

Figure 15.1 shows a range of phonological models of prosodic structure (reproduced from a detailed summary of prosodic models by Shattuck-Hufnagel &

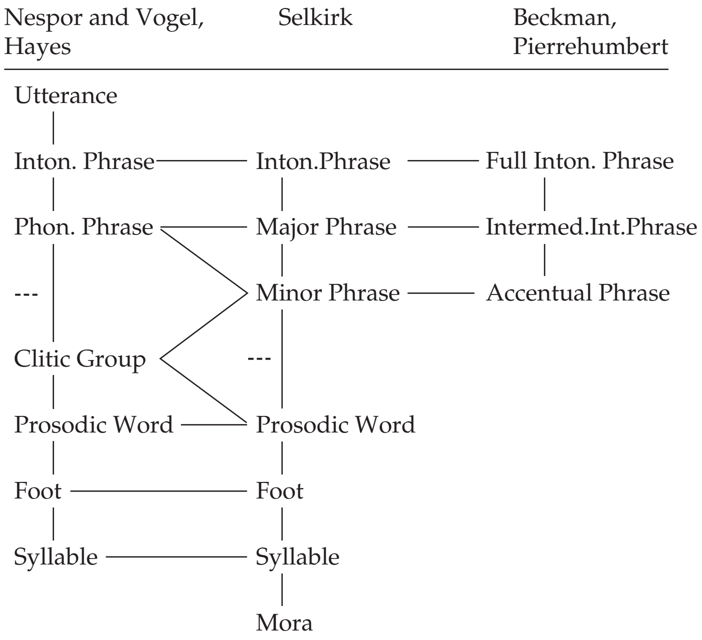


Figure 15.1 A schematic illustration of different proposals for prosodic structure, summarized in S. Shattuck-Hufnagel and A. E. Turk (1996), A prosody tutorial for investigators of auditory sentence processing, *Journal of Psycholinguistic Research*, 25, 193–247, p. 206. (Reproduced by permission of Springer)

Turk, 1996, also see Beckman & Venditti, this volume) that have influenced a generation of research from the eighties on. For example at the lowest level, in Figure 15.1, the *mora* is a subconstituent of the *syllable* and dominates either a vowel segment or segments in the syllable rhyme (see, e.g., Hayes, 1989). Each successive level represents a grouping that is nested below the constituent immediately above. Depending on the particular model, the highest constituent or grouping of the hierarchy corresponds to the *utterance*, then the full *intonational phrase* (e.g., Nespor & Vogel, 1986). In other models, the full intonational phrase is the highest level, which in turn dominates the major phrase or intermediate intonational phrase and so on down the hierarchy (e.g., Beckman & Pierrehumbert, 1986). Each unit or grouping is associated with a range of structural features and phonetic signatures, including temporal patterns that help mark out differences in hierarchical structure. For example, Japanese is often described as a moraic language, because the mora is considered by many to be an important structural and "timing" unit in the language (see section 3). It is also useful to think in terms of broad functions like *head* or *edge* marking within different levels of prosodic structure (after Beckman, 1992, 1996). This view is strongly influenced by a discussion of word-level prosody as either *culminative* (e.g., word stress), or *delimitative*, whereby prosody contributes to the marking of boundaries between phonological words (after Trubetzkoy, 1939, cited and discussed in Beckman, 1986, pp. 18–27). A range of durational patterns contribute to the marking of these two functions. In section 2, many of these "temporal signatures" of prosody (after Byrd, 2000), will be summarized with particular focus on lengthenings and shortenings that are potentially constrained by prominence or stress relations and phrasing (see also Keating, 2006, for a good overview of research in this area). As mentioned previously, not all research on the durational patterning of spoken language has been conducted within this kind of theoretical framework, and the format of the following sections is not intended to exclude any relevant findings that shed light on the topic at hand. It is not within the scope of this overview to address the central question of timing control in speech production (see relevant chapters in this volume), nor questions relating to the temporal integration of information in speech perception (readers interested in this should see the special issue of the *Journal of Phonetics*, vol. 31 (2003), and Pisoni & Remez, 2005).

2 Lengthenings and Shortenings: The Temporal Signatures of Prosody

2.1 Stress and accent

Model(s) of the prosodic hierarchy shown in Figure 15.1 suggest a multiple layered structure, and *stress* or *prominence* effects operate at different levels of this structure in many languages. In English, for example, a stressed syllable is the head of a stress foot and is identified by its quantity features at the lowest levels (i.e., it is a strong syllable that shows no vowel reduction), and at the highest levels it can be the potential location for nuclear accent (sentence stress or phrasal stress). The traditional definition of a stress foot in English is a stressed syllable plus any following unstressed syllables regardless of word boundaries (e.g., Abercrombie, 1967). In other languages the stressed or prominent syllable can be at the right edge of the foot (e.g., as in French). Another more formal definition of the foot is based on alternating strong (stressed/prominent) and weak (unstressed/non-prominent syllables), and these units do not cross word boundaries (e.g., Selkirk, 1984). Syllables can be unstressed, stressed with a full vowel (i.e., head of a stress foot), and primary stressed (i.e., head of a prosodic or phonological word) (after Selkirk, 1984; Beckman, 1986; Hayes, 1995; and Gussenhoven, 2004). Primary stressed syllables may also be marked out as the location for nuclear accent by virtue of the association of "particular tones in the intonation structure" (Beckman & Edwards, 1992, p. 374). Four levels of "prominence" in English are also proposed by Vanderslice and Ladefoged (1972) who distinguish unstressed reduced syllables, stressed syllables with full vowels, accented syllables (although accented here referred to prominent due to greater respiratory energy and articulatory force, for example), and intonationally prominent syllables due to nuclear accent. Later models of prominence propose a more integrated intonational approach, whereby accent or accentual prominence is interpreted as intonational pitch accent (see Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986; Shattuck-Hufnagel & Turk, 1996; and references in Beckman & Venditti, this volume).

Early experimental investigations of stress and accentual prominence in English and other languages like Swedish and Dutch note lengthening of syllables and their constituent segments due to presence of stress (e.g., Parmenter & Trevino, 1936; Fry, 1955; Peterson & Lehiste, 1960; House, 1961; Lehiste, 1970; Nooteboom, 1972; Lindblom & Rapp, 1973; Klatt, 1975; Umeda, 1975; Nakatani et al., 1981; Crystal & House, 1988a, 1988b, 1990; and see also Cutler, 2005 for a review of research on stress production and perception). Most of the earlier studies showed that the strongest lengthening effects are in the vowel portion of syllables. Stressing a vowel can add anything from 30 ms to more than 70 ms depending on the degree of stress (e.g., Crystal & House, 1988b). However, it has been observed many times that primary stress in some of the older studies is often also "phrasal stress" or nuclear accent (e.g., Beckman & Edwards, 1994; Turk & Shattuck-Hufnagel, 2000). In traditional experimental studies of languages like English or Dutch, words spoken in isolation, or experimental tokens inserted in carrier phrases would have carried nuclear accent, and studies based on read prose or spontaneous discourse tended to collapse primary stress and postlexical pitch accent into a single category "stressed."

Since the eighties, many studies have been at pains to separate lexical stress and nuclear accent when examining the durational patterns in a range of languages including different varieties of English (Summers, 1987; Beckman et al., 1992; Beckman and Edwards, 1992; Fletcher and McVeigh, 1993; de Jong, 1995, 2004; Turk and White, 1999; Harrington et al., 2000; Greenberg et al., 2003), Dutch (e.g., Eefting, 1991; Sluijter, 1995; Sluijter & van Heuven, 1996; Gussenhoven, 2009; Rietveld et al., 2004; and Cambier-Langveld & Turk, 1999, for Dutch and English), Swedish (Lindblom & Rapp, 1973; Bruce, 1977; Fant et al., 1991; Heldner & Strangert, 2001), Finnish (Suomi et al., 2003; Suomi, 2007), Arabic (Roach, 1982; Zawaydeh & de Jong, 1999; de Jong & Zawaydeh, 2002; Chahal, 2001, 2003), German (Kohler, 1983; Kleber & Klipphahn, 2006; Barry et al., 2007; Baumann et al., 2007), Greek (e.g., Fourakis, 1991; Fourakis et al., 1999; Arvaniti, 1992, 1994), Italian (e.g., d'Imperio & Rosenthall, 1999), and European Portuguese (Frota & Vigário, 2001). Most agree that, all things being equal, intonationally accented syllables are longer than unaccented stressed and unstressed syllables. For example, Turk and Sawusch (1997) found that that accented vowels are 21 percent longer than unaccented (stressed) vowels (206 ms vs. 163 ms) in American English. There is also some consensus that at the lower levels of the prosodic hierarchy for some of these languages, stressed syllables (that are not accented) are also longer than unstressed syllables (e.g., for English, Crystal & House, 1988b; Beckman & Edwards, 1994; de Jong, 2004; for Dutch, Rietveld et al., 2004). For example de Jong (2004) showed that unstressed vowels are less than half the duration of stressed vowels in unaccented contexts (with differences of between 70 and 100 ms). Unstressed vowels are more than 50 percent shorter than stressed unaccented syllables in Swedish (Heldner & Strangert, 2001). Interestingly for Dutch, Rietveld et al. (2004) also reported differences between unstressed and stressed syllables in "weak stress feet" (where the head syllable bears secondary stress), noting differences of around 23 ms, as well as differences between unstressed and stressed syllables in strong feet (where the head syllable of the foot is primary stressed), noting lengthening of up to 54 ms depending on the vowel. In less constrained forms of spontaneous discourse as opposed to more controlled "laboratory-style" speech, stress-related lengthening can be somewhat reduced (e.g., Umeda, 1975), although Greenberg et al. (2003) in their study of the SWITCHBOARD corpus of recorded dialogues in American English (Godfrey et al., 1992), observed that accented syllables are from 60 to 100 percent longer than unstressed syllables.

In many of the above studies, different types of accentual lengthening relating to focus conditions are also taken into account. For example, accented words in narrow or broad focus domains are compared, as well as different pragmatic focus conditions (contrastive or noncontrastive nuclear accent). In an earlier set of experiments on American English, Cooper et al. (1985) found that emphasized words are 37–41 percent longer than unemphasized words. Baumann et al. (2007) compared various durational properties of German utterances like "Marlene will eine Banane schälen" (trans. 'Marlene will peel a banana') that were responses to a series of questions which elicited either narrow or broad focus, or contrastive or noncontrastive accent (see Ladd, 1996; Beckman & Venditti, this volume). They found that the nuclear accented syllable of "Ba*na*ne" is longer when it receives contrastive accent (i.e., when the utterance is a response to the question "Does Marlene want to peel a *potato*?") than when it occurs in a narrow focus domain (i.e., when the utterance is a response to "*What* does Marlene want to peel?") or a broad focus domain (i.e., when the utterance is a response to "What's new?"). Heldner and Strangert (2001, p. 339) also showed that stressed syllables in Swedish lengthen by around 30 percent under conditions of focal accent. More broadly, Lehiste (1975) found that position in the discourse often influences the durational patterns of words. Umeda (1975) reported that semantic novelty can also be associated with durational lengthening of segments in words that are newly introduced into the discourse. Subsequent experiments by Fowler and Housum (1987) also suggest that information value has durational consequences for connected speech, with shortening in words associated with old versus new information. Eefting (1991) examined accentuation and information value in Dutch words of one and three syllables' duration and found that accentuation was the main source of lengthening in the target word, followed by accentuation plus information value, but information value alone was not sufficient to account for word-level timing patterns.

Not all languages show equal degrees of stress- or accent-related lengthening, nor do they have a three-way duration contrast between unstressed, stressed unaccented, or stressed accented syllables. Early comparative studies of stress in English and Spanish (Delattre, 1966; Oller, 1979; Dauer, 1983; Hoequist, 1983a) showed stress to be less durationally marked in Spanish than in English, although den Os (1988) noted similar degrees of stress-related lengthening in Italian and Dutch. Fant et al. (1991) observed smaller durational differences between unaccented and accented syllables in French than between stressed and unstressed syllables in Swedish. They found stress-related lengthening of between 100 and 150 ms in non-prepausal syllables (see section 2.2) in Swedish and English, but French accented syllables were only about 50 ms longer than unaccented syllables. Similarly Polish shows smaller effects of stress-related lengthening than other languages (e.g., Jassem, 1959; Dogil, 1999), although durational lengthening is observed when lexically stressed syllables are also intonationally accented (e.g., Oliver & Grice, 2003). Strong stress-related lengthening has been noted in Arabic (of around 25 percent) without any additional effect of contrastive nuclear accent or focus (e.g., de Jong & Zawaydeh, 2002).

Many studies of stress and prominence in English have also examined additional articulatory or acoustic correlates of prominence, e.g., magnitude and velocity of opening and closing articulatory gestures, vowel formant patterns, spectral tilt, vowel intensity, pitch height or pitch change (e.g., Fry, 1955; Lehiste & Peterson, 1959; Delattre, 1966; Kent & Netsell, 1971; Stone, 1981; Dauer, 1983; Beckman, 1986; Beckman & Edwards, 1994; de Jong, 1995; Erickson, 1998; Cho, 2005, 2006; Kochanski et al., 2005; see also Cutler, 2005, for an additional review of the earlier literature). Languages do not necessarily use the same combinations of phonetic features to signal stress, or at least the strength of each phonetic correlate that contributes to the signaling of stress and accentuation is not identical across languages. For example, stress in Dutch and American English is associated with spectral tilt differences as well as with longer acoustic duration and full vowels (e.g., Sluijter, 1995; Sluijter & van Heuven, 1996; Campbell & Beckman, 1997), although overall vowel intensity is more a feature of nuclear accent rather than lexical stress in English. Vowels are also more acoustically peripheral, or have more "extreme" or bigger articulatory gestures at the highest level of prosodic prominence (usually contrastive nuclear stress or accent) in many languages including German, English, and Dutch (Koopmans-van Beinum, 1980; Stone, 1981; Kelso et al., 1985; Ostry & Munhall, 1985; Summers, 1987; Beckman & Edwards, 1994; de Jong, 1995; Erickson, 1998; Harrington et al., 2000; Mooshammer & Fuchs, 2002; Cho, 2005, 2006; Hay et al. 2006; Baumann et al., 2007). Nuclear accented vowels in English are also more resistant to trans-consonantal coarticulatory effects of surrounding vowels (Cho, 2004), and articulatory gestures associated with nuclear accented syllables show generally fewer coarticulatory effects due to segmental environment (Öhman, 1967; de Jong et al., 1993). Lindblom et al. (2007) examined locus equations which are often used as an acoustic measure of the degree of coarticulation between consonants and vowels in CV sequences (see Harrington, this volume). They found small but consistent effects of "emphatic stress" on intervocalic consonants in V1.ˈCV2 sequences in American English, with the inter-vocalic velar and alveolar consonants showing longer, "deeper" stop closures under conditions of emphasis, influencing the degree of consonant–vowel coarticulation.

Studies have also explored the link between stress, vowel target undershoot, and vowel duration, inspired directly or indirectly by Lindblom's (1963) hypothesis that the shorter the vowel, the stronger the assimilatory effects of neighboring consonants on vowel formants (e.g., Delattre, 1969; Gay, 1978; Harris, 1978; Tuller et al., 1982; Summers, 1987; Fourakis, 1991; Moon & Lindblom, 1994; Padgett & Tabain, 2005; Nowak, 2006; and also see the discussion of vowel reduction in Harrington, this volume, and section 4.3 below). Moreover the extent of vowel reduction in unstressed or unaccented syllables varies among languages. Unstressed vowels are more centralized in English than in Dutch, although they exhibit similar durational patterns (van Bergem, 1993; Sluijter & van Heuven, 1996) whereas in other languages with lexical stress like Arabic, unstressed vowels are not as reduced or short as unstressed vowels in English (e.g., de Jong & Zawaydeh, 2002; Zuraiq & Sereno, 2007). Interestingly, English listeners tend to group unreduced vowels with *stressed* rather than *reduced unstressed* vowels, and the perceived strong/weak syllable distinction has more to do with spectral features and intensity of vowels, rather than acoustic vowel duration (Fear et al., 1995). Polish reportedly does not reduce vowels in unstressed position (Jassem, 1959), although some studies have found evidence of vowel formant undershoot and spectral tilt differences in unstressed syllables (e.g., Crosswhite, 2003a; Nowak, 2006). Pitch level and pitch movement are reportedly the most salient correlates of stress and accentuation in Polish, and not acoustic duration (e.g., Dogil, 1999). Ortega-Llebario and Prieto (2005) noted spectral balance and vowel quality differences between stressed and unstressed vowels, regardless of whether the stressed syllables are also nuclear accented, although durational differences are quite small (around 15 ms and well below the JND levels described by Klatt & Cooper, 1975, for stress in English at least). Stress at the lowest levels in Catalan is also associated with duration, vowel quality, and spectral tilt differences, although vowels do not neutralize to schwa in unstressed position, but still show more centralization than in Spanish, for example, and duration differences between stressed and unstressed syllables are of the order of around 35 ms, even in the absence of postlexical pitch accent (Astruc & Prieto, 2006).

A range of other segmental and syllabic duration effects are also particularly evident in prosodically prominent syllables. For example, VOT is longer in stressed and/or accented syllables in many languages (e.g., Lisker & Abramson, 1967; Keating, 1984; Cho & McQueen, 2005; Cole et al., 2007), suggesting that segments are hyper-articulated as well as generally longer in stressed and accented contexts (after de Jong, 1995, and see below for further discussion). Vowel quantity contrasts (i.e., phonological contrasts between short and long vowels) are also most evident in primary stressed or accentually prominent syllables in a wide variety of languages including Finnish (Engstrand & Krull, 1994; Suomi, 2007), Swedish (Elert, 1964; Lindblom & Rapp, 1973; Engstrand & Krull, 1994; Heldner & Strangert, 2001), German (Kohler, 1983; Jessen, 1993; Barry et al., 2007), Estonian (Lehiste, 1970; Engstrand & Krull, 1994; Traunmüller & Krull, 2003), Aleut (Taff et al., 2001), Chickasaw (Gordon, 2005), Tamil (Keane, 2006), Dutch (Nooteboom, 1973 and references therein; Rietveld et al., 2004), Arabic (de Jong & Zawaydeh, 2002). Estonian is an interesting case as it has a three way vowel quantity; Krull et al. (2006, p. 81) summarize Estonian vowel quantity as follows: "in a disyllabic word of the form $C_1V_1C_2V_2$ . . . V1 as well as C2, both singly and as a VC unit, can have three degrees of quantity: short, long, and overlong." Stress and accentuation contribute to the signaling of this contrast. The longer duration of the accented initial syllable is crucial as well as the durational ratio between the first and second syllables which is highly stable in spontaneous speech (e.g., Engstrand & Krull, 1994). Intrinsic or inherent vowel duration differences (e.g., due to vowel height), and the effects of postvocalic voicing are also most evident in primary stressed and accented syllables in a number of languages (for Dutch Nooteboom, 1972; for French Benguerel, 1971; Fletcher, 1991; for Swedish Lindblom & Rapp, 1973; for American English Klatt, 1975; Port et al., 1980; Port, 1981; de Jong, 1991, 2004) although many of these contrasts are lost or reduced in unstressed, word medial contexts (e.g., Klatt, 1975; Umeda, 1975, 1977; Crystal & House, 1988b; Rietveld et al., 2004).

Languages that have been traditionally described as having different lexical prosody from English or West Germanic languages, for example, show mixed effects of durational lengthening due to accentuation and prominence. In Japanese (which does not have lexical stress), certain words have a lexical pitch accent (see Beckman & Venditti, this volume). Pitch rather than duration is the main cue to accentual prominence (Beckman, 1986), although Hirata (2004) also noted small degrees of lengthening (around 12 percent) between unaccented and accented short and long vowels (Japanese has a phonological contrast between long and short vowels). Languages with so-called mixed prosody, i.e., that have lexical stress and lexical pitch accent like Serbo-Croatian, also show stress-related lengthening in syllables, as does Swedish (e.g., Lindblom & Rapp, 1973; Strangert, 1985; Fant et al., 1991; Heldner & Strangert, 2001; see also Bruce, 2005 for a general discussion of pitch accent variation among different varieties of Swedish). Mandarin and Thai, traditionally described as "lexical tone" languages, also have lexical stress which is cued by syllable lengthening (e.g., Potisuk et al., 1996; van Santen & Shih, 2000). In Mandarin, lengthening effects are most evident when words are pragmatically focused, although in longer words the final syllable shows the most lengthening (Chen, 2006). Results like these suggest that Mandarin actually shows similar kinds of durational patterns to English but is quite different from a restricted tone or "pitch accent" language like Japanese (see Beckman & Venditti, this volume, for a discussion of traditional pitch-based lexical prosodic typology). In less well studied languages like the Austronesian language Ma'ta (Remijsen, 2002), which has lexical tone and word stress, duration also serves as a cue to stress (as well as spectral balance and vowel quality). Curaçao Papiamentu, also a tone language, similarly has longer stressed versus unstressed syllables (Remijsen & van Heuven, 2005). Experimental studies of lexical and postlexical prosody in the indigenous Australian languages Bininj Gun-wok and Warlpiri (Fletcher & Evans, 2002; Butcher & Harrington, 2003) suggest small durational effects due to accentual (focal) prominence in these languages (see also Bishop, 2002). We will discuss language-specific stress realization and its relationship to rhythm classification in section 3.

Some of the language-specific differences observed above are clearly related to different types of lexical and/or phrasal prominence. For example, accentual prominence in French is not *lexical* like stress in west Germanic languages (e.g., English, Dutch, or German), but is associated with a more boundary-marking or demarcative versus culminative function (Grammont, 1946; Delattre, 1966; Garde, 1968; Vaissière, 1974, 1991; Crompton, 1980; Léon & Martin, 1980; Martin, 1987; Touati, 1987; Di Cristo, 2000; Post, 2000; Jun & Fougéron, 2002; Welby, 2006). French presents an interesting case in that accentual lengthening essentially contributes to the marking of prosodic prominence, on the one hand, as well as the marking of *edges* of different prosodic constituents. While there is some disagreement on the best way to model prosody in French (see Welby, 2006, for a useful discussion), most models posit a minimal right-headed group of syllables that have been referred to as rhythmic units, rhythm groups, accentual phrases, phonological phrases, prosodic words, and intonational groups. There is consensus, however that the rightmost syllable of the group bears primary accent or prominence and is usually significantly longer than group-internal syllables (Rigault, 1962; Delattre, 1966; Benguerel, 1971; Crompton, 1980; Léon & Martin, 1980; Rossi et al., 1981; Wenk & Wioland, 1982; Fletcher, 1991; Fant et al., 1991; Jun & Fougéron, 2002). This type of accentual prominence is not the same as secondary accent that that can occur word-initially and is not generally associated with lengthening (Pasdeloup, 1990; Di Cristo, 1999, 2000). There is also a larger prosodic unit or grouping akin to an intonational phrase in French (e.g., Crompton, 1980; Martin, 1987; Di Cristo, 2000; Jun & Fougéron, 2002) and final syllables associated with this unit exhibit greater lengthening than final syllables at the edge of smaller prosodic units like the accentual phrase.

2.1.1 The local and global domain(s) of accentual lengthening While vowels carry most of the lengthening effects associated with prosodic prominence, some of the earlier discussion suggests that accentual lengthening is not uniformly distributed through a syllable. This can also vary among languages. Fant et al. (1991) noted that lengthening in accented syllables in French is concentrated in the beginning of syllables, whereas in Swedish 75 percent of stress-related lengthening is located in the stressed VC (V:C or VC:) portion of a syllable. Heldner and Strangert (2001) compared the effects of focal and nonfocal accent in Swedish and showed that in focally accented CVC: syllables (i.e., where vowels are short and coda consonants are long), both the initial and final consonants lengthen, but the short vowel only lengthens minimally, whereas in CV:C syllables (i.e., with long vowels and a following short consonant), all segments are up to 31 percent longer under conditions of focal accent. Syllable-initial consonants and coda consonants also lengthen in English (see Crystal & House, 1988b, 1990; Turk & White, 1999; for a review of these effects), and in other languages including Dutch (e.g., Eefting, 1991; Cambier-Langeveld & Turk, 1999), and German (Barry et al., 2007). Conversely, Greenberg et al. (2003), in their study of the SWITCHBOARD corpus of recorded dialogues in American English (Godfrey et al., 1992), noted that degree of stress (and accentuation) has little impact on coda duration and the majority of stress-related lengthening is evident in the vowel nucleus. Accentual lengthening in Warlpiri is more evident in the rhyme portion of syllables than in the onset

(Butcher & Harrington, 2003), although lengthening effects are less extensive than in languages like Swedish or English (see section 3.3 for discussion of further implications of these differences).

Lengthening effects associated with accentual prominence can also extend beyond accented syllables to following syllables within a prosodic word in English and Dutch (see Sluijter & van Heuven, 1995; Cambier-Langeveld & Turk, 1999; Turk & White, 1999; Turk & Shattuck-Hufnagel, 2000; Cho & McQueen, 2005; Cho & Keating, 2007, for an additional comprehensive review of these effects). Baumann et al. (2007) also showed that in German lengthening of the nuclear foot (i.e., the stress foot where the nuclear accent is located) is greater under conditions of contrastive focus. Heldner and Strangert (2001) reported a similar effect in Swedish, although the effect is limited to the post-stress syllable in words that are accented due to pragmatic focus, whereas Turk and White (1999) showed that durational effects associated with accentual lengthening in English can extend across all three syllables of a word that carries initial primary stress and bears an intonational pitch accent. It was originally thought that the lengthening effect is asymmetrical in English, i.e., lengthening effects due to an accentuation are much more evident in syllables after than before the accented syllable (e.g., Turk & Sawusch, 1997). Subsequent studies have shown that there are small lengthening effects that are also evident in syllables preceding accented syllables (e.g., Cambier-Langveld & Turk, 1999; and Turk & White, 1999) and there is also evidence of articulatory modifications to consonants (including longer VOT) in an initial syllable of an accented word when the third syllable of the same prosodic word carries nuclear accent (e.g., Cho & Keating, 2007). Moreover, there is also a small rightward effect of accentual lengthening across a word boundary when the initial syllable of the following word is unstressed, suggesting that accentual effects can go beyond the prosodic word in English. Finnish also exhibits accentual lengthening effects that are not bounded by a stress foot in a polysyllabic word (Suomi, 2007). Lengthening can extend from the accented syllable across at least two following syllables, irrespective of presence or absence of secondary stress (i.e., a second foot) in the word, although secondary stressed syllables are longer than unaccented unstressed syllables.

2.1.2 Not all lengthenings are alike It has been suggested that different levels of prominence in American English are associated with different types of qualitative articulatory effects (e.g., Beckman & Edwards, 1994, p. 30). The extra duration of stressed syllables (which are also nuclear accented) is associated with larger and faster opening lip and jaw gestures compared to similar syllables that are stressed but unaccented. The process of making a stressed syllable even more prominent is called *sonority expansion*. Cho (2005, 2006) also reported that accented syllables, and in particular the opening lip gestures of /bi/ and /ba/ syllables, are associated with longer, faster, and bigger gestures across a variety of prosodic positional contexts that include intermediate phrase (major prosodic phrase or phonological phrase) initial and final, and intonational phrase initial and final positions. Other studies have also shown that the extra lengthening of nuclear accented syllables can be produced by later relative timing of the closing relative to the opening gesture in /bVb/ syllables where V is an open central or back vowel (e.g., Harrington et al., 1995; Beckman & Bretonnel Cohen, 2000). Lower tongue-body positions have been observed in accented syllables for low vowels, with nonlow front vowels having higher or more fronted tongue body and nonlow back vowels having a backer tongue body (e.g., Engstrand, 1988; de Jong, 1995; Cho, 2005). De Jong (1995) referred to this as *localized hyperarticulation*, after Lindblom's (1990) H&H model (see Harrington, this volume, and section 1.1 above). The speaker intends to produce a more peripheral vowel to differentiate it from any other vowel that might have occurred in the same position. Similar kinematic effects have been found for accented close vowels in Australian English (e.g., Harrington et al., 2000). Moreover, consonant articulation is also affected, with clear accent-related articulatory *strengthening*, including more extreme tongue movements (e.g., de Jong, 1995), or more extensive linguopalatal contact (e.g., Bombien et al., 2007; Cho & Keating, 2007). However, most articulatory studies of stress and accentuation show that there is a high degree of inter-speaker variability, and de Jong (1995) suggested that articulatory realization of prominence contrasts are subject-independent and that the "articulatory" goals are more abstract.

Lengthening due to different prosodic or phonological effects (e.g., stress vs. accent or stress, tempo variation, postvocalic voicing, contrastive vowel length) is not necessarily uniform in syllables or in the subcomponents of a syllable. For example, Summers (1987) found that lengthening strategies due to stress and postvocalic voicing are not the same in American English. While the measured vowel intervals in de-accented sequences /bab/ are almost identical to vowel intervals in accented /bap/ syllables, kinematic data for jaw movement show quite different articulatory strategies in both cases, with bigger opening and closing gestures in the accented syllables and different formant timing patterns compared to syllables exhibiting the postvocalic voicing contrast (see Figure 15.2). Stressed syllables (i.e., contrastive nuclear accent) showed the greatest lengthening effects of postvocalic voicing. The influence of voicing is apparent in jaw-raising gestures in the later portion of /bap/ versus /bab/ sequences, whereas stress affects the entire syllable.

Similarly, kinematic patterns associated with accentuation in American English can be qualitatively different from those associated with preboundary lengthening or initial strengthening (see section 2.2), or slowing down speaking tempo (Edwards et al., 1991; Beckman & Edwards, 1992, 1994; see also the discussion in section 4.3 below). While all three duration-influencing factors can result in longer jaw opening and closing gesture durations in closed syllables (e.g., /pɒp/), final lengthening and slowing down tempo are realized differently to accentuation. Longer gestures in phrase-final stressed syllables are the result of localized slowing down of gestures at the edge of the phrase (see also Byrd & Saltzman, 1998, and section 2.2 below). By contrast, a bigger gesture is also produced in phrase-final unstressed open syllables. Beckman and Edwards (1994) suggested that the kinematic differences (i.e., the differences in gestural amplitude and velocity) between unstressed and stressed unaccented syllables are more extreme

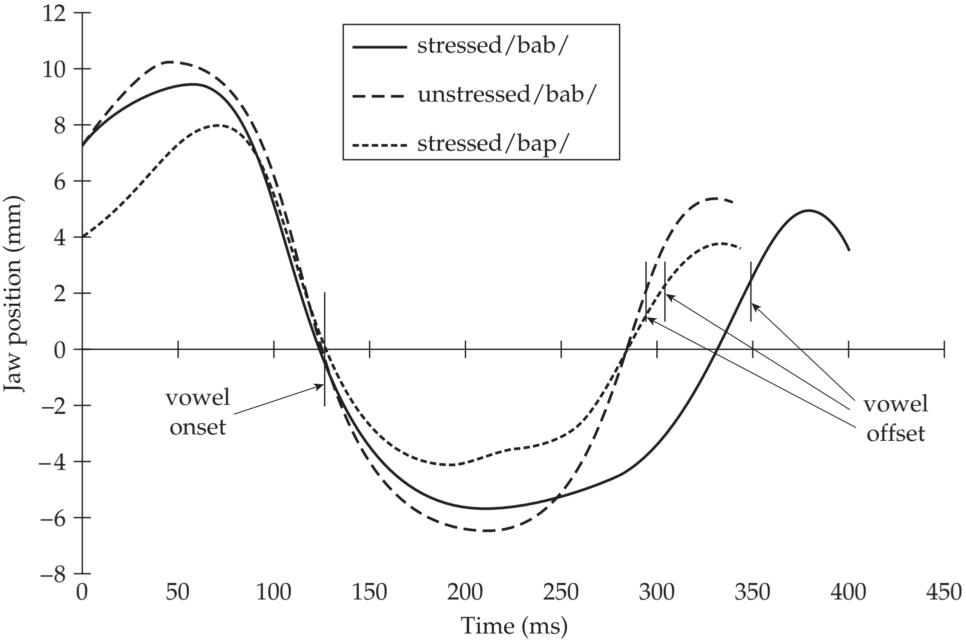


Figure 15.2 Jaw trajectories for stressed and unstressed /bab/, and stressed /bap/ syllables/ (from W. V. Summers, 1987, Effects of stress and final consonant voicing on vowel production: Articulatory and acoustic analysis, *Journal of the Acoustical Society of America*, 82, 847–63, p. 850). (Reproduced by permission of the American Institute of Physics)

than the differences between stressed unaccented and accented syllables, presumably due to the different types of "phonological content" largely to do with full versus reduced vowel quality that contribute to marking a syllable as stressed versus unstressed in English. With regard to other duration-based contrasts, lip and jaw kinematics in VC sequences contrasting short/long vowels in German suggest a localized slowing down of closing gestures to produce a longer vowel (Hertrich & Ackermann, 1997). Kroos et al. (1997) also found that short (lax) vowels in German are associated with tighter CV and VC "coupling" (or inter-gestural timing), compared to long tense vowels. By contrast, changes in inter-gestural phasing or timing do not account for reduced movement amplitude associated with de-accentuation according to Mooshammer and Fuchs (2002), which possibly suggests that other factors, like reduced articulatory force (e.g., Lindblom, 1983) or reduced intra-gestural stiffness (after Saltzman & Munhall, 1989), might account for the smaller observed amplitudes in de-accented syllables (see also Bombien et al., 2007).

Kinematic studies of lip and jaw articulation in languages with different prosodic structure like French also show that lengthenings are not necessarily produced in the same way. Recall that domain-final syllables in French are also accentually prominent. While accentual lengthening is associated with bigger and faster opening and closing gestures (e.g., Vatikiotis-Bateson & Kelso, 1993), accentual lengthening at the edge of the higher-level units (akin to the intonational phrase) is best modeled as a change in the timing relationship between opening and closing gestures (e.g., Fletcher & Vatikiotis-Bateson, 1994). Vowels in accented syllables that also bear contrastive focus are articulated with more extreme tongue gestures (e.g., Loevenbruck, 1999), or larger lip area compared to vowels in non-focal accented syllables (e.g., Ménard et al., 2006). Tabain (2003) and Tabain and Perrier (2005, 2007) also suggest there are clear articulatory and spectral, as well as durational, differences between vowels that are word final or word internal (i.e., unaccented) compared to accentual-phrase final vowels. They suggest that their results reveal a strategy of *temporal expansion* at prosodic phrase edges (which is the location for phrasal accent) rather than sonority expansion (e.g., Beckman & Edwards, 1994). Interestingly Edwards et al. (1991) also suggest that the articulatory kinematics of phrase-final syllables in English reveal a more targeted durational contrast, similar to Tabain and Perrier's (2007) notion of temporal expansion (see section 2.2) which tends to re-enforce earlier assumptions that the primary cue to accent in French is duration. Vatikiotis-Bateson and Kelso (1993) also show that in Japanese, opening and closing gestures in heavy syllables (two-mora) are longer and bigger than in light (single-mora) syllables, but the durational differences are somewhat smaller than those observed for English, although similar to those observed for non-phrase-final contrasts in French. They ascribe these differences to phonetic differences in the realization of accentual prominence across the three languages. In other words, not all lengthenings (or shortenings, for that matter), are produced in the same way by speakers, nor to the same extent in different languages.

2.2 *Durational marking of boundaries and medial shortening*

A syllable, and in particular the rhyme portion of a syllable (i.e., the vowel nucleus plus any coda consonant(s)) tends to be longer in intonational phrase-final and utterance-final position than when the same syllable is uttered in nonfinal or phrase-medial position. This durational phenomenon variously referred to as final lengthening, prepausal lengthening, domain-final lengthening, or preboundary lengthening, is considered by many to be universal, although the degree and extent of lengthening varies among languages (e.g., Delattre, 1966; Oller, 1973; Hoequist, 1983b; and references in Cambier-Langeveld, 1997). Preboundary lengthening has been observed in Russian (Zlatousova, 1975; Volskaya & Stepanova, 2004), most varieties of English (Oller, 1973; Lehiste, 1973; Lehiste et al. 1976; Klatt, 1975; Cooper & Danly, 1981; Nakatani et al., 1981; Edwards et al., 1991; Ladd & Campbell, 1991; Wightman et al., 1992; Fletcher & McVeigh, 1993; Turk & Shattuck-Hufnagel, 2007), French (Delattre, 1966; Benguerel, 1971; Crompton, 1980; Fletcher, 1991; Di Cristo & Hirst, 1997; Hirst, 1999; Jun & Fougéron, 2002; Tabain, 2003), Italian (Vayra & Fowler, 1992; d'Imperio & Gili-Favela, 2004; Avesani & Vayra, 2005; Hajek et al., 2007), Greek (Arvaniti, 1991), Czech (Dankovičová, 1997), Taiwanese (Peng, 1997), Iberian and Cuban Spanish (Delattre, 1966; Oller, 1979; Hoequist, 1983a), German (Delattre, 1966; Kohler, 1983; Kuzla et al., 2007), Dutch (e.g., Gussenhoven & Rietveld, 1992; Cambier-Langeveld, 2000), Swedish (Elert, 1964; Lindblom & Rapp, 1973; Lyberg, 1981; Fant et al., 1989; Horne et al., 1995), Finnish (Hakokari et al., 2007), Japanese (Hoequist, 1983a; Campbell, 1992a; Kaiki et al., 1992; Kaiki & Sagisaka, 1992; Venditti & van Santen, 1998), Arabic (Port et al., 1980; Chahal, 2003), Hebrew (Berkovits, 1994), Creek (Johnson & Martin, 2001), Mandarin (van Santen & Shih, 2000), Dalabon and Kayardild (Fletcher & Butcher, 2003), Warlpiri (Butcher & Harrington, 2003). Some of these studies also show that boundary-related lengthening may also be evident in the penultimate syllable in an intonational phrase, throughout a final foot or word, as well as in the final syllable. (See Lehiste, 1977; Scott, 1982; and Turk & Shattuck-Hufnagel, 2007, for English; Kohler, 1983; and Kuzla et al., 2007, for German; Berkovits, 1994, for Hebrew; Cambier-Langeveld, 2000, for Dutch; and Krull, 1997, for Estonian). Different levels of prosodic constituency may also interact with the degree of preboundary lengthening (e.g., Lindblom & Rapp, 1973; Wightman et al., 1992; Turk & Shattuck-Hufnagel, 2000). As with stress and accentuation, preboundary lengthening can also interact with a range of other durational factors including vowel quantity contrasts and speaker tempo variation (see section 4.3 below), and can manifest itself at the subsegmental and segmental levels. Various articulatory studies have shown that phrase-final syllables are associated with spatially larger and longer and, in some cases, slower closing gestures in closed syllables (e.g., Edwards et al., 1991; Fougéron & Keating, 1997; Byrd & Saltzman, 1998; Byrd et al., 2005, 2006; Cho, 2005, 2006; Tabain, 2003; Tabain & Perrier, 2005, 2007).

A range of proposals have been put forward over the years to account for preboundary lengthening. Early proposals reviewed in Oller (1979), and later rejected or modified, include the motoric planning theory of final lengthening which related it to a check-ahead mechanism that allows planning for following speech constituents. This is not unlike Lindblom's (1975) hypothesis that utterance durations reflect generative constraints geared to the size of the chunk of speech to be produced. He hypothesized that upcoming speech constituents are planned and stored in a hypothetical phrase buffer. Final lengthening reflects a general tendency to decelerate towards the end of a chunk because nothing else remains to be produced from the buffer. Lyberg (1979, 1981), by contrast, maintained that the final lengthening process is not necessarily governed by a "central" control factor as short-term memory-dependent explanations seem to suggest. He proposed that f_0 patterning relating to focus constituency and final lengthening are part of the same process in Swedish utterances. Others have suggested that final lengthening might be a listener-oriented strategy to signal different levels of constituency (e.g., Lindblom & Rapp, 1973; Klatt, 1975; Oller, 1979; Turk & Shattuck-Hufnagel, 2000). Lindblom (1968), after Öhman (1967), proposed that final lengthening is due to a relaxation of speech gestures, i.e., deceleration towards the end of the utterance. This view has also been subsequently supported by

Cooper and Sorensen (1977), Vayra and Fowler (1992), Berkovits (1994), and Tabain (2003), who describe final lengthening in terms of supra-laryngeal declination or "declension." Another view is that phrase-final lengthening in English is a *targeted* duration change which manifests as a localized slowing down of the syllable-final gesture in a phrase-final syllable (Beckman & Edwards, 1994; and Edwards et al., 1991). Later kinematic studies of preboundary lengthening draw a related conclusion, noting that the temporal spacing of word-final and word-initial gestures is adjusted when there is an intervening phrase boundary (e.g., Byrd, 2000; Byrd & Saltzman, 1998). Cho (2006, p. 540) has suggested that a combination of dynamical parameters including stiffness and inter-gestural timing can account for articulatory timing patterns at major phrase boundaries (see below).

There is little controversy these days that preboundary lengthening can be interpreted as a major perceptual cue to levels of linguistic structure in many languages along with other juncture-marking phenomena including intonational features (see also Beckman & Venditti this volume). It is perhaps here that we see the crucial role of prosodic phonology which has provided, as an alternative to syntactic models, a structural framework within which to investigate (or interpret) durational effects like preboundary lengthening (see Turk & Shattuck-Hufnagel, 2000, for a similar view). While, in the seventies, there was already some reluctance to formalize a link between syntax and final lengthening, statements such as "duration increases seem to have the primary purpose of marking syntactic units for the listener" (Klatt, 1975, p. 138) suggest that it was thought that syntactic structure can be at the very least concomitant with preboundary lengthening at the edges of larger prosodic units like intonational phrases, for example. Oller (1973, 1979) and Cooper and Paccia Cooper (1980) also clearly identified preboundary lengthening as syntactic, although Gee and Grosjean (1983) suggested it is governed by metrical rather than syntactic constraints. The general question of the prosody–syntax relationship is not straightforward, but it is more or less agreed that prosodic groupings can be influenced by syntax, although these influences vary depending on the language (see Jun, 2003, for a good summary). Moreover, a number of other studies at that time showed that final lengthening at clause boundaries is used by listeners to disambiguate syntactically ambiguous sentences in Swedish (Lindblom & Rapp, 1973), American and Southern British English (Huggins, 1975; Lehiste et al., 1976; Lehiste, 1977; Scott, 1982; Ferreira, 2000), and French (Martin, 1982). Preboundary lengthening in this context has been linked to the facilitative role of prosody in general auditory sentence processing in retrieving syntactic structure for the listener (e.g., Jun, 2003, and references therein).

2.2.1 Multiple levels of preboundary lengthening Most of the above-cited studies suggest that there are differential degrees of lengthening that help signal each prosodic level. In studies of preboundary lengthening in English, for example, three or four levels (depending on the study) are assumed to interact with segmental and syllable timing patterns – the prosodic word, the intermediate (phonological) phrase, the intonational phrase, and the utterance (e.g., for English, Wightman et al., 1992; also Ladd & Campbell, 1991; Beckman & Edwards, 1994;

Fougéron & Keating, 1997; Byrd & Saltzman, 2003; Byrd et al., 2005, 2006; Turk & Shattuck-Hufnagel, 2000, 2007). In Swedish, at least three levels have been investigated, including the prosodic word, prosodic phrase, prosodic utterance (e.g., Horne et al., 1995), and for French, three or more levels (Crompton, 1980; Martin, 1987; Di Cristo, 1999, 2000; Jun & Fougéron, 2002; Tabain, 2003; Tabain & Perrier, 2005). Most of these studies interpret differential preboundary lengthening effects as an indication of the position of a constituent in a structural prosodic hierarchy. There is mixed evidence that preboundary lengthening is greater at utterance edges than at intonational phrase edges. Klatt (1975) and Umeda (1975) found little evidence that the lengthening of a sentence-final syllable was greater than lengthening of "other phrase-final syllables in a connected discourse" in American English (referring here to syntactic phrase). Wightman et al. (1992) also found that preboundary lengthening of vowels in final syllables was no greater at constituent edges of break index values of 4 (i.e., full intonational phrases), 5 (intonation phrases followed by long silent pauses), and 6 (sentence boundaries). Preboundary lengthening may also be more pronounced at topic transition points in read speech and at the boundaries of large discourse segments (see Grosz & Hirschberg, 1992; Shattuck-Hufnagel & Turk, 1996; and Smith, 2004, for a useful discussion about implications for the relationship between discourse structure and prosodic structure).

Word-level boundary effects are evident in articulatory and acoustic studies of American English. Beckman and Edwards (1990) found that in phrase-medial positions there was word-final lengthening in the final /ə/ of "poppa" in "popp**a** posed" compared to the initial /ə/ of "pop **o**pposed." Likewise, the vowel monosyllable "pop" was longer than the vowel in the initial syllable of "poppa." The effects were consistent across accent conditions (i.e., with or without nuclear accent). Turk and Shattuck-Hufnagel (2000) also compared the durations of syllable and subsyllabic components of /tun/, /ə/, and /kwair/, in sequences like "tune a choir," "tuna choir," and "tune acquire," and concluded that durational differences are suggestive of word-level timing effects that are independent of higher-level constituents like prosodic phrase boundaries. They further proposed that the effects are amplified under conditions of pitch accent, and the boundaries between a content and a function word are weaker than between two content words, suggesting that morpho-syntactic structure plays a potential role in the micro-durational patterning of syllables.

In French, utterance-final vowels can be up to twice as long as word-medial vowels (e.g., Delattre, 1966; Crompton, 1980; Pasdeloup, 1990; Fletcher, 1991), with some studies reporting differences of more than 200 to 300 percent (e.g., Tabain, 2003). Longer vowels are produced at higher prosodic levels (i.e., accentual phrases and intonational phrases), and duration differences are often produced between vowels at accentual phrase boundaries compared to intonational or utterance boundaries. Articulatory gestures of greater magnitude are also observed in the final vowel of higher-level prosodic units compared to lower levels (e.g., Tabain & Perrier, 2005, 2007). Preboundary lengthening at accentual phrase edges is not found in Korean, although it contributes to the marking of intonational phrase edges (Jun, 1998). In Japanese, the final mora lengthens at the edge of phrases, although there is equivocal evidence for lengthening at accentual phrase edges compared to intonational phrase edges (e.g., Kaiki & Sagisaka, 1992). Ueyama (1996) compared vowel duration at the edge of accentual phrases, intermediate phrases, intonational phrases, and utterances and found that vowels are shorter at the edge of lower prosodic domains (accentual and intermediate phrases). Campbell (1992a) also reports on the tendency of sentence-final vowels to be shorter in Japanese (see also Kaiki & Sagisaka, 1992; and Kubozono, 2002). He claims that this is mainly due to the high frequency of unaccented particles like the past tense marker *-ta* in the corpus under investigation.

2.2.2 Preboundary lengthening, prominence and segmental context Strong interactions between lengthening due to accentual prominence, stress, and preboundary lengthening have been found in many acoustic durational studies of English (e.g., Klatt, 1975; Umeda, 1975; Oller, 1979; Nakatani et al., 1981; Crystal & House, 1990; Fletcher & McVeigh, 1993; Turk & Shattuck-Hufnagel, 2007, and references therein). Phrase-final syllables that are also nuclear accented show lengthening of the syllable rhyme (i.e., vowel nucleus plus any coda consonants) of up to 60 percent compared to similar syllables in phrase-medial position. In some influential early studies of languages other than English, Lindblom (1968) and Oller (1973) reported no significant difference in lengthening between utterance-final syllables that had either secondary stress or were unstressed in Swedish and English, respectively. Cooper and Danly (1981) set out to examine signs of potential interaction between vowel intensity, final voicing, and final lengthening in phrase-final and utterance-final positions. A significant effect was found for final voicing. Utterance-final monosyllabic words containing a final voiced consonant lengthened significantly more than words consisting of a final voiceless consonant. The effect was largely carried by differences in preceding vowel length. No significant effect was found for vowel intensity. Umeda (1975) noted that the longest syllables in an utterance (i.e., that are prepausal and/or primary stressed) are also where lengthening due to postvocalic voicing is most evident. Crystal and House (1988b) presented a similar although less consistent interaction between stress, vowel length, and the "lengthening-before-voicing effect in prepausal" contexts. Early studies also showed that languages that maintain vowel quality in unstressed or unaccented position, like Spanish or Japanese, do not show as much final lengthening as English, for example (e.g., Delattre, 1966; Hoequist, 1983a). The degree of preboundary lengthening (or other types of prosodic lengthening) may also depend on other phonological constraints, including vowel quantity contrasts, for example (e.g., Nooteboom, 1973; Umeda, 1975; Oller, 1979; Kohler, 1983; Beckman, 1986). In a vowel-quantity language like Creek (Johnson & Martin, 2001) the durational differences between long and short vowels are most evident in preboundary contexts, but interestingly, they are also acoustically centralized, in spite of being phonetically longer. Johnson and Martin (2001) discuss their results in relation to those of Nord (1986), who concluded that duration alone does not determine the "sharpness" of vowel quality in Swedish (unlike the earlier predictions of Lindblom's, 1963, undershoot theory discussed briefly in section 2.1). In other words, phonetically long vowels (in preboundary contexts) are not always the most acoustically peripheral. These results provide further support for the claim that not all prosodic "lengthenings" are alike, and language-specific factors have to be taken into account.

2.2.3 Measuring and modeling the domain or extent of preboundary lengthening

In the introduction to this section it was noted that lengthening effects in the vicinity of a phrase boundary have been observed predominantly in the rhyme of phrase-final syllables, although additional lengthening can occur in penultimate syllables or even the final foot. Terms such as progressive lengthening have also been used to describe the domain or extent of preboundary lengthening (see Kohler, 1983; Berkovits, 1994). Turk and Shattuck-Hufnagel (2007, pp. 445–6) summarized three kinds of models to account for the *domain* or *extent* of preboundary lengthening at major phrase breaks: Structure-based, Content-based, and Hybrid approaches. The first of these is largely determined or fixed by linguistic structure, i.e., final lengthening only affects a structurally similar region across all phrases (i.e., this might be a syllable rhyme or a final syllable, e.g., Klatt, 1975; Wightman et al., 1992). The content view, which might also be termed the "overlap view," suggests that the domain of lengthening is variable depending on the nature of the segments or syllables in the phrase-final region. The work of Byrd and colleagues (Byrd & Saltzman, 1998, 2003; Byrd, 2000; Byrd et al., 2005, 2006) is an example of this type of model. Byrd and Saltzman (2003) suggested that the local slowing down of gestures in the vicinity of major phrases can be modeled as changes to the activation time course of special prosodic or Pi-gestures that capture trans-gestural temporal patterns. Importantly, these prosodic gestures have no articulatory realization of their own, but they effectively yoke vocal tract variables (i.e., individual gestures described within an articulatory phonology framework – see Browman & Goldstein, 1990, 1993). The Pi-gesture governs the time course of activation by slowing down the clock that "controls the time-flow of an utterance." Byrd et al. (2006) further suggested that the extent of the Pi-gesture "waxes and wanes" in the vicinity of a prosodic boundary and that other prosodic effects, like the proximity of major prosodic prominences, can influence the extent of lengthening effects. As in Hebrew (Berkovits, 1994), there is evidence of progressive lengthening throughout a phrase-final disyllabic word. The hybrid view according to Turk and Shattuck-Hufnagel (2007) is evident in Cambier-Langeveld's (1997) findings. The extent of final lengthening may be fixed according to some structural domain, i.e., the final syllable rhyme, but the extent of lengthening also reflects local phonetic constraints such as inherent duration and expandibility, particularly if the final vowel is schwa-like, and therefore inherently short. She found that in cases like this, lengthening was also apparent in the penultimate vowel. This position is intermediate to the previous two in that both local and structural influences determine the domain of final lengthening.

Turk and Shattuck-Hufnagel (2007) concluded that intonational phrase-final lengthening, in English at least, is complex and is not suggestive of progressive lengthening throughout a phrase-final word. It appears that lengthening is most apparent in the phrase-final syllable rhyme with degrees of lengthening of up to 90 percent compared to nonfinal syllables. Smaller but still significant degrees of lengthening are also apparent in the primary or main stressed syllable of a phrase-final word, but are not necessarily evident in any intervening syllables, and both lengthening effects are independent of whether the phrase-final word is carrying the nuclear accent for the phrase (Turk & Shattuck-Hufnagel, 2007). However, the crucial perceptual variable for Dutch listeners appears to be the *amount* of final lengthening, not the domain over which it occurs (Cambier-Langeveld et al., 1997).

2.2.4 Initial strengthening and lengthening There are a range of qualitative as well as durational effects that also mark the *left e*dge of different levels of prosodic constituency across a range of languages (e.g., Pierrehumbert & Talkin, 1992; Dilley et al., 1996; Fougéron & Keating, 1997; Fougéron, 2001; Keating et al., 2003; Cho & Keating, 2007; and see also the summaries of relevant literature in Keating, 2006; Harrington, 2003; Byrd et al., 2006; and Cho et al., 2007). There is also fairly substantial psycholinguistic evidence to suggest that domain-initial strengthening plays an important role in spoken-word recognition, word segmentation strategies, and lexical disambiguation (e.g., Quené, 1992; Cho et al., 2007, and references therein).

Word-initial consonants are longer than word-medial consonants, all things being equal, and phrase-initial consonants are longer than word-initial consonants in a range of languages including English and French (e.g., Byrd & Saltzman, 1998, 2003; Fougéron & Keating, 1998; Keating et al., 2003; Fougéron, 2001). Korean in particular appears to show clear lengthening in initial consonants that correlates with the strength of the prosodic domain boundary (Cho & Keating, 2001), even though final lengthening is largely associated with intonational phrase boundaries (Jun, 2005). Spatial strengthening of initial consonants is also observed at higher prosodic levels in English (e.g., Keating et al., 2004), Korean (e.g., Cho & Keating, 2001), and French (Fougéron, 2001), along with articulatory lengthening. For example, consonants show a higher level of linguopalatal contact (observed using electropalatography – EPG) in intonational phrase-initial position compared to medial positions. Byrd et al. (2005) also found significant durational effects of phrase position (intonational phrase initial and final) on tongue-tip and lip-aperture gesture duration in American English, but less consistent effects for spatial strengthening. Interestingly when it does occur, phrase-final coda consonants are also strengthened and lengthened as well as phrase-initial consonants, increasing or reinforcing the "salience" of these prosodic boundaries. In their study of the Radio News Speech corpus, Cole et al. (2007) did not find that stops are more acoustically distinct in intonational phrase-initial compared to medial position, but intonational phrase-initial stops are articulated with more precision and are less variable, constituting an additional strengthening strategy to those observed in articulatory studies. The effects also appear to be partially independent of prosodic prominence. For example, Cho and Keating (2007) observed higher degrees of articulatory strengthening in consonants (i.e., greater EPG linguopalatal contact and longer VOT values) that are *localized* at the left edge of utterances. Accentuation and primary stress, by contrast, are associated with other consonantal parameters like duration and energy (in syllable-initial nasals) and longer, more acoustically peripheral vowels, which suggests that accent-related effects are more extensive than initial strengthening effects, which are mostly observed in the domain-initial consonant or vowel gesture (see also Cho, 2005).

2.2.5 Polysyllabic shortening and medial shortening Another durational constraint that can operate within, or at the edge, of prosodic units is polysyllabic shortening. It is also referred to variously as anticipatory, compensatory, foot-level, or stress-timed shortening. In a classic study, Lehiste (1972) compared the duration of the initial syllable of words like "speed," "speedy," and "speediness" and found that syllables and vowels in particular, become progressively shorter as additional syllables are appended. Mean durations of 266 ms, 150 ms and 115 ms were noted for the primary stressed (and most likely, also accented) /i:/ in the above three tokens. A number of other studies have recorded this effect in English (e.g., Barnwell, 1971; Oller, 1973; Harris & Umeda, 1974; Huggins, 1975; Klatt, 1976; Fowler, 1981; Nakatani et al., 1981; Port, 1981; Rakerd et al., 1987; Beckman & Edwards, 1990; Turk & Shattuck-Hufnagel, 2000) and in other languages, including Swedish (e.g., Elert, 1964; Lindblom & Rapp, 1973), Dutch (e.g., Nooteboom, 1972, 1973), German (e.g., Kohler, 1986). The shortening effect also tends to level off beyond a certain word length (e.g., Oller, 1973; Fowler, 1981; Nakatani et al., 1981; Port, 1981) and it is somewhat asymmetric in that appending syllables to the left of a target syllable does not have the same effect (Lindblom & Rapp, 1973). Generally, it is thought to be a word-level shortening effect, although some have also claimed it can also operate within a higher-level prosodic domain like a phrase or utterance, for example (e.g., Lehiste, 1975; although see White, 2002, for an alternative view). It has also been interpreted as a potential stress-foot-level as well as word-level timing constraint, with suggestions that shortening of the initial stressed syllable may be a result of temporal compression processes relating to the attainment of isochrony (i.e., equal spacing of stress beats) in the production of stress feet in languages like English or German, for example (Huggins, 1975; Lehiste, 1977; Kohler, 1986), although this view is no longer widely supported (see section 3.1 for further discussion of this). Lindblom and Rapp (1973) also suggested that the shortening effect is part of the same speech production mechanism involved in determining preboundary lengthening. The mechanism involves looking ahead to what remains to be produced in a particular stretch of speech, and appropriate temporal compensation takes place accordingly. In the case of a final syllable there is little temporal adjustment because it is at the end of the particular constituent. Two influential later treatments of polysyllabic shortening (Beckman & Edwards, 1990; Turk & Shattuck-Hufnagel, 2000) are worth discussing here. Beckman and Edwards (1990) also argued that the effects are related, insofar as syllable durations are adjusted relative to some kind of prosodic constituent (i.e., a stress foot), but preboundary lengthening is a more clear-cut "edge" effect (as discussed in the preceding section) which marks out constituent boundaries like the phonological/prosodic or intonational phrase.

Lehiste (1972) found that word boundaries are largely ignored and polysyllabic shortening effects are present in comparisons of “speed” and “speedy” with “speed,” “speed kills” or “the speed increased.” Others have suggested that word boundaries may be important. Turk and Shattuck Hufnagel (2000) cited the findings of Huggins (1975) who found mainly within-word shortening effects. He compared sequences like “cheese **a**bounded” with “cheese bounded,” and “chees**es** bounded” with “cheese bounded,” and found polysyllabic shortening was much more evident when the extra syllable was added to the target word “chees**es**” than when the extra syllable was added across the word boundary (i.e., **a**bounded), which maintains effectively the same Abercrombian foot structure (i.e., where feet can be constructed across word boundaries). While some evidence for this was found in their study, Turk and Shattuck-Hufnagel (2000) proposed a complex range of interactions between word-final lengthening, polysyllabic shortening, and accent-related lengthening to account for their results. They suggested, like Beckman and Edwards (1990), that potentially different articulatory strategies are employed in syllable articulation in words depending on location of (prosodic) word boundaries, type of syntactic word (i.e., content versus function), and the near-vicinity of an intonational prominence or pitch accent. They also point out that unless a syllable is removed from the constituent above the word as syllables are added to the latter, compression may be operating at the higher-level constituent (phonological or intonational phrase), as indeed predicted much earlier by Lehiste (1980). She found that the shorter the utterance, the longer test-word durations tend to be. Conversely, test words shortened as a function of total length of utterance. White (2002) also suggested that polysyllabic shortening only occurs in pitch accented words in English, claiming that it is more to do with the word being the “locus” of accentual lengthening, “with lengthening greatest at word edges; variation in the distribution of accentual lengthening according to word length results in shortening of subconstituents in words of more syllables” (White, 2002, p. 275).

Another important general finding with regard to word-level shortening effects is that there are smaller shortening or compression levels for total number of syllables in the word, compared to initial syllable nucleus compression (Nooteboom, 1972; Lindblom & Rapp, 1973). Klatt (1976) incorporated a shortening factor of around 15 percent for segments in polysyllabic words in English. Nooteboom (1973) also found that patterns of unstressed syllable nucleus durations in Dutch words illustrate different degrees of shortening dependent on their position in a word. A syllable nucleus preceding a stressed syllable is somewhat shorter than a syllable nucleus preceding an unstressed syllable, for example. He hypothesized that this three-syllable pattern of short, very short, and long vowel durations is potentially universal. The durational buildup of words in other languages seems to indicate similar patterning. Lindblom and Rapp (1973) found this pattern was present, but less pronounced, in Swedish words. They further hypothesized that medial vowels are more temporally compressed than other segments due to the joint effects of anticipatory and backwards shortening in the production of syllables and words. Bruce (1983) also found patterns of durational alternation in adjacent unstressed syllables in Swedish. Early results for American English also seem to reflect these patterns, with medial vowels consistently shorter than vowels in other utterance or word positions (Lehiste, 1973; Oller, 1973; and Klatt, 1975). Nakatani et al. (1981) found that at each stress level, i.e., unstressed, secondary stress, and primary stress (most likely also nuclear accent), word-initial syllables were always longer than medial syllables but not as long as final syllables. Umeda (1975) also found that unstressed word-initial vowels exhibited much wider range of durational variation than word-medial vowels, which seems to support in part Nooteboom's hypothesis.

Similar patterns of nonuniform shortening have been observed in languages that have a different prosodic structure from English, Dutch, or Swedish. Unaccented medial vowels or syllables in French polysyllabic words show patterns of non-uniform shortening (O'Shaughnessy, 1981; Duez & Nishinuma, 1985; Fletcher, 1991), as they do in Italian (e.g., Farnetani & Kori, 1986; Vayra et al., 1999; Hajek et al., 2007). Estonian and Finnish do not show word-level temporal patterning of this kind even though they both have lexical stress (e.g., Lehiste, 1970; Asu & Nolan, 2006, 2009; Suomi, 2007). This calls to mind Nooteboom's (1973, p. 40) hypothesis that languages that have a "high functional load on phonological quantity patterns" may constrain the amount of permissible segmental shortening due to number of syllables in a unit, or the amount of segmental and syllabic lengthening due to positional variables. One also has to consider whether some of the earlier studies were also picking up on initial strengthening at the left edge of higher-level prosodic boundaries, particularly in studies based on isolated tokens.

3 Speech Timing: A Rhythmic Dimension

> *For a vocal and auditory communication system to function well, it must have some temporal constraints on transmission units. Such constraints necessarily impose a rhythmic structure that makes it possible for speakers to produce sounds efficiently, and for hearers to listen efficiently.*
>
> Oller (2000, p. 80)

Few speech researchers would disagree with this statement, although one of the longest debates in phonetics revolves around whether an independent rhythmic constraint underlies surface ordering or organization of speech intervals in spoken language, and if it does, how we can access it in our experiments. One prevailing view since the seventies is that rhythm is subordinate to syntactic or hierarchical prosodic constraints, and surface durational patterns of speech segments are largely due to the higher-level syntactic and/or phonological structures of language and can more or less effectively be modeled without reference to any kind of direct low-level rhythm component (e.g., the duration models of Klatt, 1976, 1979; and

Carlson & Granström, 1986, have no direct rhythmic component). Other quantitative duration models explicitly invoke rhythmic constraints (e.g., Campbell & Isard, 1991; Barbosa, 2007). Another view strongly echoes Kozhevnikov and Chistovich (1965), and sees rhythm as "the regular effects of temporal control over larger domains than the segment – mora, syllable, foot – without having to turn these domains into durational units" (Kohler, 2003, p. 8). Since the late nineties, there has also been renewed interest in developing metrics to measure linguistic rhythm in the acoustic signal in order to quantify perceived rhythmic differences among languages (e.g., Low, 1998; Ramus et al., 1999; Grabe & Low, 2002; Ramus, 2002; White & Mattys, 2007b). In order to understand this apparent paradox, it is useful to re-examine the sources of evidence for a rhythm component in spoken language, and to review research on linguistic rhythm typologies.

3.1 What is speech rhythm?

Psychologists such as Lashley (1951), Woodrow (1951), Lenneberg (1967), and Fraisse (1963) influenced a range of research on speech rhythm in the seventies (e.g., Martin, 1972; Allen, 1972a, 1972b, 1975) with the exploration of concepts of objective and subjective rhythm in auditory perception, and neurological features of motor behavior. In brief, listeners have a tendency to impose rhythmic structure on temporally unstructured material. Examples of such unstructured material include successions of beeps or drips from a leaking tap. There is a tendency to hear these stimuli as grouped, particularly if the succession of beeps, etc., is neither too fast nor too slow (i.e., between 0.1 and 3.0 seconds). There is a further subjective tendency to group these events into equal time intervals. In short sequences that that have equal temporal spacing, there is an additional "objective" tendency to hear the first pulse in these groups as being more perceptually salient than a following pulse if the first is louder or higher in pitch. If every third pulse is longer than the preceding pulses, then there is a tendency to hear the longer pulse as ending the rhythmic group (Allen, 1975, p. 77). Rhythm in this sense refers not only to sequences of like events, but to the alternation of strong and weak elements. Moreover, researchers also claim that some form of grouping relating to successions and alternations can be related to motor activity (see Allen, 1975, for an earlier review; and Cummins & Port, 1998; Port, 2003; and Cummins, 2003, for later reviews).

There also appears to be a preferred temporal rate of activity for the performance of motor tasks. Allen (1975, p. 79), citing a study by Woodrow (1951), wrote that "personally preferred rates have been found to range around an average of about two acts per second". Speech rhythms are also composed of successions and alternations of events that have a specific temporal paradigm. For example, some speech researchers originally related the 6 Hz cycle posited by Lenneberg (1967) for aspects of infant motor behaviour to signs of underlying rhythmic properties of spoken language. Ohala (1975) took up Lenneberg's hypothesis and attempted to find evidence of an underlying speech rhythm of 6 Hz. He measured the intervals between several thousand jaw openings, and the intervals between four thousand successive drops in oral pressure accompanying the release of voiceless obstruents. There was a clustering of data at around 250 ms but no robust evidence of a rhythmic strategy such as that proposed for a more "simple" motor activity.

The "rhythm hypothesis" outlined by Kozhevnikov and Chistovich (1965), whereby "utterances can be considered as prosodically although not segmentally and durationally identical," inspired a great deal of duration and articulatory timing research in the seventies (e.g., Lindblom & Rapp, 1973; Nooteboom, 1973). It was thought that the relative durations of neighboring syllables may exhibit some kind of abstract temporal invariance, whereas at the level of the microstructure, i.e., the acoustic segments that make up the units, there is a great deal of durational variance relating to a number of duration-influencing variables (e.g., position in syllable, foot, word, or phrase, level of prosodic prominence, speaker tempo). According to this view, speech rhythm is not a function of the sequence of absolute time intervals corresponding to syllable or foot-like units in a language. Syllables maintain constant relative durations to each other at the expense of the constituent segments. The absolute durations of intervals vary as a consequence of segmental composition and other duration-influencing factors, thus influencing the absolute timing of the higher-level timing unit, although the "abstract" relationship between the intervals remains constant or nearly constant.

Perhaps what clearly emerged from earlier experimental work in the seventies and eighties was a view of speech rhythm as rhythm of alternation, rather than succession of like events or strict periodicity. This view of rhythm has been formalized in linguistic theory as metrical phonology (Liberman & Prince, 1977; Hayes, 1995) and is considered to be an important principle in the phonological organization of many languages. However, the temporal view of rhythm as periodicity or at least quasi-periodicity is still alive and well in many studies of speech rhythm that explore the relationship between speech rhythm and other rhythmical activities such as limb movement, following the work of Kelso and colleagues (e.g., Kelso et al., 1985; Kelso, 1995). For example, Port and colleagues suggest that evidence of underlying rhythmic organization in speech can be uncovered through the use of rhythmic cycling tasks (see Port, 2003; Cummins & Port, 1998; Tajima, 1998; Cummins, 2002). A speaker repeats a short utterance in synchrony with a two-tone metronome whose tone and temporal spacing are subsequently manipulated and any resulting effects on durational patterning in the utterances (including spacing of stressed syllables) are observed. Cummins and Port (1998) propose that rhythm in speech involves "entrainment" of different metrical levels, i.e., rhythm in speech is interpreted as hierarchically ordered and reflects principles of organization and "periodic oscillation" similar to those that underlie locomotion (see also Cummins, 2009). Port (2003) also suggests that different oscillation patterns might reflect language-specific rhythm (see next section). Other studies stress the importance of dynamical systems in modeling speech rhythm and suggest that temporal structures emerge from self-organizing principles of dynamical systems and can be modeled accordingly (e.g., Barbosa, 2007).

3.2 Language-specific rhythm

Much of the experimental phonetic research on timing and rhythm has devoted itself to investigating rhythmic differences among languages. Pike (1946) and Abercrombie (1967) proposed that the majority of the world's languages can be divided into two main "timing" or "rhythm" categories, depending on the nature of the unit that seems to recur at equal intervals of time. Historically, languages like English, German, or Arabic, have been classified as stress-timed because stressed syllables were the source of the dominant rhythmic beat that recurs at regular intervals; whereas in French, Spanish, Yoruba, and Italian, individual syllables, whether accentually prominent or not, have been determined to be the primary generator of rhythm, hence their classification as "syllable-timed." The typology was further expanded to include a category "mora-timed" because the mora has historically been considered to be the primary timing or rhythmic unit in Japanese, for example (e.g., Han, 1962).

The notion of isochrony was important to earlier rhythm studies, although it has been more or less abandoned as a heuristic for classifying languages into a specific rhythmic category. The principle in its most extreme form is rooted very much in notions of rhythm as periodicity (see Couper-Kuhlen, 1993; Scott, 1998; Port, 2003), and states that temporal distance between stresses or stress beats (or stressed vowel-onsets) in stress-timed languages should remain reasonably constant irrespective of the number of syllables in the stress foot. In other words, there should be a degree of syllable compression within a foot. In addition, the Abercrombian foot can consist of a silent beat or silent stress followed by an unstressed syllable or syllables (e.g., Abercrombie, 1964) which also supposedly contributes to regularizing the time between stress beats in stress-timed languages like English. In syllable-timed languages, by contrast, stressed syllables should recur at unequal intervals, given that the duration of the inter-stress interval is dependent on the additive durations of the component syllables: in other words, there should be no foot compression, and syllables should be more or less isochronous. Similarly in mora-timed languages, the mora is the timing unit that should recur at more or less regular intervals, and segment durations will be compressed to regularize mora duration. As support of this theory of rhythm in speech production, Abercrombie (1964) called upon the work of Stetson (1951). Stetson claimed that in "syllable-timed" languages the production of syllables is accompanied by chest pulses and in "stress-timed" languages the stressed syllable is accompanied by a reinforced chest pulse. Subsequent articulatory investigation of these phenomena by Ladefoged (1967) and Ohala (1975) failed to show this, although since the eighties and nineties there has been renewed interest in Stetson's theories, and in particular a re-assessment of the role of the syllable in the temporal organization of speech (e.g., Kelso et al., 1985; Vatikiotis-Bateson & Kelso, 1993; Kelso, 1995).

Notions of "pure" isochrony in stress-timed languages were seriously challenged in the late seventies and eighties (see Dauer, 1983; Arvaniti, 1994; Cummins & Port, 1998; Ramus et al., 1999; Grabe & Low, 2002, for additional reviews of the relevant literature). Experimental investigations of isochrony in spoken English, for example, failed to show that stressed syllables or "stress beats" recur at equal intervals. Classe (1939) (cited in Cummins & Port, 1998, p. 146), found isochrony only among inter-stress intervals that were more or less identical in terms of syllable count, segmental content, and grammatical structure. Stress feet in stress-timed languages vary in physical duration and a number of studies have shown there is a monotonic relationship between foot duration in ms and foot length in syllables or number of phonemes in stress languages like English, Dutch, Swedish, and Arabic (e.g., Faure et al., l980; Nakatani et al., 1981; Roach, 1982; Beckman, 1982; Dauer, 1983; Strangert, 1985; den Os, 1988; Crystal & House; 1990; Fant et al., 1991; Williams & Hiller, 1994). This has been shown for spontaneous speech, read speech, and reiterant speech (i.e., where successions of like syllables, e.g., /ma/ or /ba/, mimic real utterances). Moreover, this monotonic relationship is also observed in stress or accent groups in syllable-timed languages like French, Spanish, Italian, or Telugu (Roach, 1982; Dauer, 1983; den Os, 1988; Fant et al., 1991; Fletcher, 1991; Pamies-Bertran, 1999) and there is no experimental evidence that inter-stress intervals in languages conventionally described as stress-timed are more equal than inter-stress or accent intervals in syllable-timed languages. Moreover, Fant et al. (1991) showed that inter-stress spacing is similar in Swedish (550 ms), English (565 ms), and French (555 ms). Roach (1982) also found that out of a group of stress-timed languages (English, Russian, Arabic) and syllable-timed languages (French, Telugu, Yoruba), English shows the most variance in inter-stress interval duration.

A further assumption of older views of syllable timing is that durations of open and closed syllables should be regularized in much the same way as durations of stress feet in stress-timed languages. In a classic examination of cross-linguistic duration patterns, Delattre (1966) found no evidence of differential degrees of temporal compensation in closed syllables in his analysis of French, German, Spanish, and English. Roach (1982) compared the standard deviations of syllable duration in French (75.5 ms) and English (86.0 ms) and concluded that it is equally variable in both languages. Similarly, as segments are added to simple CV syllables and more complex syllables are formed, there is no evidence that the degrees of shortening experienced by the constituent segments in French are more pronounced than in German, for example. Pamies-Bertran (1999) also found that syllables in Spanish lengthen as a result of increased syllable complexity. Similarly in Finnish, bimoraic syllables are longer than monomoraic syllables and there is no evidence of syllable-level isochrony (Lehtonen, 1970; Suomi & Ylitalo, 2003). Syllable duration appears to be as variable in syllable-timed languages as in stress-timed languages (Delattre, 1966, Crompton, 1980; Roach, 1982; den Os, 1988; Wenk & Wioland; 1982, Dauer, 1983; Fant et al., 1991; Fletcher, 1991). Durational adjustments within morae to compensate for short or long adjacent morae have also been interpreted as evidence of mora-timing in Japanese in some studies (e.g., Port et al., 1987). Other studies show a wide range of durational variation among morae depending on segment identity, suggesting that there is no evidence of mora-based isochrony in Japanese (e.g., Beckman, 1982; Hoequist, 1983b). The reader should also see the discussion in Beckman (1986) and Tajima (1998) for differing views. Warner and Arai (2001) suggest that word duration in spontaneous Japanese is best predicted on the basis of mora duration (see also Hirata, 2004), although they reject any notion of isochronous rhythm.

3.2.1 Isochrony: A tendency In the seventies and eighties, isochrony was re-interpreted as a perceptual phenomenon or a *tendency* in spoken English and "stress-timed" languages in general. Listeners, when asked to tap in time to syllables beginning with a particular plosive in controlled English utterances, tend to tap more regularly than the measured inter-stress intervals (e.g., Donovan & Darwin, 1979; see also Allen, 1975). This tendency to overestimate short intervals and underestimate long intervals ties in with the findings reported in studies of objective and subjective rhythm (Classe, 1939; Woodrow, 1951; Fraisse, 1963) discussed earlier. The notion of the "perceptual-center" or P-Center also became a popular one with regard to exploring the perceptual basis of isochrony (e.g., Morton et al., 1976; Marcus, 1981; Couper-Kuhlen, 1993; Pompino-Marschall, 1989; also see Scott, 1998 for a review). P-Centers are generally defined as perceptually salient "periodically recurring" events (at around 500 ms) that occur close to the vowel onset of stressed syllables in English, for example, although the location can be perturbed by factors like onset or coda complexity. Earlier studies suggested that the alignment of P-Centers influences the perception of periodic rhythm or isochronous speech.

Supporters of a weak isochrony hypothesis also referred to speech error research (e.g., Cutler, 1981) in which examples of syllable additions or omissions were interpreted as a means of evening up the number of syllables in stress feet. It was also claimed that isochrony at the foot or syllable level is also disrupted by other timing factors like preboundary lengthening or word-initial lengthening (e.g., Nooteboom, 1973; Huggins, 1975; Lehiste, 1977; Scott, 1982; Kohler, 1986). Lehiste (1977) proposed that this disruption of isochrony acts as a boundary cue to the listener, and that rhythmic structure in an utterance can be modified to indicate syntactic information. Other factors that can disrupt isochrony include number of constituent syllables in a foot, syllable complexity, and accentual lengthening in the final foot, if the head of that foot is also the location of the nuclear accent in the intonational phrase in languages like English and Dutch (e.g., Cambier-Langeveld & Turk, 1999; Turk & White, 1999). Of course, the problem is that many of these isochrony-disrupting factors also operate in syllable-timed and mora-timed languages (e.g., Beckman, 1982; Hoequist, 1983b; den Os, 1988; Fletcher, 1991; and Campbell, 1992b).

Foot-level compression in stress-timed languages has also been interpreted as evidence of weak isochrony (e.g., Huggins, 1972; Lehiste, 1973). That is, even though stress foot duration increases as more unstressed syllables are added, individual syllable durations should be compressed accordingly (see earlier discussion). However, Roach (1982) found that stress-timed languages (e.g., English, Arabic) and syllable-timed languages (French and Telugu) could not be distinguished using this metric. Furthermore, Nakatani et al. (1981) found that there was no evidence of foot-level compression in their corpus of American English reiterant speech. They measured sentences containing metrical feet of two, three, or four syllables in length and found that feet increase in duration monotonically with increasing syllable number, and rejected even a "weak isochrony" hypothesis for American English. Crystal and House (1990) backed up these findings in a corpus of read prose, reporting a more or less additive relationship between stress foot length in syllables, stressed syllable duration, and unstressed material in feet ranging from one to five syllables. Similar findings for read Swedish prose were reported by Fant et al. (1989), who noted correlations of 0.92 between stress foot duration and syllable count. By contrast, Kohler (1986) reported mixed results for German, showing that one- and two-syllable feet of the same complexity approached isochrony, but three-syllable feet did not show evidence of foot compression. Williams and Hiller (1994) also suggested that the very small foot-level compression (up to 39 ms) in their read sentence corpus of RP English was a much less significant factor than other durational factors such as number of segments per syllable, or presence of lexical stress or intonational accent. Port et al. (1987) also noted very small effects of syllable compression as a function of increasing word length in Japanese, but strong temporal compensation at the level of the mora, and concluded that this constituted evidence of the validity of the mora as the rhythmic timing unit of Japanese. However, Beckman (1982) did not observe strong mora-level temporal compression in her study of Japanese. O'Shaugnessy (1981) observed some degree of overall shortening in the constituent syllables of French words as they increased in length, whereas Duez and Nishinuma (1985) showed no effects of foot compression in French. They did find, however, a rhythm of alternation among adjacent unaccented syllables that was similar to effects previously reported for Swedish (e.g., Bruce 1983). Fletcher (1991) also found a clear monotonic relationship between foot or "rhythm group" length in syllables and overall duration in ms in a corpus of spontaneous French, with correlations of between 0.74 and 0.94, although nonuniform shortening was also observed in unaccented syllables.

Another related durational pattern, often interpreted as evidence of weak isochrony, is polysyllabic shortening in polysyllabic words (see also section 2.2). Recall that this is where initial stressed syllables, and in particular vowels, appear to shorten as syllables are added to the word, usually within the same foot. It was originally suggested that patterns of initial stressed-syllable shortening should be absent in languages that do not have the same rhythmic structure as German, Swedish, or English (e.g., Kohler, 1986). However, initial stressed-syllable shortening has also been observed in Italian, a syllable-timed language (e.g., Vayra et al., 1983, 1999; Hajek et al., 2007). Recall that polysyllabic shortening does not occur in Finnish or Estonian, which are two languages that have been variously described as syllable-timed or mora-timed. These results present ambiguous evidence of weak isochrony or rhythmic influences due to the confounds of final lengthening, syllable complexity, and general word-level timing effects (see Fletcher, 1991; Beckman, 1992; Turk & Shattuck-Hufnagel, 2000; Hajek et al., 2007; and studies reviewed in section 2.2). The initial stressed monosyllable in many early experiments was often nuclear-accented and phrase-final, or an isolated word or nonsense word and therefore also phrase-final and nuclear-accented, and thus long to start with. Once an unstressed syllable is appended, the stressed syllable is no longer in final position and shortens by up to 50 percent or more. Recall from section 2.4 that most studies of polysyllabic shortening found that it is often negligible beyond the addition of a second syllable. In fact, Crystal and House (1990) examined stressed vowel duration in non-prepausal feet and found no evidence of polysyllabic shortening in their large corpus of American English. Certainly many researchers these days adopt the less prescriptive position of Nooteboom (1972), Ohala (1975), Lindblom (1983), and Kohler (1986, 2003) who suggest that inter-stress intervals may be important production units in languages like English or Dutch, but they are not necessarily isochronous in nature. Many of the durational effects reported in earlier studies may also be more indicative of syllable-, word-, or phrase-level duration patterns and prominence effects, or are described more generally as compression effects or polysyllabic shortening (e.g., see review of these effects in Turk & Shattuck-Hufnagel, 2000, 2007, as well as discussion in section 2.2 above).

As Beckman (1992, p. 459) suggests:

> The interesting question to ask, therefore, is not what happens to the larger prosodic unit as subunits are added, but rather which is the consistently l**onger** subunit. For that we need to go inside the larger unit . . . and examine the actual length of component syllables as a function of their position within the hierarchy of stresses and phrases.

Some of these durational patterns were reviewed in sections 2.1 and 2.2, and formal models of prosody and prosodic typology have provided an additional framework within which to consider a range of prosodic effects on speech timing (see Jun, 2005; and Beckman & Venditti, this volume). Crystal (1969), Scott et al. (1985), and Vatikiotis-Bateson (1988), among others, suggest that rhythm typologies are somewhat Anglo-centric anyway. Nevertheless, there is still a high level of support for them at least in an informal descriptive sense, even if a strict categorical view has been replaced by a more gradient view for the most part (e.g., Dauer, 1983; Grabe & Low, 2002; and see discussion below). There must be something to rhythm typology, otherwise we would have moved on long ago. Certainly, the psycholinguistic literature or work on first language acquisition suggests that word segmentation processes provide some evidence for rhythm classes. For example, Cutler and colleagues have shown quite convincingly that strong syllable onsets (i.e., metrically stressed syllables) contribute to a metrical segmentation strategy in English, whereas languages like French ("syllable-timed") or Japanese ("mora-timed") use syllable-based or mora-based word segmentation strategies, respectively (e.g., Cutler et al., 1986; Cutler & Carter, 1987; Mehler et al., 1981; Cutler & Norris, 1988; Cutler & Butterfield, 1992; Otake et al., 1993; Otake et al., 1996; see also Cutler & McQueen, this volume). However, listeners may have at their disposal more than one segmentation strategy (e.g., Murty et al., 2007). Psycholinguists also suggest that infants (even newborns) are sensitive to rhythmic differences and use rhythm to help differentiate languages of different rhythm classes (e.g., Mehler et al., 1996; see also the review of relevant literature in Ramus et al., 1999, pp. 266–7; Port, 2003). Studies of infant babbling (e.g., Levitt & Aydelott-Utman, 1992) also conclude that syllable duration patterns can reflect "rhythmic" differences.

Even if many languages have accented or stressed elements that alternate with weaker or less marked phonetic elements, it seems that the perception of these events is heavily influenced by the linguistic and general phonetic characteristics of the language in question, and the first language of the listener. In two influential cross-language rhythm studies of the early eighties, Roach (1982) and Dauer (1983) claimed it is precisely the different phonetic manifestation of the phonological and phonotactic structure of languages like French, Spanish, or Arabic that contributes to an impression of "syllable timing" or "stress timing" as opposed to the absolute duration of intervals (see also Bertinetto, 1989, for a similar view). They proposed that the perception of rhythmic differences among languages is more to do with what goes on within, rather than across, inter-stress intervals. Reprising Classe (1939), Dauer (1983) concludes that the conventional rhythm categories have little to do with physical time either in the acoustic or articulatory domain, with the characteristic "rhythm" of a language determined largely by the phonological patterning of vowels, syllable structure, and stress, overlaid on a grid determined by perceptual universals and motor behavior. She claimed it is necessary to look at what is happening within a rhythmic unit (a suggestion echoed by Fant et al., 1991), rather than the way rhythmic units (syllables, morae, or stress feet) follow each other in time. Absolute and relative timing of syllables and larger units such as inter-stress or inter-accent intervals are determined mainly by a culmination of the above factors. The perceptual impression of a particular kind of rhythm is therefore only indirectly related to absolute durations of speech events.

3.2.2 Stress and prominence According to Abercrombie (1967), the majority of "stress-timed" languages have *lexical* stress, and successive stress groups or feet tend to consist of one to four syllables (e.g., Uldall, 1971, and Williams & Hiller, 1994, for English; den Os, 1988, for Dutch; Kohler, 1986, for German; Strangert, 1985, and Fant et al., 1991, for Swedish). In syllable-timed languages, inter-stress intervals tend to be longer or more variable in length. For example, inter-stress intervals (or inter-accent intervals in the case of languages without lexical stress like French), can consist of up to eight syllables in Spanish (e.g., Dauer, 1983; Hoequist, 1983a) and French (Pasdeloup, 1990; Fletcher, 1991; Jun & Fougéron, 2002), and up to ten syllables in Greek (e.g., Dauer, 1983; see also Arvaniti, 1994). Neither unit length in syllables nor overall physical duration is a reliable heuristic to sort a language into a particular rhythm class, however. Jun (2003) notes that accentual phrases in Korean (which is unclassified according to the traditional rhythm typologies) tend to be short, i.e., around three syllables. The majority of inter-stress intervals consist of two to four syllables in Dauer's (1983) Spanish data, and accentual phrases (or rhythmic units) in French also tend to be less than five syllables in length (Boudreault, 1970; Fonàgy, 1979; Wenk and Wioland, 1982; Pasdeloup, 1990; Fletcher, 1991; Jun & Fougéron, 2002), although den Os (1988) found that a large proportion of inter-stress intervals in Italian (also traditionally classified as syllable-timed) were between five and six syllables in length (see also Bertinetto, 1981). Of course, it may not be appropriate to compare foot length in Germanic or "stress languages" with groupings that are representative of some other kind of structural grouping like the accentual phrase (see also Arvaniti, 2009).

Not all researchers agree with classifications of one language or another as syllable-, stress-, or mora-timed. French is a good example of this (e.g., Grammont, 1946; Fónagy, 1979; Wenk & Wioland, 1982; Di Cristo, 2000). According to Grammont, the rhythmic units of French are not individual syllables, accented or unaccented, but groups of syllables that form right-headed iambic rhythmic feet or "groupes rhythmiques." Accordingly, Wenk and Wioland (1982) describe French rhythm as "trailer-timed," whereas English with its left-headed trochaic feet is described as "leader-timed." However, many syllable-timed languages like Greek, Italian, or Spanish have left-headed feet (e.g., Arvaniti, 1994), suggesting there is no simple relationship between headedness of stress feet or accentual phrases and the perception of a language as stress-timed or syllable-timed. Abercrombie (1967, p. 97) states "in order to have this immediate and intuitive apprehension of speech rhythm . . . it is necessary that the speaker and hearer should have the same mother tongue," suggesting a potential reason for the divergence of opinion about the syllable-timed classification of French, for example. Judgments of stress or prominence can be difficult when confronted with an unfamiliar or less well studied language, and native-speaker influences or intuitions can result in premature categorization of a language into one rhythm class or another.

In a related vein, the relationship between the acoustic and articulatory parameters that make up stressed and accented syllables and those of the surrounding unaccented syllables has often been cited as another possible source of perceived difference between languages classified as syllable- or stress-timed. In section 2.1, we saw that stress or prominence effects are not realized identically across languages. It is also extremely difficult to conduct a cross-linguistic study of this kind given the fundamental differences in prosodic structure among languages (see Barry et al., 2009, for a good relevant discussion). Nevertheless earlier cross-linguistic studies showed, for example, that stress is less marked durationally in Spanish compared to English (Delattre, 1966; Oller, 1979; Dauer, 1983; Hoequist, 1983b; Ortega-Llebario & Prieto, 2005). Fant et al. (1991) also noted that stressed/unstressed syllable duration differences are more evident in Swedish and English than in French, with differences of 100–150 ms compared to 50 ms. By contrast, den Os (1988) found that the degree of syllabic lengthening due to presence or absence of accent in Italian (syllable-timed) and Dutch (stress-timed) is roughly similar. Stressed syllable nuclei in English, Dutch, and German are usually associated with increases in duration and vocal effort compared to surrounding unstressed syllables whereas in French, duration is the main correlate of *accent primaire*: f_0 is also an important cue (Rigault, 1962; Vaissière, 1974, 1991; Léon &

Martin, 1980; Rossi et al., 1981; Di Cristo, 1999, 2000; Post, 2000; Welby, 2006), but others suggest duration is more important than f_0 (e.g., Delattre, 1965; Wenk & Wioland, 1982; Benguerel, 1971). The ratio of lengthening between the syllable preceding the accented syllable and the accented syllable itself is of great importance in cueing whether a syllable in French is accented or not (Benguerel, 1971). Fant et al. (1991) suggested there could be three factors – the lack of "nonterminal" stress (at least in their read corpus), the predominance of terminal stress (i.e., *accent primaire*), and the lack of significant vowel reduction in French – that lead to perceptions of greater regularity of successive syllables by non-native speakers.

The avoidance of stress clash in languages like English, formalized in metrical phonology as the rhythm rule (e.g., Liberman & Prince 1977; Hayes, 1995; see also Beckman et al., 1990; Beckman & Edwards, 1994; Vogel et al., 1995, and references therein), is often cited as evidence of the importance of alternation in the specification of English rhythm. In other words, in sequences like "thir′teen" and "′thirteen ′men," the primary stress shifts from the second syllable of "thirteen" to the initial syllable in "thirteen men." However, Pointon (1980) points out that Spanish also exhibits stress shift and rhythmical alternation, e.g., "Jo′sé" versus "′José An′tonio." No-one would argue that Spanish is stress-timed on the basis of this. However, Spanish does not have the same tendency as English to move stresses around to break up potentially long inter-stress intervals, a feature that is shared with other syllable-timed languages. For example, Arvaniti (1994) investigated rhythmic properties of Greek, a syllable-timed language which, like Spanish, has stress but does not have a strictly alternating rhythm. Stressed syllables are relatively sparse compared to English and stress clash is not always resolved and syllable-timed languages tend be less "eurythmic." She suggests that rhythm in both types of language is organized around prominent syllables, but differences relate to different settings for the hierarchical structure of rhythm, rather than to isochrony tendencies, syllable structure, or the acoustic correlates of stress and prominence (see also Arvaniti, 2009).

3.2.3 Vowel reduction A number of authors refer to vowel reduction when outlining the contrasts between stress-timed, syllable-timed, or mora-timed languages (Delattre, 1969; Roach, 1982; Dauer, 1983; Hoequist, 1983b; den Os, 1988; Beckman, 1992; Ramus et al., 1999; Frota & Vigário, 2001; Grabe & Low, 2002; White & Mattys, 2007a, 2007b). In fact, vowel reduction is only one of a number of reduction processes that operate to "maximise the difference between stressed and unstressed syllables in a stress-timed language" (Dauer, 1983, p. 57). Many languages experience some formant target undershoot in unstressed syllables uttered at normal tempo (see section 2.1 and Harrington, this volume). Dauer (1983, p. 57) pointed out that syllable-timed languages "do not regularly have reduced variants of vowels in unstressed position," suggesting that the issue here has more to do with what is meant by "vowel reduction" than with evidence of underlying rhythm per se. It is obvious that languages have different patterns of phonological (and phonetic) vowel reduction in unstressed or unaccented syllables (see the discussion in Crosswhite, 2003b). Romance languages like Spanish, French, or Italian typically do not centralize unstressed vowels to a schwa-like vowel, as does English, for example. Koopmans-van Beinum (1980) and den Os (1988) compared spectral vowel reduction in Dutch and Italian, and found significant differences between these languages. On the other hand, unstressed or unaccented vowels in Italian, French, and Spanish show some formant undershoot as well as being shorter (e.g., Delattre, 1969; den Os, 1988; Bertinetto & Fowler, 1989; Farnetani, 2000), and patterns of vowel space reduction are observed in short versus long vowels in these languages (Gendrot & Adda-Decker, 2007). Also languages within a rhythm category like English and German can show greater differences in the degree of vowel centralization in stressed versus unstressed syllables than languages across the rhythm divide like French and German. Reduced variants of French and German vowels can be closer (at least in terms of formant values) to their stressed or accented counterparts than unstressed vowels are to vowels in stressed syllables in English (e.g., Delattre, 1969). These results tend to suggest that vowel reduction alone does not determine whether a language is syllable-timed or stress-timed.

3.2.4 Syllable structure and syllable duration Syllable structure and syllable duration within inter-stress and inter-accent intervals are two factors that potentially contribute to an impression of syllable timing versus stress timing (Roach, 1982; Dauer, 1983). Abercrombie (1967) suggested that stress-timed languages show a considerable degree of syllable duration variation due to a greater range of permissible syllable types. Delattre (1966) compared the maximum range of syllable duration variation under the varied conditions of stress, utterance position, and syllable structure in English, Spanish, German, and French, and found that English shows the most overall durational variation across all factors with syllable duration ratios of 1 : 3.39 compared to Spanish at 1 : 1.77, with French in the middle at 1 : 2.48. Subsequent studies have more or less supported these early results. For example, Crystal and House (1990) found that syllable durations in American English can vary between 120 ms (in open syllables) and 483 ms (in non-prepausal stressed CCVCC syllables), compared to 516 ms for prepausal syllables of similar complexity, with the lowest value dropping to 70 ms for unstressed open syllables. Hoequist (1983b) noted durational variation of between 101 ms and 169 ms in Spanish syllables, whereas Fletcher (1991) observed durational variation from 80 ms for unaccented syllables to a maximum of 250 ms in accented syllables in non-prepausal environments in her spontaneous French corpus.

Dauer (1983) concluded that the impression of greater syllabic regularity in syllable-timed languages like Spanish comes from the repetition of structurally similar syllables. In terms of distribution of syllable type across languages, Dauer found that nearly 70% of syllables in Spanish corpora were open, compared to 44% in English. Similar proportions of open to closed syllables have been noted in French, ranging from 70% to 89% (e.g., Fletcher, 1991). Ramus et al. (1999) cited Nespor (1990), who reported that 60% of all syllables were open in her Italian corpus, compared to 43% for Dutch (Levelt & van de Vijver, 1998, also cited in

Ramus et al., 1999). European Portuguese has 59% CV syllables (cited in Frota & Vigário, 2001), but it is classified as a mixed rhythm language, i.e., it shares elements of syllable timing and stress timing. As with vowel reduction in non-prominent syllables, it is actually hard to talk about syllable structure in isolation from other sources of phonological variation like stress or prominence. Syllable structure and stress are closely linked in the case of quantity-sensitive languages (e.g., Hayes, 1995). Although complex syllable structure and stress seem to co-occur in many languages, this co-occurrence serves to reinforce the perceptual impression of "stressedness." The fact that French or Spanish do not have as many complex syllable structures as English or German could mean that stress and open syllables do not have the same perceptual "impact" as stressed heavy syllables to a non-native speaker's ear at least.

The different acoustic and durational patterns of stress, syllable structure, and levels of vowel reduction may in fact contribute to perceived rhythmic differences among languages. Nevertheless it is still difficult to come up with objective measurements to quantify degrees of vowel reduction, syllable elision, and accented or stressed syllable salience in order to establish a threshold between different categories. Languages within the categories differ in precisely these ways. For a number of years researchers have argued that all languages exhibit elements of syllable timing and stress timing (e.g., Crystal, 1975; Roach, 1982). Moreover typical features of mora-timed languages, like gemination, vowel quantity contrasts, and tonal features, have also been investigated in similar kinds of ways with mixed results (see Beckman, 1986, for an early review). Languages like Tamil have been analyzed as stress-timed, syllable-timed, or mora-timed or as belonging to none of these categories (see Keane, 2006, for an overview on previous and current research on Tamil rhythm). Estonian has also traditionally been analyzed as syllable-timed (e.g., Eek & Help, 1987), although we noted earlier that it is also a language that has lexical stress with a complex vowel quantity system, and its rhythm has also been classified as stress-timed (e.g., Asu & Nolan, 2005, 2006, 2009).

As mentioned above, there have been various proposals to deal with this apparent impasse. Dauer (1987) suggested that linguistic rhythm can be viewed more as a gradient continuum depending on the combination of structural features that are thought to be prime "rhythm" indicators. Some languages will have more stress-timed characteristics, some will have more syllable-timed features, and so on. Others have suggested that rather than a rhythmic continuum, there is a set of intermediate languages (e.g., Ramus et al., 1999, after Nespor, 1990). An example that is often cited is Polish, a stress-accent language with extremely complex syllable structure but relatively little vowel reduction. Catalan is also classed as an intermediate language because it shares many structural features with Spanish but has vowel reduction in unstressed contexts. Ramus et al. (1999) concluded that more work on less well studied languages might lead to the emergence of more linguistic rhythmic categories. Other proposals have also been put forward to suggest at least five rhythm classes (Auer, 1993), whereas Cummins and colleagues (e.g., Cummins & Port, 1998, and Cummins, 2002, 2003; see also Tajima, 1998) propose a more hierarchical view of language rhythm that returns to fundamental principles of motor rhythms (see Kohler, 2009, for an interesting overview). Nevertheless, there has until now been a high level of acceptance of the gradient view as proposed by Dauer (1987) or a mixed view of language rhythm (e.g., Asu & Nolan, 2009). Since the late 1990s a range of duration metrics or measures were developed to "sort" languages into prototypical rhythm classes like English or Spanish, or to assist in the classification of previously unclassified languages, with varying degrees of success. These measures are summarized in the following section.

3.3 Alternative metrics

Since the late nineties, there has been a resurgence of interest in rhythm classes. A series of phonetic duration measures building largely on psycholinguistic evidence of rhythm perception, have been developed to avoid reliance on "phonological" constructs like stress feet or syllables. Syllable division in languages like English is not always straightforward (see, e.g., Blevins, 1995) and these duration measures were designed to get around this particular issue (e.g., Ramus et al., 1999; Ramus, 2002). One aim was to see whether languages already classified as either mora-timed, syllable-timed, or stress-timed can be separated according to these relatively simple metrics. A second aim was to see how languages which have either been described as having "mixed" or "intermediate" rhythm (e.g., like Catalan) pattern in relation to other languages belonging to the other rhythm classes. Two of the Ramus et al. (1999) measures, %V (percentage of vocalic interval duration in an utterance) and ΔC (consonant duration standard deviation in an utterance), appear to group languages into traditional rhythm classes when plotted against each other. In Figure 15.3, stress-timed languages (e.g., English, Dutch, German, and Polish) cluster in the top left corner of the plot, whereas syllable-timed languages (Spanish, French, Italian, and Catalan, which is actually classified as a mixed rhythm language) cluster below the stress-timed languages, with Japanese, showing the highest value of %V and lowest ΔC, to the bottom right of the plot. Two additional measures, ΔV (vowel standard deviation) and C% (percentage of consonant interval duration), were also investigated by Ramus et al. (1999) and the former measure was found to relate to the degree of vowel reduction and the shortening of vowels in unstressed syllables in the languages in question. These measures therefore largely reflect patterns of segmental inventory (i.e., whether a language has short versus long vowels, or unstressed reduced short vowels) and phonotactic differences (particularly syllable structure) between the languages; patterns that were previously attributed by Roach (1982) and Dauer (1983, 1987), among others, to be some of the prime contributors to *perceived* language rhythm. As a result, it is not surprising that the prototypical stress-timed languages in the Ramus et al. study show *lower* values for %V and *higher* values for ΔC reflecting syllable complexity. They suggested that this measure results in Polish being grouped with English and Dutch (larger number of complex syllable types), whereas Catalan clusters with Spanish (fewer syllable types). They also pointed out that the addition of more languages to their study might start to blur

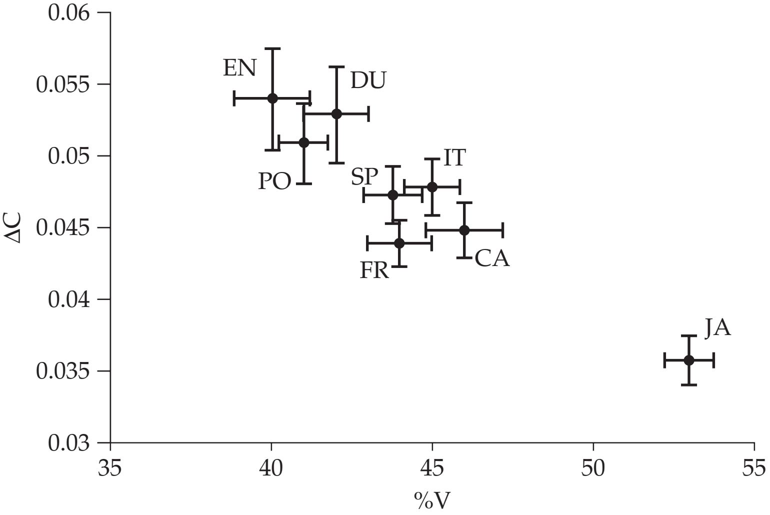


Figure 15.3 Distribution of languages using %V (proportion of vowel interval duration/utterance) and ΔC (standard deviation of consonant interval duration) (from F. Ramus, M. Nespor, & J. Mehler, 1999, Correlates of linguistic rhythm in the speech signal, *Cognition*, 73, 265–92, p. 273). (Reproduced by permission of Elsevier)

the boundaries between the language "clusters" thrown up by their metrics, or further clusters may emerge that do not necessarily adhere neatly to any of the traditional three rhythm categories. Ramus (2002) later suggested that an alternative interpretation of these measures is also possible, namely that they highlight rhythmic differences among languages, but do not necessarily define classes.

The Ramus measures have been applied to a number of different languages, often in conjunction with the *pairwise variability index* (PVI) which is another frequently used computation in recent rhythm class studies (e.g., Low, 1998; Low et al., 2000; Deterding, 2001; Grabe & Low, 2002; Asu & Nolan, 2005, 2006, 2009; Keane, 2006; White & Mattys, 2007a, 2007b). Grabe and Low (2002) proposed two PVI measures – a raw pairwise variability index *rPVI* (1) and a speaker-tempo normalized index: *nPVI* (2) – to look at *sequential* timing variability in *consecutive* intervocalic intervals or adjacent segment intervals in 11 languages that either have a conventional rhythmic label (e.g., English as stress-timed and French as syllable-timed), or are intermediate languages (e.g., Polish), or had not as yet been assigned to one of the three categories (e.g., Malay). They sought to tap into differences of rhythmic alternation among adjacent or near adjacent elements in the speech stream.

$$rPVI = \left(\sum_{k=1}^{m-1} |(d_k - d_{k+1})| \right) / (m-1) \tag{1}$$

$$nPVI = 100 \times \left(\sum_{k=1}^{m-1} |(d_k - d_{k+1})/((d_k - d_{k+1})/2)| \right) / (m-1) \qquad (2)$$

Grabe and Low (2002) also computed %V and ΔC statistics for the corpus. Like Ramus et al. (1999), they found that some languages tend to cluster where one would predict, i.e., with other typical stress-timed languages (most notably the Germanic languages, Dutch, British English, and German; illustrated in Figure 15.4). Specifically, a high vocalic PVI suggests stresstiming, where as a low vocalic PVI suggests syllable timing. However, as pointed out by Grabe and Low (2002) themselves, and later by Keane (2006, pp. 306–7), the different metrics produce quite different outcomes for languages like Thai or Tamil, which pattern with stress-timed languages on the basis of PVI measures, but with syllable-timed languages using the Ramus et al. (1999) metrics.

The PVI metrics have been adjusted in various ways to try and overcome some of these issues. Deterding (2001) used a *syllable PVI* in his comparison of spontaneous British English and Singapore English, which is often claimed to exhibit syllable timing (e.g., Low, 1998). He concluded that there are differences in syllable-to-syllable variation, with British English exhibiting a higher level of variation irrespective of speaker tempo, although he suggests that the higher incidence of

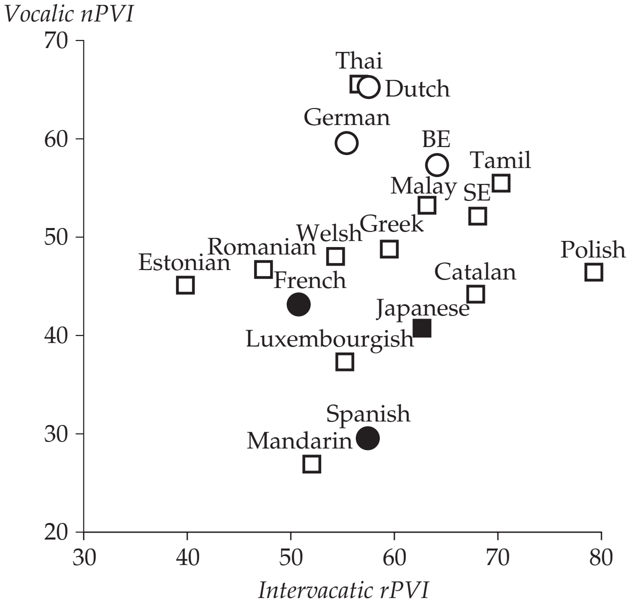


Figure 15.4 Distribution of languages using pairwise variability indices (from E. Grabe & E. L. Low, 2002, Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7* (pp. 515–43), Berlin: Mouton de Gruyter, p. 530). (Reproduced by permission of Walter de Gruyter)

reduced syllables in British English compared to Singapore English is a strong contributing factor. Barry et al. (2003) employed another PVI index *PVI-CV* which measures consonant + vowel sequences, which others have related to the notion of the "articulatory" syllable postulated by Kozhevkinikov and Chistovich (1965). Barry et al. (2003) applied this measure and the Grabe and Low (2002) and Ramus et al. measures in their comparison of read and spontaneous German, Bulgarian, and Italian and found that %V and PVI-CV achieved the best language-specific differentiation, although they also concluded that these metrics are not necessarily the best way to model typological rhythm (see also Barry et al., 2009). Keane (2006) also showed that Tamil cannot be assigned easily to any particular rhythm class using the combined Grabe and Low (2002) and Ramus et al. (1999) metrics.

Asu and Nolan (2005) added a normalized *foot PVI* in their study of Estonian and English rhythm, claiming that the other PVIs do not capture other types of "rhythmic" effects like stress-foot compression. While fully acknowledging the problems of defining feet in two different ways (i.e., a full vowel nucleus defines the head of the foot in English whereas in Estonian the lack of significant vowel reduction makes this impossible), they claim that the combination of normalized *syllable PVI* (after Deterding, 2001) and normalized *foot PVI* show that Estonian and British English have similar foot-level timing, but different syllable-level timing patterns which they suggest are due to differential foot compression tendencies in the two languages. Estonian has little vowel reduction and compresses syllables in a foot relatively evenly, whereas strong vowel reduction tendencies in English are largely responsible for foot-internal syllable-timing patterns. They concluded that Estonian is a mixed rhythm language showing elements of syllable timing and stress timing.

Combinations or variations of these measures and other and similarly inspired metrics have focused specifically on languages of "mixed" or "intermediate" rhythm, or have sought to assign previously unclassified languages to one of the three rhythmic categories (see for example the aforementioned study by Keane, 2006; see also Gibbon & Gut, 2001; Gibbon & Romani Fernandes, 2005; and Arvaniti, 2009, for addtitional cross-language comparisons). For example, Frota and Vigário (2001) compared European and Brazilian Portuguese and found that the former, classified traditionally as stress-timed, shows combined features of syllable timing (on the basis of high %V) and stress timing (on the basis of high ΔC). Brazilian Portuguese, traditionally classified as having syllable-timed rhythm or mixed syllable-timed–stress-timed rhythm, patterns more closely with mora-timed languages on the basis of high %V and syllable-timed languages (i.e., low ΔC). They explained their results in terms of the phonotactic and intonational differences between the two varieties. For example, vowel deletion in unstressed environments in European Portuguese results in consonant clusters (and therefore a higher degree of consonant duration variability and reduced V%), whereas in Brazilian Portuguese, consonant clusters are usually broken up via vowel epenthesis resulting in an opposite pattern. Interestingly, they also suggested that perceived rhythmic differences are due to other factors such as differing intonational and tonal properties between the two varieties. For example Brazilian

Portuguese exhibits similar intonational patterns to nonstress accent languages like Korean and Japanese.

Some of these rhythm studies have used combinations of speech rate or tempo-controlled measures, others have not (see, e.g., Dellwo, 2004, 2006; White & Mattys, 2007b). Ramus (2002) suggested that tempo must be controlled in corpora for measures like %V or ΔC to be useful. Barry et al. (2003) also suggested that languages tend to become "more syllable-timed" at faster tempo, and that both ΔC and ΔV decrease as speaker tempo increases. This is not surprising, as one of the effects of fast tempo is to shorten vowel and consonant segments (see section 4.3 below). By contrast Dellwo and Wagner (2003), in their study of a large corpus of French, English, and German, found that ΔC decreases with increased tempo and that %V is relatively constant. Barry et al. (2003) also found that %V is immune to changes in tempo in Italian and German. White and Mattys (2007b) showed that V% is the most efficient measure that discriminates syllable-timed languages (i.e., French and Spanish) from stress-timed languages (English and Dutch) across speaking rates. Two further measures, VarcoV and VarcoC (the variation coefficient of the standard deviations of vowel and consonant interval durations, i.e., for example ((VarcoC = ΔC * 100)/mean C)), were applied by Dellwo and colleagues to a range of languages in order to "normalize" speaking rate (e.g., Dellwo et al., 2004; Dellwo, 2006). They found in that VarcoC performed better than ΔC at discriminating French from German or English across all speech rates. In a later study Dellwo et al. (2007) compared percentage of voiced intervals (%VO) and the variance of the standard deviation of unvoiced intervals (VarcoUV) in two stress-timed languages (English and German), and two syllable-timed languages (Italian and French). Both measures resulted in effective separation of the two pairs of languages. They found that the two stress-timed languages showed higher variability in unvoiced interval duration and a smaller overall percentage of voiced intervals compared to the two syllable-timed languages. They concluded that this metric is easier to implement (as voicing detection is somewhat easier to automate than consonant and vowel interval labeling). Furthermore, they cite the work of Ramus and others (e.g., Ramus et al., 1999) suggesting that infants may be sensitive to variations in overall voicing variation. However, measures like this still do not necessarily show anything other than patterns of voicing in a language, and may have little to do with rhythm as such.

Various combinations of the new measures have also been applied to the speech of second language speakers (e.g., White & Mattys, 2007a, 2007b; Gut, 2003). Comparative studies have also been carried out on varieties of English including Nigerian English (Gut & Milde, 2002), Pakeha and Maori English (e.g., Warren, 1997; Szakay, 2006), varieties of British English (White & Mattys, 2007a), as well as Singapore English (Low et al., 2000; Deterding, 2001). Most conclude with discussions of the phonotactic and prosodic differences between the different target languages or varieties, which suggests that the measures in themselves are tapping indirectly or directly into the phonetic and structural patterns of a language that contribute to perceived rhythm. For example, White and Mattys (2007a) showed that VarcoV correlates best with ratings of "accentedness" of Spanish speakers of English, and they suggest this measure reflects lower variability in the vowel duration cycle of Spanish compared to English. White and Mattys (2007b) also commented on the relative insensitivity of the metrics to structural and phonetic issues relating to stress, which has long been suggested as one of the potential sources of perceived rhythm differences (e.g., Dauer, 1983, 1987; Bertinetto, 1989; Fletcher, 1991; Fant et al., 1991; Arvaniti, 1994, Arvaniti, 2009). Many of the studies have also commented on the degree of variability among speakers of the same language as well as among languages of supposedly identical rhythm "classes" (see Keane, 2006, and her discussion of Tamil rhythm, for example; and White & Mattys, 2007b). Some of the rhythm studies since the late nineties have also included perception results. Ramus et al. (2003) used re-synthesis to test the perceptual reality of rhythm classes in a series of listening discrimination experiments. Sentences from the original eight languages that were part of the Ramus et al. (1999) study were re-synthesized to produce what is called flat "sasasa" speech (Ramus & Mehler, 1999). This involves removing segmental information by replacing all consonants with /s/ and vowels with /a/, and flattening pitch to 230 Hz, but retaining acoustic duration patterns and intensity fluctuations. They compared different pairs of languages that either were clear examples of syllabletiming versus stress timing (i.e., English vs. Spanish), or not so clear examples (e.g., Polish vs. English), as well as examining a within-class pair (Dutch vs. English). French listeners were able to easily discriminate Spanish from English, Polish from Catalan, Polish from English, and Polish from Spanish. Two pairs of languages; Dutch and English, and Catalan and Spanish, were not well discriminated by listeners and Ramus et al. (2003) concluded that Catalan must therefore be syllable-timed.

One could question whether the new rhythm measures are actually tapping into rhythm directly, let alone sorting previously unclassified languages into rhythm classes or settling them into a slot on the rhythm class continuum. As mentioned in the introduction to this chapter, Kohler (2003) questioned the usefulness of trying to uncover evidence of underlying rhythm that is essentially not going to be revealed by the measurement of acoustic intervals (see also Oller, 2000; Cummins, 2002, 2009; Barry et al., 2009; Arvaniti, 2009; and Kohler, 2009, for interesting discussion on this topic). It is certainly true that many of the newer measures throw up some unexpected clusters of languages that have quite distinct lexical prosody and postlexical prosody, e.g., Thai and Dutch pattern closely together using PVI measures, although, as pointed out by Grabe and Low (2002), Thai has been previously classified as stress-timed. However, like Tamil, Thai patterns with syllable-timed languages if the Ramus et al. metrics are applied. Analyses of Czech have also thrown up similar paradoxes (see, e.g., Dankovičová & Dellwo, 2007). While traditionally classified as syllable-timed, Czech also patterns strongly with stress-timed languages depending on the duration measure. The mixed results for Czech reflect the different sensitivities of the metrics, which seems to be a typical outcome of many experimental studies based on the traditional rhythm classes. On the other hand, issues of cross-language durational variation continue to be an important component of experimental phonetics research, which is a very good thing.

Perceived rhythmic differences among languages are also clearly related to higher-level prosodic prominence and/or patterns of prominence. The temporal effects within larger prosodic groupings than the syllable, foot, or accentual phrase or the alternation of intonational pitch accents (phrase-level stress), as well as lexical stress, may also contribute to perceived rhythmic differences among languages as suggested by Klatt (1976), Beckman (1992, p. 460), and Jun (2005), among others (see also Cummins & Port, 1998; Arvaniti, 2009). For example, Beckman (1992) has suggested that differences in tune–text alignment (i.e., whether a language has lexical tone or postlexical pitch accent) may influence the ways in which we hear rhythmic differences (see Beckman & Venditti, this volume, for a summary of tune−text differences among languages). Kohler (2003, p. 8) proposed a "pitch accent domain of timing," suggesting that pitch marking of certain syllables in words as intonationally prominent can be associated with a degree of temporal regularity (see also Barry et al., 2009 for a similar conclusion relating to the perception of linguistic rhythm). Germanic languages and many other so-called "stress languages" show that unaccented syllables can forgo complete reduction or deletion, whereas pitch-accented syllables can often carry the "bare bones" of rhythm in connected speech, which is not always the case in languages like Italian or Spanish, insofar as extreme reductions are not as common in unaccented contexts (see also Arvaniti, 1994, 2009). Jun (2005, p. 442) suggests that the perception of a language's prosody is influenced by higher-level prosodic structure, i.e., the perception of postlexical accentual prominence or the marking of constituent edges using either pitch accents, boundary tones, or some combination thereof. Perceived rhythmic variability is greater in these larger prosodic or "macro-rhythmic" units compared to smaller "micro-rhythmic" units, i.e., the mora, syllable, or foot. These macro-rhythmic units will also have size constraints, (as in Korean, where accentual phrase length is three to five syllables on average, but no longer than seven). When utterances are long, they are more likely to be broken into two intonational phrases, whereas short utterances tend to be produced as a single intonational phrase, and this can influence perceived rhythm. A fruitful direction for future research may be to focus on temporal patterns within or across these larger macro-rhythmic units. Kohler (2009) also suggests it may be time for a paradigm shift and that future rhythm research should focus on broader issues of perception and communicative function. There continues to be great interest in the complex area of speech rhythm and the interested reader should examine the various contributions to a special issue of the journal *Phonetica* (vol. 66, 1−2, 2009).

4 Tempo and Pausing

4.1 Introduction

Rhythm and tempo are often linked in studies of speech (as in studies of music), largely because both contribute to the perceived temporal flow of connected speech (e.g., Fraisse, 1963; Bertinetto, 1981; den Os, 1988; Cummins, 2002). Abercrombie (1967, p. 96) defined tempo as the "speed of speaking which is best measured by rate of syllable succession . . . it is . . . varied from time to time by the individual speaker." Most speakers have a particular habitual tempo (e.g., Tsao & Weismer, 1997), and variability of inter- (and intra-) speaker tempo has also been studied for a range of different languages (e.g., for German: Butcher, 1981; Künzel, 1997; Trouvain, 2004; for English: Miller et al., 1984; Crystal & House, 1990; for Dutch: Eefting, 1988; den Os, 1988; Verhoeven et al., 2004; Quené, 2007, 2008; for French: Malécot et al., 1972; Grosjean & Deschamps, 1975; Grosjean, 1979; Duez, 1981, 1982; Fletcher, 1987; Bartkova, 1991; Zellner, 1994, 1996; Fougéron & Jun, 1998; for Czech: Dankovičová, 1997; for Japanese: Koiso et al., 1998).

We saw in the preceding discussion that some of the newer studies of rhythm normalize or use tempo-controlled algorithms to reduce the effects of speech rate or tempo variation in their measures. However, tempo is a factor worthy of investigation in its own right precisely because of the way in which it interacts with the types of duration patterns we have described in earlier sections of this review. A range of factors influence tempo variation (see Trouvain, 2004, ch. 2 for a useful review). Certain styles of discourse are associated with slower or faster tempo, and it can also reflect a speaker's emotional state (Goldman-Eisler, 1968; Butcher, 1981; van Bezooijen, 1984). For example, tempo is generally more variable in spontaneous speech and spoken interaction compared to read prose (e.g., Butcher, 1981). Slowing down or deceleration near important or salient sections, or adjacent to dialogue turn transition points, can signal elements of the information structure of the discourse (e.g., Eefting, 1991; Nooteboom & Eefting, 1994; Hirschberg & Nakatani, 1996; Koopmans-van Beinum & van Donzel, 1996; Koiso et al., 1998; and references in Quené, 2007). Quené (2007, p. 353) suggested that the last-mentioned role of tempo can be related to the hypo-speech–hyper-speech continuum (Lindblom, 1983, 1990) in that speakers may deliberately slow down articulation rate to improve the communicative channel between hearer and listener (see also Eefting, 1991, for a similar view).

Emotions like sadness, grief, boredom tend to be associated with slower tempo, whereas the converse is often the case for anger, happiness, or frustration (see Trouvain, 2004, for a useful summary). Tempo differences have also been related to sociolinguistic milieu and regional differences (Duez, 1981; Verhoeven et al., 2004; Quené, 2008), and sex and age (e.g., Malécot et al., 1972; Whiteside, 1996; Verhoeven et al., 2004). For example, comparisons of Dutch spoken in the Netherlands and Flanders Dutch show that the former is spoken more rapidly and is less variable, and that there is also a tendency for males to speak faster than females (Verhoeven et al., 2004; Quené, 2008). Older people speak more slowly than younger adults (e.g., Malécot et al., 1972) and in a major review and study of the phonetic effects of aging, Schötz (2006) suggested that speech and reading rates can decelerate by up to 20–25 percent with age, and longer pauses, greater hesitancy, and lower articulation rates contribute to the perception of aging. Quené (2008) also reported an age effect in his re-analysis of Dutch varieties, but suggests it may be more to do with the measurement of tempo in earlier studies which are picking up on the tendency of older people to produce shorter phrases.

Tempo variation has also been examined in relation to hearing and language impairment. Early studies report intelligibility benefits of speech produced at slower speech rates (generally also "clear speech") for hearing-impaired adults and normal-hearing adults in noise (e.g., Picheny et al., 1986; Uchanski et al., 1996). This benefit can also extend to clear speech produced at fast speech rates (e.g., Krause & Braida, 2002). However, Uchanski et al. (2002) claimed that it is important not to confuse normally produced slow speech with clear speech as the latter is also associated with a range of additional acoustic-phonetic modifications as well as durational variation (Bradlow, 2002; Bradlow et al., 2003; Smiljanic & Bradlow, 2008, and references therein). Tempo variability is also often associated with non-native language proficiency, although there is not always a straightforward relationship between perceived fluency (and lack thereof) and objective quantitative measures of tempo (e.g., Cucchiarini et al., 2002). There is also a commonly held view in the wider community that some languages or language varieties "are spoken faster" than others, although the quantitative evidence to back this up is highly dependent on the particular measure of objective tempo (see below, and also see Kowal et al., 1983; den Os, 1988; Roach, 1998). Moreover, slowing down tempo does not always have a facilitative communicative role in certain communicative situations. For example, Derwing and Munro (2001) found that slowing down tempo excessively, or increasing the amount of pause time (e.g., Munro & Derwing, 1995) does not necessarily result in easier comprehension on the part of the second language learner (see also Bradlow & Bent, 2002, who suggested there is only a small clear speech effect for second language listeners).

4.1.1 Measuring tempo There are a number of measures that are used to monitor speaking tempo in prosody and timing studies. These include speaking rate or speech rate, which is the number of syllables uttered per second of total time available to the speaker (including pauses), and articulation rate, which is the number of syllables uttered per second of articulation speaking time minus pauses (e.g., Grosjean & Deschamps, 1975; Duez, 1981; Butcher, 1981; Fletcher, 1987; den Os, 1988), although tempo can also refer to articulation rate excluding pause (e.g., Quené, 2008). The most common stretch of speech over which to calculate articulation rate or speech rate is a stretch of speech between two pauses (see section 4.2 below for definitions of what constitutes a pause). For example, Miller et al. (1984), Crystal and House (1990), and Koopmans-van Beinum and van Donzel (1996) compared average syllable duration (ASD) per "run of pause-free speech" or "interpausal run," whereas Tauroza and Allison (1990) and Krause and Braida (2002) measured words per minute. Koiso et al. (1998) compare average duration of morae in inter-pause units in Japanese, which they call "sub-utterance units." Others have compared articulation rate (syllables/s) within prosodically defined units like prosodic words and intonational phrases as well as inter-pause stretches or "chunks" (e.g., Dankovičová, 1997; Fougéron & Jun, 1998), or have measured phone duration/second within inter-pause stretches (e.g., Trouvain et al., 2001) or both phone/s and syllable/s (e.g., Pfitzinger, 1999), phones and stresses per second (e.g., Fant et al., 1989), or onsets of vowels per second (e.g., Allen, 1972a, 1972b). Determining what constitutes fast versus slow tempo on the basis of these measures is problematic. What is often perceived by the listener as slow or fast speech, is not just a question of whether articulation rates are high or low. Intonational and prosodic variation and connected speech phenomena like assimilations, deletions, and re-syllabifications can influence perceived tempo (e.g., Goldman-Eisler, 1968; Butcher, 1981; Kohler, 1986; Roach, 1998). A related question is whether to use phonological syllables, phonetic syllables, phonemes, or phonetic segments as the countable unit in articulation rate calculation. This is of particular interest in cross-linguistic studies of tempo change (e.g., Dauer, 1983; Roach, 1998). Many studies have used phonological syllables/s, which presumes that the phonological structures of a language have been fully worked out.

The most popular measures tend to be based on syllable rates/s or ASD. Some representative studies are as follows. Mean articulation rates for French spoken at normal tempo have been reported at around 5.29 syll/s (Grosjean & Deschamps, 1975), 5.73 syll/s (Malécot et al., 1972), 5.9–6.2 syll/s (e.g., Fletcher, 1987, 1988), or 5.2–6.0 syll/s (Fougéron & Jun, 1998). Articulation rates of RP English (read or spontaneous speech) can range from 5.2 syll/s (e.g., White & Mattys, 2007b) to 5.9 syll/s (e.g., Dauer, 1983). A comparison of Orkney and Edinburgh English revealed mean articulation rates of 5.49 and 5.43 syll/s for read speech and 6.02 and 5.52 syll/s for spontaneous speech for each respective dialect (Hewlett & Rendall, 1998). Mean articulation rates for German read speech range from 5.3 syll/s (Trouvain, 2004), 5.84 (Butcher, 1981), to 6.04 syll/s (e.g., Künzel, 1997). Den Os (1988) compared speech rate in Italian and Dutch and showed that measures based on phonological syllables and phonetic segment succession per second coincide best with perceived tempo differences in both languages. Den Os (1988) found no significant differences in articulation rate between the two languages with articulation rates of 6.1 syll/s in Dutch and 6.4 syll/s in Italian for normal tempo, although significant ASD differences of 263 ms and 213 ms have been observed for Dutch spoken in the Netherlands compared to Flanders (Quené, 2008), backing up articulation rate differences (syllables/s) of around 16 percent observed by Verhoeven et al. (2004). Articulation rates for Greek range from 7.1–8.0 syll/s (e.g., Dauer, 1983), and for Spanish rates of between 6.1 and 8.0 syll/s (Dauer, 1983) and 8 syll/s (White & Mattys, 2007b) have been reported. Fant et al. (1991) report ASD values of 195 ms, 165 ms, and 215 ms for Swedish, French, and English, respectively. The percentage increase in articulation rate (syll/s) variation, where speakers are asked to increase their tempo from slow through to fast, also varies from 13–38 percent and 13–30 percent depending on the speaker (e.g., Fletcher, 1987; Fougéron & Jun, 1998, for French), whereas average articulation rate increases of around 24–25 percent were reported by Butcher (1981) for German, and by den Os (1988) for Dutch and Italian. Crystal and House (1990) similarly reported mean ASD values of 233 ms for a fast talker and 209 ms for a slow talker in their study of American English; suggesting, once again, that tempo is highly variable and speaker-dependent for the most part.

It has been suggested, however, that the faster succession of predominantly open syllables in languages like Spanish, for example, contributes to the perception of syllable timing in this language. Once again, there is a high level of inter (and intra-) speaker variation in most studies (e.g., Dauer, 1983), and these measures are not really sensitive enough to tap into the other structural features of a language, outlined earlier in this chapter, that can influence the succession of syllables within inter-pausal chunks or structural units like tone groups or intonational phrases. Verhoeven et al. (2004) also suggested that cross-linguistic comparisons of speech rates based on syllables per second or words per minute are difficult particularly when languages have different syllable and word structures, although it is fine to compare syllable rates within a language because structural differences are largely minimized (at least in terms of syllables or words). For these reasons, some cross-linguistic studies opt to measure phonemes per second or phones per second (e.g., Osser & Peng, 1964, cited in Roach, 1998) to circumvent structural differences among languages. For example, Osser and Peng (1964) compared American English and Japanese and found no differences using a measure based on phones/s. Measures based on phones/s also do not suggest a major difference between stress-timed languages and syllable-timed languages (see Roach, 1998, for discussion of this), although den Os (1988) observed significant differences between Italian and Dutch for measures based on phonemes/s, with higher values for Dutch versus Italian (e.g., 14.2 phonemes/s compared to 12.0 phonemes/s at normal tempo).

Irrespective of which measure is ultimately used, any quantitative study needs to differentiate "subjective tempo" (the intended tempo of an utterance) from "objective" tempo (the measurement thereof), and subjective, perceived tempo (after Butcher, 1981; den Os, 1988; Trouvain, 2004). Articulation rate and speaking rate (both measured as syll/s) have both been found to correlate with perceived tempo differences (e.g., van Bezooijen, 1984). Quené (2007) examined the JND for tempo variation within ten-second fragments of read speech for Dutch. The fragments were accelerated uniformly from 80% to 95% or decelerated similarly between 105% and 120%. He reported JND values of around 5% for perceived tempo variation, although intra-speaker speech rate variation also often exceeds 5%. Kato et al. (2003) also found that the ability of Japanese listeners to discriminate different speaking rates within four-mora words is extremely high, with slightly lower JNDs of 3.5%.

It is also important to distinguish global from local tempo or dynamic rate variation, although the use of these terms can be quite confusing (see Dankovičová, 1997, for a good discussion). Many of the tempo measures outlined above can be thought of as global because they effectively collapse micro-temporal variation within the domain across which articulation rate is calculated, i.e., the inter-pausal stretch, intonational phrase, or tone group. However, speaking rate or articulation rate in earlier studies (e.g., Goldman-Eisler, 1968; Grosjean & Deschamps, 1975), was usually measured across entire speech corpora, and these measures have also been termed “global” compared to those that restrict the domain of articulation rate calculation (e.g., Miller et al., 1984; Crystal & House, 1990; Pfitzinger, 1999; Kato et al., 2003). Local rate effects therefore refer to ASD or average phoneme duration measures within restricted domains or some other kind of measure like P-center succession (e.g., see Pompino-Marschall et al., 1982; Scott, 1998). Recall from section 2.2 that preboundary lengthening can be modeled as a local slowing down of tempo in the phrase-final syllable rhyme (e.g., Edwards et al., 1991; Byrd & Saltzman, 1998). This is precisely the kind of effect that is often called a local rate effect (e.g., den Os, 1988; Fant et al., 1991). In fact many of the duration influencing variables we have surveyed in this chapter could be classed as “local” rate effects in this respect. Variation in articulation rate depends on text-specific factors (i.e., number of stresses or accentual prominences, prosodic phrasing, and so on) as well as speaker-specific factors. At the end of the day it is important to make clear that rates of syllable or segment succession are *measures* of temporal patterns that refer to information structure, prosodic structure, speech style, or subjective adjustment of speaking tempo.

4.2 Pause

Pause makes up an important component of the temporal structure of speech and pauses contribute to the perception of global tempo (Goldman-Eisler, 1968; van Bezooijen, 1984). Pauses are often defined as either filled or unfilled/silent. Silent pauses show no voiced component in the acoustic waveform and some researchers distinguish short intra-segmental or so called “articulatory pauses” with an upper threshold of around 100 ms (after Butcher, 1981), corresponding to the closure phase of a voiceless stop, from inter-lexical pauses that tend to be longer (e.g., Zellner, 1994). Filled pauses are disfluencies that consist generally of voiced material that can correspond to elongated single vowels like “uh” in English, or portions of syllables (see Shriberg, 2001, for a good survey). Depending on the type or corpus analyzed (read speech, news broadcast speech, spontaneous speech), the percentage of pause time, as opposed to speaking time, can vary from 15–30% (Fant & Kruckenberg, 1989, for Swedish), from 16%–35% (Butcher, 1981, for German), and from 6%–38% (e.g., Fletcher, 1988, for French read speech; see also the comprehensive summary in Trouvain, 2004, p. 8). In a classic psycholinguistic study of pausing, Goldman-Eisler (1968) determined that silent intervals of between 200 and 250 ms are perceived as audible pauses, and 200 ms seems to be a threshold measurement that has been used in subsequent studies of pausing (Grosjean & Deschamps, 1975; Grosjean, 1980; Fletcher, 1987; Zellner, 1994). Intervals of either 100 ms, 130 ms, or 150 ms, or less in the case of “articulatory” pauses, have been measured by others (e.g., Butcher, 1981; Hieke et al., 1983; Fant & Kruckenberg, 1989; Dankovičová, 1997; Tsao & Weismer, 1997; Hansson, 2002; Wennerstrom & Siegel, 2003; Dankovičová et al., 2004; Trouvain, 2004). Butcher (1981) described three broad pause frequency distributions correlating with how listeners perceived them: “inaudible” pauses of between 100 and 200 ms, short pauses of between 500 and 600 ms, and long pauses of between

1,000 and 1,200 ms. Trouvain (2004) also found median pause durations of between 400 and 600 ms in German, supporting earlier findings by Butcher (1981). Kirsner et al. (2002) used Log Normal Distribution (LND) and defined two modes of distribution with means of 67 ms and 579 ms with a 173 ms threshold. Campione and Véronis (2002) similarly analyzed pause durations in a corpus of read and spontaneous French speech. They suggested that pauses can be grouped into three main categories: brief (< 200 ms), medium (200–1,000 ms). and long (> 1,000 ms). One advantage of the LND is that it supposedly avoids arbitrary decisions regarding pause duration classification (e.g., Kirsner et al., 2002).

Pauses occur in connected speech for a number of reasons. They may relate to respiration and the need to inhale during spoken communication, and much has been written about the "juncture," "planning," or "cognitive" functions of pauses in speech production and speech processing (e.g., Goldman-Eisler, 1968; Butcher, 1981; Levelt, 1989; Ferreira, 1993; Dankovičová et al., 2004), and their role in talk and interaction (e.g., Couper-Kuhlen, 1993). The relationship between syntax and pausing has been investigated by a number of researchers, and pauses tend to occur at major syntactic boundaries (e.g., Oller, 1973; Klatt, 1975; Cooper & Paccia-Cooper, 1980; Butcher, 1981; Ferreira, 1991; Horne et al., 1995; Yang, 2007). In many of these studies, these pauses (as well as other junctural phenomena including preboundary lengthening) also tend to coincide with boundaries of higher-level prosodic constituents like intonation phrases or tone groups (e.g., Butcher, 1981; Gee & Grosjean, 1983; Ferreira, 1993; Krivokapić, 2007). Moreover, there also appears to be a correlation between pause duration and prosodic boundary strength (Strangert, 1991; Zellner, 1994; Choi, 2003). Pauses also tend to be longer at the end of major discourse segments or units (e.g., Hirschberg & Nakatani, 1996; Swerts, 1997) and are longest at the boundary of discourse paragraphs (e.g., Strangert, 1991) or at places of topic shift in the spoken discourse (e.g., Smith, 2004). Some report a positive correlation between phrase length and pause duration (e.g., Grosjean & Collins, 1979; Zvonik & Cumminsm, 2002, 2003). For example, short pauses of less than 300 ms separate short phrases of less than 10 syllables (e.g., Zvonik & Cumminsm, 2003) and longer phrases are separated by longer pauses (Grosjean & Collins, 1979; Krivokapić, 2007).

Pausing also gives the speaker time to plan an upcoming utterance (Goldman-Eisler, 1968; Butcher, 1981; Levelt, 1989). Ferreira (1991) showed that speech "planning-based" pauses are longer before more complex syntactic material, whereas what she terms "timing-based" pauses (after already spoken material), tend to reflect prosodic structure. There is also a relationship between pause placement, prosodic structure, and syntactic disambiguation across a range of languages (e.g., Price et al., 1991; Jun, 2003). In general, tasks that require greater cognitive load on the speaker or that require them to perform a more complex task other than reading from a prepared script result in longer pauses (e.g., Goldman-Eisler, 1968; Grosjean & Deschamps, 1975; Butcher, 1981). For example, Grosjean and Deschamps (1975) found that pauses are more than twice as long during description tasks (1,320 ms) than during interviews (520 ms), and Grosjean (1980) also reported lower articulation rates during more cognitively demanding tasks. Some language-specific differences have also been suggested in earlier studies. For example, Grosjean and Deschamps (1975) found a difference between the incidence of "non-breath" and "breath" (i.e., respiratory) pauses in their corpus comparing English and French radio interviews. Apparently the French speakers inserted more respiratory pauses in their speech than the English speakers, as a proportion of all pauses in general. A difference between the two languages was also evident with respect to inter-pausal run, i.e., the average length (measured in either syllables or ms) of stretches of speech between pauses. There was a silent pause every 9.7 syllables in the English corpus, whereas the French data averaged a pause per 14.85 syllables. Subsequent studies of American English by Crystal and House (1990) reported average inter-pausal stretches of between 6.0 and 10.7 syllables, showing similar levels of variability reported by Miller et al.'s (1984) re-analysis of the original Grosjean and Deschamps (1975) corpus.

Subjective studies of tempo variation are generally based on read speech and require participants to self-select their speaking tempo (i.e., slow, normal, fast) or to speak in time with a metronome which is manipulated accordingly. Much of the increase in speaker tempo from slow through to fast is manifested as a reduction in pause time in relation to total speaking time (e.g., Grosjean & Deschamps, 1975; Grosjean, 1980; Butcher, 1981; Fletcher, 1987; den Os, 1988; Trouvain, 2004), although pause behavior alone does not necessarily account for perceived tempo variation as suggested in earlier studies (e.g., Goldman-Eisler, 1968). Modifying tempo also involves modifying the amount of articulation time (e.g., Grosjean & Deschamps, 1975; Butcher, 1981; Crystal & House, 1990; Edwards et al., 1991). Pause variables (i.e., pause occurrence, pause type, pause length) are also not always "utilized" in symmetrical ways to speed up or to slow down (e.g., Butcher, 1981). Speakers tend to insert more pauses when slowing down tempo but the pauses themselves are not always longer (e.g., Butcher, 1981; Grosjean, 1980). Speeding up reading tempo can result in fewer pauses and/or reduced pause duration (e.g., Grosjean, 1980; Butcher, 1981; Fletcher, 1987; Fougéron & Jun, 1998; Trouvain, 2004). Most pauses during fast (read) speech are physiological or respiratory pauses, with reduced numbers of "juncture" pauses (e.g., Grosjean & Lane, 1977; Butcher, 1981). These varying strategies reflect at the very least a high degree of inter-speaker variability, as well as "text"-specific factors, like prosodic phrasing and information structure as mentioned above. For example, the reduced number of juncture pauses in earlier studies also relates to prosodic restructuring at fast tempo. A number of studies based on read speech show that subjects produce fewer major or minor prosodic groupings at fast rates of speech (e.g., Vaissière, 1983; Fletcher, 1988; Jun, 1996; Fougéron & Jun, 1998; Trouvain, 2004). In conversational speech, pauses tend to be somewhat more variable in duration and placement, but some studies have shown that there is a complex relationship between length of pause and turn-taking. Wennerstrom and Siegel (2003) found that "turn shift" in conversation was more likely after either short pauses (lower than 500 ms) or long pauses (around 1.5 s), but pauses of intermediate length were more likely to occur mid-turn, supporting earlier research on pausing in talk and interaction (e.g., Jefferson, 1988; Wilson & Zimmerman, 1986).

4.3 Tempo: Lengthenings and shortenings

There is a certain amount of consensus in the timing literature that when speakers slow down or speed up speaking tempo, speech segment durations are lengthened or shortened (e.g., Peterson & Lehiste, 1960; Lindblom, 1963; Kozhevnikov & Chistovich, 1965; Nooteboom & Slis, 1972; Gay, 1978; Port, 1981; Crystal & House, 1982, 1990; Tuller et al., 1982; Kohler, 1986; den Os, 1988; Engstrand, 1988; Flege, 1988; Fletcher, 1988; Fourakis, 1991; Vatikiotis-Bateson & Kelso, 1993; Janse et al., 2003). Many of these studies are based on the study of controlled experimental tokens placed in focal position in a phrase, whereas others are based on the comparison of more extensive read material (e.g., Fletcher, 1987; den Os, 1988; Crystal & House, 1990). For the most part, tempo-related articulation rate variation does not always manifest itself as a simple horizontal time compression or expansion of acoustic intervals corresponding to consonant or vowel segments. Between normal and fast tempi, most of the above studies agree that vowel segments shorten more than consonants with compression values of around 30 percent (e.g., Fourakis, 1991, for American English, although Port, 1981, also found that when speakers are asked to slow down their neutral speaking tempo, all consonant and vowel segments increased in duration by a more or less constant ratio).

There is also a general claim that long segments shorten more than short segments at fast rates (e.g., den Os, 1988; Crystal & House, 1982). For example, at fast rates, phonologically long vowels shorten more than phonologically short vowels in some languages that have vowel quantity contrasts, and conversely at slow tempo, phonologically long vowels stretch more than short vowels and contribute more to increased overall word duration (e.g., Nooteboom & Slis, 1969; Port, 1981; Hirata, 2004). Others have reported more or less equal degrees of shortening in long and short vowels due to increased tempo (see, e.g., Magen & Blumstein, 1993, for Korean). Speaking rate can also influence the perception of vowel quantity in some languages. At fast tempi in English, the degree of vowel lengthening before voiced obstruents is reduced (e.g., Port et al., 1980; Smith, 2002), although there is no difference in shortening of tense and lax vowels (e.g., Gay, 1978). Different types of prosodic factors also interact with tempo in a variety of ways. Peterson and Lehiste (1960) and Port (1981) found that stressed syllable durations in their American English data are less affected than unstressed syllable durations by increases in tempo, although Gay (1978) and Tuller et al. (1982) showed that unstressed and stressed vowel durations undergo a similar amount of compression from slow to fast tempi. Later studies that control for stress versus intonational accent have found a more complex relationship. For example, at fast rates, stressed syllables in Dutch reduce to 64 percent of their normal rate duration, whereas unstressed syllables reduce to 45 percent (Janse et al., 2003). Interestingly, syllables carrying sentence stress (i.e., nuclear accent) also shorten at fast rates reducing the durational difference between accented and unaccented syllables. In other words, the duration cue to nuclear accent is less robust at fast rates. Janse et al. explained these differences in terms of the different cues to lexical stress and nuclear accent. It is imperative to maintain the durational correlates of the lexical stress contrast at fast rate, whereas other cues like f_0 excursion cue sentence stress at all rates (see also Fougéron & Jun, 1998; Ladd et al., 1999; Caspers, 2003, for specific examination of some of the interactions between speech rate and intonational variation). Lehiste (1970) also claimed that some languages reduce stressed syllable durations less than unstressed syllable durations, while others spread the effects of tempo increase in fast speech. This view has received support from subsequent studies, including Janse et al. (2003). Fourakis (1986) also showed that stressed and unstressed Greek vowels shorten by the same amounts (25 percent) from slow to fast tempo. Accented vowels in French vowels compress by 28 percent and unaccented vowels by 24 percent at fast speech rates from slow through to fast tempo, although intonational phrase-final vowels shorten less (Fletcher, 1988; see also Pasdeloup, 1990). Fougéron and Jun (1998) also found that the longest syllables (intonational phrase-final) shorten *less* than accentual-phrase final or nonfinal syllables at fast rates for two out of three speakers. Smith (2002) also noted that intonational phrase-final vowels in English resist temporal compression compared to nonfinal vowels, although other studies show that the degree of shortening of phrase-final accented and unaccented syllables (articulatory duration) varies among speakers at fast rates (e.g., Edwards et al., 1991). In other words, the durational contrast at the phrase edge is maintained at fast rates.

Hirata (2004) showed that while duration ratios of short to long vowels are affected by rate changes in Japanese, overall word duration ratios and vowel-to-word ratios are little affected. Port et al. (1987) also showed that at fast tempi, morae in longer Japanese words shorten proportionately more than morae in short words. These results are reminiscent of elements of Kozhevnikov and Chistovich's (1965) theory of rhythmic invariance outlined earlier in this review. They proposed that changing articulation rate due to tempo increase does not affect the relative durations of syllables within a *syntagma* or a word, even though consonant and vowel durations are affected. This notion of relative invariance lay at the heart of earlier studies of tempo and other types of temporal effects (e.g., Nooteboom & Slis, 1969; Lindblom & Rapp, 1973; Port et al., 1980; Kohler, 1986; Pickett et al., 1999). For example, Nooteboom and Slis (1969) tested this theory on Dutch nonsense word data and found that the relative durations of syllables in words uttered at fast tempo did not differ from those at neutral tempo. Similar results are reported by Kohler (1986) for German, with the relative durations of two- and three-syllable feet remaining invariant across tempi. Lindblom and Rapp (1973, p. 27) also write of the "occasional incompatibility between the rhythm and tempo of syllable imitation and sluggishness in articulatory performance." The degree of temporal compression that takes place within a speech unit such as a word or stress foot is often in conflict with "a mechanism aiming at keeping the durations of acoustic segments large enough to obviate extensive changes in movement rate and the associated neural control" (Lindblom & Rapp, 1973, p. 28). There is still a degree of sympathy for the Kozhevnikov and Chistovich view of temporal organization in speech, particularly at a more abstract level (see, e.g., Kohler, 2003).

4.3.1 Tempo: The fine phonetic detail In addition to the kinds of segmental and syllabic duration effects summarized in the previous section, there is a range of articulatory or subsegmental changes that take place as a result of speeding up or slowing down tempo. Many early studies of speech rate effects on vowel and consonant articulation were influenced by Lindblom's (1963) model of durational undershoot (see section section 2.1, and also outlined in detail in section 2.5 in Harrington, this volume), with later studies influenced by the H&H model of listener-oriented communication (e.g., Lindblom, 1983, 1990; Moon & Lindblom, 1994), and models of articulatory dynamics (see, e.g., Kelso et al., 1985; Saltzman & Munhall, 1989; Browman & Goldstein, 1990; Byrd & Saltzman, 1998). Recall that Lindblom (1963) suggested that irrespective of whether vowel duration is determined by stress or tempo, formant undershoot will occur in shorter segments due to mechanical inertia of the articulators in response to the articulatory demands of surrounding consonants. This model of vowel production has been much applauded for its elegance and simplicity, and articulatory or spectral undershoot or reduced articulatory gesture magnitude is indeed observed at fast speech rates (e.g., Lindblom, 1964; Kent & Moll, 1972; Flege, 1988; Vatikiotis-Bateson & Kelso, 1993; Moon & Lindblom, 1994). It is not the only consequence of changing speaking rate however. Gay (1978) found that his American English-speaking subjects were able to maintain full vowel quality in unstressed position at a rapid speaking rate when requested to do so. Vowel formant targets were also consistently reached in stressed fast rates, even though they were often of similar duration to unstressed vowels in slow readings. Vowel formant transitions were also less vulnerable to shortening than steady state portions at fast rates. In addition, Tsao et al. (2006) showed that the vowel space of habitually fast talkers (of English spoken in the upper Midwest of the USA), is not "compressed" relative to the vowel space of habitually slow talkers, although the latter show more variability in their F1/F2 vowel space which does not necessarily correlate with durational variation. Van Son and Pols (1992) also found that formant frequencies were not significantly affected at fast rates in Dutch, with the exception of overall F1 raising, but there was little evidence of rate-induced formant undershoot and relatively small levels of shortening in both stressed and unstressed vowels compared to other studies with no major difference between the two types of vowel. One should also recall that Dutch vowels do not reduce as much as English vowels in unstressed contexts, so this may also have been a factor (see section 2.1). As also observed by Fourakis (1991) for American English, greater shortening effects were attributed to stress than speeding up tempo.

Kuehn and Moll (1976) also showed that potential trade-offs exist between strategies of articulatory displacement (in the case of undershooting an articulatory target) and articulatory velocity (increasing the speed of movement) at faster rates. They found that some speakers increased velocity and therefore their articulations exhibited relatively little undershoot. Other speakers maintained constant velocity while showing decreased articulatory displacement at fast rates. These differing strategies were also observed by Hertrich and Ackermann (2000) for German, and Edwards et al. (1991) for American English. Engstrand (1988) found that labial gestures overlap more with vowel gestures at fast speech rates, which suggests that an additional strategy, involving changing inter-gestural timing or phasing, can also be associated with speech rate variation. Shaiman (2001) drew a similar conclusion, suggesting that while changes in vowel duration due to speech rate are mainly brought about by either changes to lip and jaw velocity or displacement, changing inter-gestural timing is a further strategy (see also Munhall & Löfqvist, 1992). Consonant clusters and consonant sequences show consistent durational shortening at fast rates, but less consistent patterns of either increased articulatory overlap or spatial modification (e.g., Hardcastle, 1985; Byrd & Tan, 1996; Ellis & Hardcastle, 2002). Almost all articulatory studies that examine tempo as a separate factor report high levels of inter-speaker variability, in line with studies of more global measures based on syllables per second or pausing. Articulatory studies of tempo effects and prosodic prominence or stress in languages like English mostly concur that changing stress or level of prominence and articulation rate are not "equivalent motor transformations" (Tuller et al., 1982, p. 1541; see also Ostry & Munhall, 1985). Articulatory strategies associated with stress and accentuation can be more consistent, compared to speech rate (e.g., Tuller et al., 1982; Beckman & Edwards, 1994), although high levels of inter-speaker variability are also observed (e.g., recall the discussion in section 2.1). Speakers regularly adapt their production strategies depending on the needs of a particular communicative situation (see, e.g., Schulman, 1989; Moon & Lindblom, 1994; Janse et al., 2003; and the discussion in Smiljanic & Bradlow, 2008).

5 Concluding Comments

As a result of this vast body of research on prosody and timing, we now know a lot more about universal and language-specific segmental and syllable duration patterns, and have greater insights into the nature of articulatory timing in relation to prosody and prosodic structure. The list of experiments on prosody and speech timing on a range of different languages grows yearly. But do we know anything more about prosody and timing as a result of five decades or more of experimental phonetic research? The answer to this question is yes, but there is still more work to be done, particularly on less well studied languages, and on different varieties within a language. It is also important to bear in mind that the temporal signatures of prosody do not always manifest themselves as a simple squeezing or expanding of particular acoustic intervals. The continuing development of larger more varied speech corpora, as well as the refinement of articulatory monitoring devices and further developments in speech perception research will enable closer examination of the phonetic detail of spoken language. Prosody and the temporal organization of speech will continue to be the subject of a great deal of experimental phonetic research in the years to come.

REFERENCES

Abercrombie, D. (1964) Syllable quantity and enclitics in English. In D. Abercrombie, F. Fry, P. A. D. McCarthy, N. C. Scott, & J. Trim (eds.), *In Honour of Daniel Jones*. (pp. 216–22). London: Longmans.

Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Allen, G. D. (1972a) The location of rhythmic stress beats in English: An experimental study, I. *Language and Speech*, 15, 72–100.

Allen, G. D. (1972b) The location of rhythmic stress beats in English: An experimental study, II. *Language and Speech*, 15, 179–95.

Allen, G. D. (1975) Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3, 75–86.

Arvaniti, A. (1991) Rhythmic categories: A critical evaluation on the basis of Greek data. In *Proceedings of the 12th International Congress of Phonetic Sciences*, vol. 2 (pp. 298–301). Université de Provence: Service des Publications.

Arvaniti, A. (1992) Secondary stress: Evidence from modern Greek. In G. J. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 398–419). Cambridge: Cambridge University Press.

Arvaniti, A. (1994) Acoustic features of Greek rhythmic structure. *Journal of Phonetics*, 22, 239–68.

Arvaniti, A. (2009) Rhythm, timing and the timing of rhythm. *Phonetica*, 66, 46–63.

Astruc, L. & Prieto, P. (2006) Acoustic cues of stress and accent in Catalan. In R. Hoffmann & H. Mixdorff (eds.), *Proceedings of Speech Prosody 2006* (pp. 337–40). Dresden: TUD Press.

Asu, E. L. & Nolan, F. (2005) Estonian rhythm and the pairwise variability index. *Proceedings, FONETIK 2005*, Department of Linguistics, Göteborg University, 29–32.

Asu, E. L. & Nolan, F. (2006) Estonian and English rhythm: A two-dimensional quantification based on syllables and feet. In R. Hoffmann & H. Mixdorff (eds.), *Proceedings of Speech Prosody 2006*. (pp. 249–52). Dresden: TUD Press.

Asu, E. L. & Nolan, F. (2009) The pairwise variability index and coexisting rhythms in languages. *Phonetica*, 66, 29–45.

Auer, P. (1993) Is a rhythm-based typology possible? A study of the role of prosody in phonological typology. *KontRI Working Paper* 21, University of Hamburg.

Avesani, C. & Vayra, M. (2005) Accenting, deaccenting and information structure in Italian dialogue. In *SIGdial6-2005*, 19–24.

Bailly, G., Benoit, C., & Sawalis, T. R. (1992) *Talking Machines: Theories, Models, and Designs*. Amsterdam: North-Holland.

Barbosa, P. (2007) From syntax to acoustic duration: A dynamical model of speech rhythm production. *Journal of Phonetics*, 49, 725–42.

Barnwell, T. P. (1971) *An algorithm for segment durations in a reading machine context*. *Massachusetts Institute of Technology Research Laboratory of Electronics Technical Report*, 479.

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003) Do rhythm measures tell us anything about language type? In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2693–6). Barcelona/ Australia: Causal Productions.

Barry, W. J., Andreeva, B., & Steiner, I. (2007) The phonetic exponency of phrasal accentuation in French and German. *Proceedings of Interspeech 2007*, Antwerp, 1010–3.

Barry, W. J., Andreeva, B., & Koreman, J. (2009) Do rhythm measures reflect perceved rhythm? *Phonetica*, 66, 78–94.

Bartkova, K. (1991) Speaking rate modelization in French application to speech synthesis. *Proeedings of the International Congress of Phonetic Sciences*, Aix-en-Provence, 3, 482–5.

Baumann, S., Becker, J., Grice, M., & Mücke, D. (2007) Tonal and articulatory marking of focus in German. In J. Trouvain & W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1029–32.

Beckman, M. E. (1982) Segment duration and the "mora" in Japanese. *Phonetica*, 39, 113–35.

Beckman, M. E. (1986) *Stress and Non-Stress Accent*, Netherlands Phonetic Archives 7. Dordrecht: Foris.

Beckman, M. E. (1992) Evidence for speech rhythm across languages. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 457–63). Amsterdam: IOS Press.

Beckman, M. E. (1996) The parsing of prosody. *Language and Cognitive Processes*, 11, 17–67.

Beckman, M. E. (1997) Speech models and speech synthesis. In J. P. van Santen, R. Sproat, J. Olive, & J. Hirschberg (eds.), *Progress in Speech Synthesis* (pp. 185–209). New York: Springer.

Beckman, M. E. & Bretonnel Cohen, K. (2000) Modeling the articulatory dynamics of two levels of stress contrast. In M. Horne (ed.), *Prosody: Theory and Experiment: Studies Presented to Gösta Bruce* (pp. 169–200). Dordrecht: Kluwer.

Beckman, M. E. & Edwards, J. (1990) Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of speech.* (pp. 152–78). Cambridge: Cambridge University Press.

Beckman, M. E. & Edwards, J. (1992) Intonational categories and the articulatory control of duration. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 356–75). Amsterdam: IOS Press.

Beckman, M. E. & Edwards, J. R. (1994) Articulatory evidence for differentiating stress categories. In P. A. Keating (ed.), *Phonological Structure and Phonetic Form: Phonology and Phonetic Evidence* (pp. 7–33). Cambridge: Cambridge University Press.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992) Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 68–86). Cambridge: Cambridge University Press.

Beckman, M. E. & Pierrehumbert, J. (1986) Intonational structure in English and Japanese. *Phonology Yearbook*, 3, 255–310.

Benguerel, A. (1971) Duration of French vowels in unemphatic stress. *Language and Speech*, 14, 383–91.

Bergem, D. R. van (1993) Acoustic vowel reduction as a function of sentence accent, word stress and word class. *Speech Communication*, 12, 1–23.

Berkovits, R. (1994) Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, 37, 237–50.

Bertinetto, P. M. (1981) *Strutture prosodiche dell' italiano.* Firenze: Accademia della Crusca.

Bertinetto, P. M. (1989) Reflections of the dichotomy stress versus syllable timing. *Revue de phonétique appliqué*, 91–3, 99–130.

Bertinetto, P. M. & Fowler, C. A. (1989) On sensitivity to durational modifications in Italian and English. *Rivista di Linguistica*, 1, 69–94.

Bezooijen, R. van (1984) *Characteristics and Recognizability of Vocal Expressions and Emotions*, Netherland Phonetic Archives 5. Dordrecht: Foris.

Bishop, J. (2002) Stress accent without phonetic stress. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 179–82.

Blevins, J. (1995) The syllable in phonological theory. In J. A. Goldsmith (ed.), *The Handbook of Phonological Theory* (pp. 206–44). Oxford: Blackwell.

Bochner, J. H., Snell, K. B., & MacKenzie, D. J. (1988) Duration discrimination of speech and tonal complex stimuli by normally hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 84, 493–500.

Bombien, L., Mooshammer, C., Hoole, P., Rathcke, T., & Kühnert, B. (2007) Articulatory strengthening in initial German /kl/ clusters under prosodic variation. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 457–60.

Boudreault, M. (1970) Le rythme en langue franco-canadienne. In P. R. Léon, G. Faure, & A. Rigault (eds.), *Prosodic Feature Analysis/Analyse des faits prosodiques*, Studia Phonetica 3 (pp. 21–31). Montreal/Paris/Brussels: Didier.

Bradlow, A. (2002) Confluent talker- and listener-oriented forces in clear speech production. In C. Gussenhoven & N. Warner (eds.) *Laboratory Phonology 7.* (pp. 237–74). Berlin: Mouton de Gruyter.

Bradlow, A. & Bent, T. (2002) The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, 112, 272–84.

Bradlow, A., Kraus, N., & Hayes, E. (2003) Speaking clearly for children with learning disabilities: Sentence perception in noise. *Journal of Speech, Hearing, and Language Research*, 46, 80–97.

Browman, C. P. & Goldstein, L. (1990) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (eds.), *Papers in Laboratory Phonology, I: Between the Grammar and Physics of Speech* (pp. 341–76). Cambridge: Cambridge University Press.

Browman, C. P. & Goldstein, L. (1993) Dynamics and articulatory phonology. In T. van Gelder & B. Port (eds.), *Mind as Motion* (pp. 175–93). Cambridge, MA: MIT Press.

Bruce, G. (1977) *Swedish Word Accents in Sentence Perspective*. Lund: CWK Gleerup.

Bruce, G. (1983) On rhythmic alternation. *Working Papers of Lund*, 25, 25–52.

Bruce, G. (2005) Intonational prominence in varieties of Swedish revisited. In. S-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 410–29). Oxford: Oxford University Press.

Butcher, A. (1981) Aspects of the speech pause: Phonetic correlates and communicative functions. *Aipuk* (Arbeitsberichte Institut für Phonetik Kiel), 15.

Butcher, A. & Harrington, J. (2003) An instrumental analysis of focus and juncture in Warlpiri. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, 321–4.

Byrd, D. (2000) Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, 57, 3–16.

Byrd, D., Krivokapić, J., & Lee, S. (2006) How far, how long: On the temporal scope of prosodic boundary effects. *Journal of the Acoustical Society of America*, 120, 1589–99.

Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005) Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America*, 118, 3860–73.

Byrd, D. & Saltzman, E. (1998) Intragestural dynamics of multiple phrasal boundaries. *Journal of Phonetics*, 26, 173–99.

Byrd, D. & Saltzman, E. (2003) The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31, 149–80.

Byrd, D. & Tan, C. C. (1996) Saying consonant clusters quickly. *Journal of Phonetics*, 24, 263–82.

Cambier-Langeveld, T. (1997) The domain of final lengthening in the production of Dutch. In H. de Hoop & J. Coerts (eds.), *Linguistics in the Netherlands* (pp. 13–24). Amsterdam: John Benjamins.

Cambier-Langeveld, T. (2000) Temporal marking of accents and boundaries. Doctoral dissertation, University of Amsterdam, Holland Institute of Generative Linguistics, Netherlands Graduate School of Linguistics.

Cambier-Langeveld, T., Nespor, M., & Heuven, V. van (1997) The domain of final lengthening in production and perception in Dutch. *Proceedings of Eurospeech 1997*, 931–4.

Cambier-Langeveld, T. & Turk, A. (1999) A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics*, 27, 255–80.

Campbell, N. (1992a) Segmental elasticity and timing in Japanese speech. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production, and Linguistic Structure* (pp. 403–18). Amsterdam:IOS Press.

Campbell, N. (1992b) Syllable-based segmental duration. In. G. Bailly, C. Benoit, & T. R. Sawalis (eds.), *Talking Machines: Theories, Models and Designs* (pp. 209–10). Amsterdam: North-Holland.

Campbell, N. & Isard, S. (1991) Segment durations in a syllable frame. *Journal of Phonetics* 19, 37–47.

Campbell, N. & Beckman, M. E. (1997) Stress, prominence, and spectral tilt. *Proceedings of the ESCA Workshop on Intonation: Theory, Models, and Applications*. Athens, Greece. 67–70.

Campione, I. & Véronis, J. (2002) A large-scale multilingual study of silent pause duration. *Proceedings of Eurospeech 2002*, 199–202.

Carlson, R. & Granström, B. (1986) A search for durational rules in a real-speech data base. *Phonetica*, 43, 140–54.

Caspers, J. (2003) Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, 31, 251–76.

Chahal, D. (2001) Modeling the intonation of Lebanese Arabic using the Autosegmental-Metrical Framework: A comparison with English. Doctoral dissertation, University of Melbourne.

Chahal, D. (2003) Phonetic cues to prominence in Arabic. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2067–70). Barcelona/Australia: Causal Productions.

Chen, Y. (2006) Durational adjustment under corrective focus in Standard Chinese. *Journal of Phonetics*, 34, 176–201.

Cho, T. (2004) Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, 32, 141–76.

Cho, T. (2005) Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a, i/ in English. *Journal of the Acoustical Society of America*, 117, 3867–78.

Cho, T. (2006) Manifestation of prosodic structure in articulation: Evidence from lip kinematics in English. In L. M. Goldstein, D. H. Whalen, & C. T. Best (eds.), *Laboratory Phonology 8: Varieties of Phonological Competence* (pp. 519–48). Berlin/New York: Mouton de Gruyter.

Cho, T. & Keating, P. (2001) Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, 29, 155–90.

Cho, T. & Keating, P. (2007) Effects of initial position versus prominence in

English. *UCLA Working Papers in Phonetics*, 106, 1–33.

Cho, T. & McQueen, J. (2005) Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, 33, 121–57.

Cho, T., McQueen, J., & Cox, E. (2007) Prosodically driven detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, 35, 210–43.

Choi, H. (2003) Prosody-induced acoustic variation in English stop consonants. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2661–4). Barcelona/Australia: Causal Productions.

Classe, A. (1939) *The Rhythm of English Prose*. Oxford: Blackwell.

Coker, C. H., Umeda, N., & Browman, C. P. (1973) Automatic synthesis from text. *IEEE Transactions, Audio Electroacoustics*, AU-21, 293–7.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007) Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from radio news speech. *Journal of Phonetics*, 35, 180–209.

Coleman, J. (1992) Synthesis-by-rule without segments or rewrite rules. In G. Bailly, C. Benoit, & T. R. Sawalis (eds.), *Talking Machines: Theories, Models, and Designs* (pp. 43–60). Amsterdam: North-Holland.

Cooper, W. E. & Danly, M. (1981) Segmental and temporal aspects of utterance-final lengthening. *Phonetica*, 38 106–15.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985) Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, 77, 2142–56.

Cooper, W. E. & Paccia-Cooper, J. (1980) *Syntax and Speech*. Cambridge, MA: Harvard University Press.

Cooper, W. E. & Sorensen, J. M. (1977) Fundamental frequency contours at syntaxtic boundaries. *Journal of the Acoustical Society of America*, 62, 683–92.

Couper-Kuhlen, E. (1993) *English Speech Rhythm*. Philadelphia, PA: John Benjamins.

Crompton, A. (1980) Timing patterns in French. *Phonetica*, 37, 205–34.

Crosswhite, K. (2003a) Spectral tilt as a cue to word stress in Polish, Macedonian, and Bulgarian. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 767–70). Barcelona/Australia: Causal Productions.

Crosswhite, K. (2003b) Vowel reduction. In B. Hayes, R. Kirchner, & D. Steriade (eds.), *Phonetically Based Phonology* (pp. 191–231). Cambridge: Cambridge University Press.

Crystal, D. (1969) *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Crystal. D. (1975) *The English Tone of Voice*. London: Edward Arnold.

Crystal, T. H. & House, A. S. (1982) Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, 72, 705–16.

Crystal, T. H. & House, A. S. (1988a) Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, 83, 1553–73.

Crystal, T. H. & House, A. S. (1988b) Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, 83, 1574–85.

Crystal, T. H. & House, A. S. (1990) Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, 88, 101–12.

Cucchiarini, C., Strik, H., & Boves, L. (2002) Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111, 2862–73.

Cummins, F. (2002) Speech rhythm and rhythmic taxonomy. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 121–6.

Cummins, F. (2003) Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31, 139–48.

Cummins, F. (2009) Rhythm as an affordance for the entrainment of movement. *Phonetica*, 37, 16–28.

Cummins, F. & Port, R. (1998) Rhythmic constraints on "stress-timing" in English. *Journal of Phonetics*, 26, 145–71.

Cutler, A. (1981) The reliability of speech error data. *Linguistics*, 19, 561–82.

Cutler, A. (2005) Lexical stress. In. D. Pisoni & R. Remez (eds.), *The Handbook of Speech Perception* (pp. 264–89). Oxford: Blackwell.

Cutler A. & Butterfield, S. (1992) Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218–36.

Cutler, A. & Carter, D. M. (1987) The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133–42.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986) The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385–400.

Cutler A. & Norris, D. (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–21.

Dankovičová, J. (1997) The domain of articulation rate in Czech. *Journal of Phonetics*, 25, 287–312.

Dankovičová, J. & Dellwo, V. (2007) Czech speech rhythm and the rhythm class hypothesis. In J. Trouvain & W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, 1241–4.

Dankovičová, J., Piggot, K., Wells, B., & Peppé, S. (2004) Temporal markers of prosodic boundaries in children's speech production. *Journal of the International Phonetic Association*, 34, 17–36.

Dauer, R. M. (1983) Stress-timing and syllable-timing re-analysed. *Journal of Phonetics*, 11, 51–62.

Dauer, R. M. (1987) Phonetic and phonological components of language rhythm. *Proceedings of the 11th International Congress of Phonetic Sciences*, Tallinn, Estonia, 5, 447–50.

Delattre, P. (1962) Some factors of vowel duration and their cross-linguistic validity. *Journal of the Acoustical Society of America*, 34, 1141–3.

Delattre, P. (1965) *Comparing the Phonetic Features of English, French, German, and Spanish: An Interim Report*. Heidelberg: Julius Groos.

Delattre, P. (1966) A comparison of syllable length conditioning among languages. *IRAL*, 4, 183–98.

Delattre, P. (1969) An acoustic and articulatory study of vowel reduction in four languages. *IRAL*, 7, 295–325

Dellwo, V. (2004) The Bonn Tempo-Corpus & Bonn Tempo-Tools: A database for the study of speech rhythm and rate. In *Proceedings of the 8th ICSLP*, Jeju Island, Korea, 4 pp.

Dellwo, V. (2006) Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski. & I. Szigeti (eds.), *Language and Language-Processing* (pp. 231–41). Frankfurt: Peter Lang.

Dellwo, V., Fourcin, A. & Abberton, E. (2007) Rhythmical classification of languages based on voice parameters. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1129–32.

Dellwo, V. & Wagner, P. (2003) Relationships between speech rhythm and rate. In M. J. Solé, D. Recasens, &

J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 471–4). Barcelona/Australia: Causal Productions.

Derwing, T. & Munro, M. (2001) What speaking rates do non-native listeners prefer? *Applied Linguistics*, 22, 324–37.

Deterding, D. (2001) The measurement of rhythm: A comparison of Singapore and British English. *Journal of Phonetics*, 29, 217–30.

Di Cristo, A. (1999) Vers une modélisation de l'accentuation du francais: première partie. *French Language Studies*, 9, 143–79.

Di Cristo, A. (2000) Vers une modélisation de l'accentuation du français: seconde partie. *French Language Studies*, 10, 27–44.

Di Cristo, A. & Hirst, D. J. (1997) L'accentuation non-emphatique en français: stratégies et paramètres. in J. Perrot (ed.), *Polyphonie pour Ivan Fónagy* (pp. 71–101). Paris: L'Harmattan, Paris.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996) Glottalization of vowel-initial syllables as a function of prosodic structure. *Journal of Phonetics*, 24, 423–44.

D'Imperio, M. & Gili-Favela, B. (2004) How many levels of phrasing? Evidence from two varieties of Italian. In J. Local, R. Ogden, & R. Temple (eds.), *Phonetic Interpretation, Papers in Laboratory Phonology VI* (pp. 130–44). Cambridge: Cambridge University Press.

D'Imperio, M. & Rosenthall, S. (1999) Phonetics and phonology of main stress in Italian. *Phonology*, 16, 1–28.

Dogil, G. (1999) The phonetic manifestation of word stress in Lithuanian, Polish, and German and Spanish. In H. van der Hulst (ed.), *Word Prosodic Systems of the Languages of Europe* (pp. 273–310). Berlin: Mouton de Gruyter.

Donovan, A. & Darwin, C. J. (1979) The perceived rhythm of speech. *Proceedings of the 9th International Congress of the Phonetic Sciences*, Copenhagen, 2, 268–72.

Duez, D. (1981) Pauses silencieuses et pauses non-silencieuses dans trois types de message oraux. *Travaux de l'Institut de Phonétique d'Aix*, 8, 85–114.

Duez, D. (1982) Silent and non-silent pauses in three speech styles. *Language and Speech*, 25, 11–28.

Duez, D. & Nishinuma, Y. (1985) Some evidence on rhythmic patterns of spoken French. *PERILUS III*, University of Stockholm, Institute of Linguistics, 30–40.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991) The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89, 369–82.

Eefting, W. (1988) Temporal variation in natural speech: Some explorations. In *Proceedings of Speech '88: 7th FASE Symposium*, 503–7.

Eefting, W. (1991) The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments. *Journal of the Acoustical Society of America*, 89, 412–24.

Eek, A. & Help, T. (1987) The interrelationship between phonological and phonetic sound changes: A great rhythm shift of Old Estonian. *Proceedings of the 11th Congress of Phonetic Sciences*, Tallinn, Estonia, 6, 218–33.

Elert, C.-C. (1964) *Phonologic Studies of Quantity in Swedish*. Uppsala: Almqvist & Wiksell.

Ellis, L. & Hardcastle, W. J. (2002) Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data. *Journal of Phonetics*, 30, 373–96.

Engstrand, O. (1988) Articulatory correlates of stress and speaking rate in Swedish CV utterances. *Journal of the Acoustical Society of America*, 88, 1863–75.

Engstrand, O. & Krull, D. (1994) Durational correlates of quantity in Swedish, Finnish and Estonian: Cross language evidence for a theory of adaptive dispersion. *Phonetica*, 51, 80–91.

Erickson, D. (1998) Effects of contrastive emphasis on jaw opening. *Phonetica*, 55, 147–69.

Fant, G. & Kruckenberg, A. (1989) Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, 2, 1–83.

Fant, G., Kruckenberg, A., & Nord, L. (1989) Stress patterns, pauses and timing in prose reading. *STL-QPSR*, 1, 7–12.

Fant, G., Kruckenberg, A., & Nord, L. (1991) Durational correlates of stress in Swedish, French and English. *Journal of Phonetics*, 19, 351–65.

Farnetani, E. (2000) Hyper- to hypo-articulated vowels: Articulatory-acoustic relations. *Proceedings of the 5th Seminar on Speech Production Models and Data. Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Models*. Kloster Seeon, Bavaria, 217–20.

Farnetani, E. & Kori, S. (1986) Effects of syllable and word structure on segmental durations in spoken Italian. *Speech Communication*, 5, 17–34.

Faure, G., Hirst, D., & Chafcouloff, M. (1980) Rhythm in English: Isochronism, pitch, and perceived stress. In L. R. Waugh & C. H. van Schooneveld (eds.), *The Melody of Language* (pp. 71–9). Baltimore, MD: University Park Press.

Fear, B., Cutler, A., & Butterfield, S. (1995) The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America*, 1893–904.

Ferreira, F. (1991) Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30, 210–33.

Ferreira, F. (1993) Creation of prosody during sentence production *Psychological Review*, 2, 233–53.

Ferreira, F. (2000) Syntax in language production: An approach using tree-adjoining grammars. In L. Wheeldon (ed.), *Aspects of Language Production* (pp. 291–330). Cambridge, MA: MIT Press.

Fischer-Jørgensen, E. (1954) Acoustic analysis of stop consonants. *Miscellanea Phonetica*, 2, 42–59.

Flege, J. E. (1988) Effects of speaking rate on tongue position and velocity of movement in vowel production. *Journal of the Acoustical Society of America*, 84, 901–16.

Fletcher, J. (1987) Some micro and macro effects of tempo change on timing in French. *Linguistics*, 25, 951–67.

Fletcher, J. (1988) An acoustic study of timing in French. Doctoral dissertation, University of Reading.

Fletcher, J. (1991) Rhythm and final lengthening in French. *Journal of Phonetics*, 19, 193–212.

Fletcher, J. & Butcher, A. (2003) Local and global influences on vowel formants in three Australian languages. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (905–8). Barcelona/Australia: Causal Productions.

Fletcher, J. & Evans, N. (2002) An acoustic intonational study of intonational prominence in two Australian languages. *Journal of the International Phonetic Association*, 32, 123–40.

Fletcher, J. & McVeigh, A. (1993) Segment and syllable duration in Australian English. *Speech Communication*, 13, 355–65.

Fletcher, J. & Vatikiotis-Bateson, E. (1994) Prosody and intrasyllabic timing in French. *Ohio State University Working Papers in Linguistics*, 43, 41–6.

Fónagy, I. (1979) L'accent français: accent probabilitaire. *Studia Phonetica*, 15, 123–33.

Fougéron, C. (2001) Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, 29, 109–36.

Fougéron, C. & Jun, S.-A. (1998) Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics*, 26, 45–69.

Fougéron, C. & Keating, P. A. (1997) Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 101, 3728–40.

Fourakis, M. (1986) An acoustic study of the effects of tempo and stress on segmental intervals. *Phonetica*, 43, 172–88.

Fourakis, M. (1991) Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, 90, 1816–27.

Fourakis, M., Botinis, A., & Katsaiti, M. (1999) Acoustic characteristics of Greek vowels. *Phonetica*, 56, 28–43.

Fowler, C. A. (1981) A relationship between coarticulation and compensatory shortening. *Phonetica*, 38, 35–40.

Fowler, C. A. & Housum, J. (1987) Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26, 489–504.

Fowler, C. A., Smith, M. R., & Tassinary, L. G. (1986) Perception of syllable timing by prebabbling infants. *Journal of the Acoustical Society of America*, 79, 814–25.

Fraisse, P. (1963) *The Psychology of Time*. New York: Harper & Row.

Frota, S. & Vigário, M. (2001) On the correlates of rhythmic distinction: The European/Brazilian Portuguese case. *Probus*, 13, 247–75.

Fry, D. B. (1955) Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27, 765–8.

Fry, D. B. (1958) Experiments in the perception of stress. *Language and Speech,* 1, 126–52.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975) Auditory perception of duration of speech and non-speech stimuli. In G. Fant & M. A. A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 197–219). London: Academic Press.

Fujisaki, H. & Sudo, H. (1971) Synthesis by rule of prosodic features of connected Japanese. *Proceedings of the 7th International Congress on Acoustics*, Akadémiai Kiadó, Budapest, 3, 133–6.

Garde, P. (1968) *L'accent.* Paris: Presses Universitaires de France.

Gay, T. (1978) Effect of speaking rate on vowel formant structures. *Journal of the Acoustical Society of America*, 63, 223–30.

Gay, T. (1981) Mechanisms in the control of speech rate. *Phonetica*, 38, 148–58.

Gee, J. P. & Grosjean, F. E. (1983) Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15, 411–58.

Gendrot, C. & Adda-Decker, M. (2007) Impact of vowel inventory size on formant values of oral vowels: An automated formant analysis from eight languages. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1417–20.

Gibbon, D. & Gut, U. (2001) Measuring speech rhythm in varieties of English. *Proceedings of EUROSPEECH 2001*, Aalborg, 91–4.

Gibbon, D. & Romani Fernandes, F. (2005) Annotation-mining for rhythm model comparison in Brazilian Portuguese. *Proceedings of Interspeech/Eurospeech 2005*, 329–32.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992) SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 517–20.

Goldman-Eisler, F. (1968) *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Gordon, M. (2005) Intonational phonology of Chickasaw. In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 301–330). Oxford: Oxford University Press.

Grammont, M. (1946) *Traité pratique de pronunciation francaise*. Paris: Librairie Delagrave.

Grabe, E. & Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In. C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7* (pp. 515–43). Berlin: Mouton de Gruyter.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003) Temporal properties of spontaneous speech: A syllable centric perspective. *Journal of Phonetics*, 31, 465–85.

Grosjean, F. (1979) A study of timing in a manual and a spoken language: American sign language and English. *Journal of Psycholinguistic Research*, 8, 379–405.

Grosjean, F. (1980) Comparative studies of temporal variables in spoken and sign languages: A short review. In H. Dechert & M. Raupach (eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler* (pp. 307–12). The Hague: Mouton.

Grosjean, F. & Collins, M. (1979) Breathing, pausing and reading. *Phonetica*, 36, 98–114.

Grosjean, F. & Deschamps, A. (1975) Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica* 31, 144–84.

Grosjean, F. & Lane, H. (1977) Pauses and syntax in American Sign Language. *Cognition*, 5, 101–17.

Grosz, B. & Hirschberg, J. (1992) Some intonational characteristics of discourse structure. In *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff, 429–32.

Gussenhoven, C. (2004) *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C. (2009) Vowel quantity, syllable duration, and stress in Dutch. In K. Hanson & S. Inkelas (eds.), *The Nature of the Word: Essays in Honour of Paul Kiparsky* (pp. 181–98). Cambridge, MA: MIT Press.

Gussenhoven, C. & Rietveld, A. C. M. (1992) Intonation contours, prosodic structure and preboundary lengthening. *Journal of Phonetics*, 20, 283–303.

Gut, U. (2003) Non-native speech rhythm in German. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2437–40). Barcelona/Australia: Causal Productions.

Gut, U. & Milde, J.-T. (2002) The prosody of Nigerian English. *Proceedings of Speech Prosody 2002*, Aix-en-Provence pp. 367–70.

Haggard, M. (1973)Abbreviation of consonants in English pre-vocalic and post-vocalic clusters. *Journal of Phonetics*, 1, 9–24.

Hajek, J., Stevens, M., & Webster, G. (2007) Vowel duration, compression, and lengthening in stressed syllables in Italian. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1057–60.

Hakokari, J, Saarni, T., Salakoski, T., Isoaho, J., & Aaltonen, O. (2007) Measuring relative articulation rate in Finnish utterances. In J. Trouvain & W. J. Barry (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1105–8.

Han, M. (1962) The feature of duration in Japanese. *Onsei no Kenkyuu*, 10, 65–80.

Hansson, P. (2002) Prosodic phrasing and articulation rate variation. *Proceedings of Fonetik, TMH-QPSR*, 44, 173–6.

Hardcastle, W. J. (1985) Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication*, 4, 247–63.

Harrington, J. (2003) Consonant strengthening and lengthening in various languages: Comments on three papers. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in Laboratory Phonology 6: Phonetic Interpretations* (pp. 183–93). Cambridge: Cambridge University Press.

Harrington, J., Fletcher, J., & Beckman, M. E. (2000) Manner and place conflicts in the articulation of accent in Australian English. In. M. Broe & J. Pierrehumbert (eds.), *Papers in Laboratory Phonology 5: Language Acquisition and the Lexicon* (pp. 40–51). Cambridge: Cambridge University Press.

Harrington, J., Fletcher, J., & Roberts, C. (1995) Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics*, 23, 305–22.

Harris, K. S. (1978) Vowel duration changes and its underlying physiological mechanisms. *Language and Speech*, 21, 354–61.

Harris, M. S. & Umeda, N. (1974) Effect of speaking mode on temporal factors in speech: Vowel duration. *Journal of the Acoustical Society of America*, 56, 1016–18.

Hay, J., Sato, M., Coren, A., Moran, C., & Diehl, R. (2006) Enhanced contrast for vowels in utterance focus: A cross language study. *Journal of the Acoustical Society of America*, 119, 3022–33.

Hayes, B. (1989) The prosodic hierarchy in meter. In P. Kiparsky and G. Youmans (eds.), *Rhythm and Meter* (pp. 201–60). Orlando, FL: Academic Press.

Hayes, B. (1995) *Metrical Stress Theory: Principles and Case Studies.* Chicago: The University of Chicago Press.

Heldner, M. & Strangert, E. (2001) Temporal effects of focus in Swedish. *Journal of Phonetics*, 29, 329–61.

Hertrich, I. & Ackermann, H. (1997) Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures. *Journal of the Acoustical Society of America*, 102, 523–36.

Hertrich, I. & Ackermann, H (2000) Lip–jaw and tongue–jaw coordination during rate-controlled syllable repetitions. *Journal of the Acoustical Society of America*, 107, 2236–47.

Hewlett, N. & Rendall, M. (1998) Rural versus urban accent as an influence on the rate of speech. *Journal of the International Phonetic Association*, 28, 63–71.

Hieke, A., Kowal, S., & O'Connell, M. (1983) The trouble with "articulatory" pauses. *Language and Speech*, 26, 203–14.

Hirata, Y. (2004) Effects of speech rate on vowel length distinction in Japanese. *Journal of Phonetics*, 32, 565–89.

Hirschberg, J. & Nakatani, C. (1996) A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the ACL*, Santa Cruz, 286–93.

Hirst, D. J. (1999) The symbolic coding of duration and timing: An extension to the INTSINT system. *Proceedings Eurospeech '99*, 1639–42.

Hoequist, C., Jr. (1983a) Syllable duration in stress-, syllable-, and mora-timed languages. *Phonetica*, 40, 203–37.

Hoequist, C., Jr. (1983b) Durational correlates of linguistic rhythm categories. *Phonetica*, 40, 19–31.

Horne, M., Strangert, E., & Heldner, M. (1995) Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 170–3.

House, A. S. (1961) On vowel duration in English. *Journal of the Acoustical Society of America*, 33, 1174–8.

Huggins, A. W. F. (1972) Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America*, 51, 1270–8.

Huggins, A. W. F. (1975) On isochrony and syntax. In G. Fant & M. A. A. Tatham (eds.), *Auditory Analysis and the*

Perception of Speech (pp. 455–64). Orlando, FL: Academic Press.

Janse, E., Nooteboom, S., & Quene, H. (2003) Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication*, 41, 287–301.

Jassem, W. (1959) The phonology of Polish stress. *Word*, 15, 252–69.

Jefferson, G. (1988) Notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In D. Roger & P. Bull (eds.), *Conversation: An Interdisciplinary Perspective* (pp. 166–96). Clevedon, UK: Multilingual Matters.

Jessen, M. (1993) Stress conditions on vowel quality and quantity in German. *Working Papers of the Cornell Phonetics Laboratory*, 8, 1–27.

Johnson, K. & Martin, J. (2001) Acoustic vowel reduction in Creek: Effects of distinctive length and position in the word. *Phonetica*, 58, 81–102.

Jong, K. J. de (1991) An articulatory study of consonant-induced vowel duration changes in English. *Phonetica*, 48, 1–17.

Jong, K. J. de (1995) The supraglottal articulation of prominence in English: Linguistic stress as localised hyperarticulation. *Journal of the Acoustical Society of America*, 97, 491–504.

Jong, K. J. de (2004) Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics*, 32, 493–516.

Jong, de K. J., Beckman, M. E., & Edwards, J. R. (1993) The interplay between prosody and coarticulation. *Language and Speech*, 36, 197–212.

Jong, K. J. de & Zawaydeh, B. A. (2002) Comparing stress, lexical focus, and segmental focus: Patterns of variation in Arabic vowel duration. *Journal of Phonetics*, 30, 53–75.

Joos, M. (1948) Acoustic phonetics. *Language*, 24, 1–136.

Jun, S.-A. (1996) *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure*. New York: Garland.

Jun, S.-A. (1998) The accentual phrase in the Korean prosodic hierarchy. *Phonology*, 15, 189–226.

Jun, S.-A. (2003) Prosodic phrasing and attachment preferences. *Journal of Psycholinguistic Research*, 32, 219–49.

Jun, S.-A. (2005) Prosodic typology. In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 430–58). Oxford: Oxford University Press.

Jun, S.-A. & Fougéron, C. (2002) Realization of accentual phrase in French intonation. *Probus*, 14, 147–72.

Kaiki, N. & Sagisaka, Y. (1992) The control of segmental duration in speech synthesis using statistical methods. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production, and Linguistic Structure* (pp. 391–402). Tokyo: Ohmsa.

Kaiki, N., Takeda, K., & Sagisaka, Y. (1992) Linguistic properties in the control of segmental duration for speech synthesis. In G. Bailly, C. Benoit, & T. R. Sawalis (eds.), *Talking Machines* (pp. 255–63). Amsterdam: North-Holland.

Kato, H, Tsuzaki, M., & Sagisaka, Y. (1998) Acceptability for temporal modification of single vowel segments in isolated words. *Journal of the Acoustical Society of America*, 104, 540–9.

Kato, H., Tsuzaki, M., & Sagisaka, Y. (2003) Functional differences between vowel onsets and offsets in temporal perception of speech: Local change detection and speaking rate discrimination. *Journal of the Acoustical Society of America*, 113, 3379–89.

Keane, E. (2006) Rhythmic characteristics of formal and colloquial Tamil. *Language and Speech*, 49, 299–332.

Keating, P. A. (1984) Phonetic and phonological representation of stop consonant voicing. *Language*, 60, 286–319.

Keating, P. A. (2006) Phonetic encoding of prosodic structure. In. J. Harrington & M. Tabain (eds.), *Speech Production*. (pp. 167–83). New York: Psychology Press.

Keating, P. A., Cho, T., Fougéron, C., & Hsu, C. (2003) Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in Laboratory Phonology 6: Phonetic Interpretations* (pp. 145–63). Cambridge: Cambridge University Press.

Kelso, J. A. S. (1995) *Dynamic Patterns*. Cambridge, MA: MIT Press.

Kelso J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay, B. (1985) A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics and dynamic modelling. *Journal of the Acoustical Society of America*, 77, 266–80.

Kent, R. & Moll, K. L. (1972) Cinefluorographic analyses of selected lingual consonants. *Journal of Speech and Hearing Research*, 15, 453–73.

Kent, R. & Netsell, R. (1971) Effects of stress contrasts on certain articulatory parameters. *Phonetica*, 24, 23–44.

Kirsner, K., Dunn, J., Hird, K., Parkin, T., & Clark, C. (2002) Time for a pause . . . In *Proceedings of the Ninth Australian Conference on Speech Science and Speech Technology*, Melbourne, 52–7.

Klatt, D. H. (1975) Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics*, 3, 129–40.

Klatt, D. H. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59, 1208–21.

Klatt, D. H. (1979) Synthesis by rule of segmental durations in English sentences. In. B. Lindblom & S. Öhmann (eds.), *Frontiers of Speech Communication Research* (pp. 287–300). New York: Academic Press.

Klatt, D. H. (1987) Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82, 737–93.

Klatt, D. H. & Cooper, W. E. (1975) Perception of segment duration in sentence contexts. In A. Cohen & S. Nooteboom (eds.), *Structure and Process in Speech Perception* (pp. 69–89). Heidelberg: Springer.

Kleber, F. & Klipphahn, N. (2006) An acoustic investigation of secondary stress in German. *AIPUK* (Arbeitsberichte Institut für Phonetik Kiel), 37, 1–18.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005) Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118, 1038–54.

Kohler, K. J. (1983) Prosodic boundary signals in German. *Phonetica*, 40, 89–134.

Kohler, K. J. (1986) Invariance and variability in speech timing: From utterance to segment in German. In J. S. Perkell & D. H. Klatt (eds). *Invariance and Variability in Speech Processes* (pp. 268–89). Hillsdale, NJ: Lawrence Erlbaum.

Kohler, K. J. (2003) Domains of temporal control in speech and language: From utterance to segment. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 7–10). Barcelona/Australia: Causal Productions.

Kohler, K. J. (2009) Rhythm in speech and language. *Phonetica*, 66, 29–45.

Koiso, H., Shimojima, A., & Katagiri, Y. (1998) Collaborative signaling of informational structures by dynamic speech rate. *Language and Speech*, 41, 323–50.

Koopmans-Van Beinum, F. J. (1980) Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions. Doctoral Dissertation, Academishe Pers B. V., Amsterdam.

Koopmans-Van Beinum, F. J. & Donzel, M. E. van (1996) Discourse structure and its influence on local speech rate. *Proceedings* (Institute of Phonetic Sciences, Amsterdam University) 20, 1–11.

Kowal, S., Wiese, R., & O'Connell, D. C. (1983) The use of time in storytelling. *Language and Speech*, 26, 377–92.

Kozhevnikov, V. A. & Chistovich, L. A. (1965) *Speech Articulation and Perception*. Washington, DC: Joint Publications Research Service.

Krause, J. & Braida, L. (2002) Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America*, 112, 2165–72.

Krivokapicć, J. (2007) Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, 35, 162–79.

Kroos, C., Hoole, P., Kühnert, B., & Tillmann, H. G. (1997) Phonetic evidence for the phonological status of the tense–lax distinction in German. *Forschungsberichte des FIPKM* (Instituts für Phonetik und Sprachliche Kommunikation der Universität München), 35, 17–26.

Krull, D. (1997) Prepausal lengthening in Estonian: Evidence from conversational speech. In I. Lehiste & J. Ross (eds.), *Estonian Prosody: Papers from a Symposium* (pp. 136–48).

Krull, D., Traunmüller, H., & Bertinetto, P. (2006) Local speaking rate and perceived quantity. An experiment with Italian listeners. Lund University Department of Linguistics and Phonetics, *Working Papers*, 52, 81–4.

Kuehn, D. P. & Moll, K. L. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4, 303–20.

Kubozono, H. (2002) Temporal neutralization in Japanese. In. C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7* (pp. 171–201). Berlin: Mouton de Gruyter.

Künzel, H. J. (1997) Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics*, 4, 48–83.

Kuzla, C., Cho, T., & Ernestus, M. (2007) Prosodic strengthening of German fricatives in duration and assimilatory voicing. *Journal of Phonetics*, 35, 301–20.

Ladd, D. R. (1996*) Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R. & Campbell, N. (1991) Theories of prosodic structure: Evidence from syllable duration. In *Proceedings of the 12th International Congress of Phonetic Sciences*, Aix-en-Provence, 2, 290–3.

Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A. (1999): Constant "segmental anchoring" of F0 movements under changes in speech rate. *Journal of the Acoustical Society of America*, 106, 1543–55.

Ladefoged, P. (1967) *Three Areas of Experimental Phonetics*. Oxford: Oxford University Press.

Lashley, K. S. (1951) The problem of serial order in behavior. In L. A. Jeffress (ed.), *Cerebral Mechanisms in Behavior* (pp. 112–36). New York: Wiley.

Lehiste, I. (1970) *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, I. (1972) The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, 51, 2018–24.

Lehiste, I. (1973) Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America*, 54, 1228–34.

Lehiste, I. (1975) The phonetic structure of paragraphs. In A. Cohen & S. Nooteboom (eds.), *Structure and Process in Speech Perception* (pp. 195–203). Berlin: Springer.

Lehiste, I. (1977) Isochrony reconsidered. *Journal of Phonetics*, 5, 253–63.

Lehiste, I. (1980) Interaction between test word duration and length of utterance. In L. R. Waugh & C. H. Schooneveld (eds.), *The Melody of Language* (pp. 169–76). Baltimore, MD: University Park Press.

Lehiste, I., Olive, J. P., & Street, L. (1976) The role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, 60, 1199–202.

Lehiste, I. & Peterson (1959) Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, 31, 428–35.

Lehtonen, J. (1970) *Aspects of Quantity in Standard Finnish*. Jyvaskyla: Jyvaskyla University Press.

Lenneberg, E. (1967) *Biological Foundations of Language*. London: Wiley.

Léon, P. & Martin, Ph. (1980) Des accents. In L. R. Waugh & C. H. van Schooneveld (eds.), *The Melody of Language* (pp. 177–85). Baltimore MD: University Park Press.

Levelt, C. & Vijver, R. van de (1998) Syllable types in cross-linguistic and developmental grammars. Paper presented at the Third Biannual Utrecht Phonology Workshop, 11–12 June.

Levelt, W. J. M. (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levitt, A. & Aydelott-Utman, J. G. (1992) From babbling towards the sound system of English and French: A longitudinal two-case study. *Journal of Child Language*, 19, 19–49.

Liberman, M. & Prince, A. (1977) On stress and linguistic rhythm, *Linguistic Inquiry*, 8, 249–336.

Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773–81.

Lindblom, B. (1964) A note on segment duration in Swedish polysyllables. *Speech Transmission Laboratory Quarterly Progress Status Report*, 2, 1–5.

Lindblom, B. (1967) Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress Status Report*, 4, 1–29.

Lindblom, B. (1968) Temporal organization of syllable production. *Speech Transmission Laboratory Quarterly Progress Status Report*, 2–3, 1–5.

Lindblom, B. (1975) Some temporal regularities of spoken Swedish. In G. Fant & M. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 387–96). New York: Academic Press.

Lindblom, B. (1983) The economy of speech gestures. In P. MacNeilage (ed.), *The Production of Speech* (pp. 217–46). New York: Springer.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H and H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 403–40). London: Kluwer.

Lindblom, B., Aguele, A., Sussman, H., & Eir Cortes, E. (2007) The effect of emphatic stress on vowel coarticulation. *Journal of the Acoustical Society of America*, 121, 3802–13.

Lindblom, B. & Rapp, K. (1973) Some temporal regularities of spoken Swedish. *Papers in Linguistics from the University of Stockholm*, 21, 1–59.

Lisker, L. & Abramson, A. S. (1967) Some effects of context on voice onset time in English stops. *Language and Speech*, 10, 1–28.

Local, J. & Ogden, R. (1996) A model of timing for nonsegmental phonological structure. In J. P. van Santen, R. Sproat, J. Olive, & J. Hirschberg (eds.), *Progress in Speech Synthesis* (pp. 109–22). New York: Springer.

Loevenbruck, H. (1999) An investigation of articulatory correlates of the accentual phrase in French. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 967–70). Berkeley, CA: University of California.

Low, E. L. (1998) Prosodic prominence in Singapore English. Unpublished doctoral dissertation, University of Cambridge.

Low, E. L., Grabe, E., & Nolan, F. (2000) Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, 43, 377–401.

Lyberg, B. (1979) Final lengthening: Partly a consequence of restrictions on the speed of fundamental frequency change? *Journal of Phonetics*, 7, 187–96.

Lyberg, B. (1981) Some consequences of a model for segment duration based on F0. *Journal of Phonetics*, 9, 97–103.

Magen, H. S. & Blumstein, S. E. (1993) Effects of speaking rate on the vowel length distinction in Korean. *Journal of Phonetics*, 21, 387–409.

Malécot, A., Johnston, R., & Kizziar, P.-A. (1972) Syllabic rate and utterance length in French. *Phonetica*, 26, 235–51.

Marcus, S. M. (1981) Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, 30, 247–56.

Martin, J. G. (1972) Rhythmic (hierarchical) versus serial structure in speech and other behaviour. *Psychological Review*, 79, 487–509.

Martin, Ph. (1975) Analyse phonologique de la phrase francaise. *Linguistics*, 146, 35–68.

Martin, Ph. (1982) Phonetic realizations of prosodic contrasts in French. *Speech Communication*, 1, 283–94.

Martin, Ph. (1987) Prosodic and rhythmic structures in French. *Linguistics*, 25, 925–49.

Mehler, J., Dehaene-Lambertz, G., Dupoux, E., & Nazzi, T. (1996) Coping with linguistic diversity: The infant's viewpoint. In J. Morgan & K. Demuth (eds.), *Signal to Syntax* (pp. 101–16). Hillsdale, NJ: Lawrence Erlbaum.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U., & Sequi, J. (1981) The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behaviour*, 20, 398–05.

Ménard, L., Loevenbruck, H., & Savariaux, C. (2006) Articulatory and acoustic correlates of contrastive focus in French children and adults. In J. Harrington & M. Tabain (eds.), *Speech Production* (pp. 225–51). New York: Psychology Press.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984) Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41, 215–25.

Möbius, B. & Santen, J. P. van (1996) Modeling segmental duration in German text-to-speech synthesis. *Proceedings of the International Conference on Spoken Language Processing* (Philadelphia, PA), 2395–8.

Moon, S. J. & Lindblom, B. (1994) Interaction between duration, context, and speaking style in English stressed vowels, *Journal of the Acoustical Society of America*, 96, 40–55.

Mooshammer, C. & Fuchs, S. (2002) Stress distinction in German: Simulating kinematic parameters of tongue tip gestures. *Journal of Phonetics*, 30, 337–55.

Morton, J., Marcus, S. M., & Frankish, C. (1976) Perceptual centres (P-centres). *Psychological Review*, 83, 405–8.

Munhall, K. & Lofqvist, A. (1992) Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics*, 20, 111–26.

Munro, M. J. & Derwing, T. M. (1995) Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38, 289–306.

Murty, L., Otake, T., & Cutler, A. (2007) Perceptual tests of rhythmic similarity, I: Mora rhythm. *Language and Speech*, 50, 77–99.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981) Prosodic aspects of American English speech rhythm. *Phonetica*, 38, 84–106.

Nespor, M. (1990) On the rhythm parameter in phonology. In I. M. Roca (ed.), *Logical Issues in Language Acquisition* (pp. 157–75). Dordrecht: Foris.

Nespor, M. & Vogel, I. (1986) *Prosodic Phonology*. Dordrecht: Foris.

Nooteboom, S. G. (1972) Production and perception of vowel duration; a study of durational properties of vowels in Dutch. Doctoral dissertation, Utrecht University.

Nooteboom, S. G. (1973) The perceptual reality of some prosodic durations. *Journal of Phonetics*, 1, 25–45.

Nooteboom, S. G. & Eefting, W. (1994) Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica*, 51, 92–8.

Nooteboom, S. G. & Slis, I. H. (1969) A note on rate of speech. *IPO Annual Progress Report*, 4, 58–60.

Nooteboom, S. G. & Slis, I. H. (1972) The phonetic feature of vowel length in Dutch. *Language and Speech*, 15, 301–16.

Nord, L. (1986) Acoustic studies of vowel reduction in Swedish. *Speech Transmission Laboratory Quarterly Progress Status Report*, 4, 19–36.

Nowak, P. M. (2006) Vowel reduction in Polish. Doctoral dissertation, University of California, Berkeley.

Ohala, J. (1975) The temporal regulation of speech. In G. Fant & M. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 431–54). London: Academic Press.

Öhman, S. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41, 310–20.

Oliver, D. & Grice, M. (2003) Phonetics and phonology of lexical stress in Polish verbs. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2027–30). Barcelona/Australia: Causal Productions.

Oller, D. K. (1973) The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, 54, 1235–47.

Oller, D. K. (1979) Syllable timing in Spanish, English, and Finnish. In H. Hollien & P. Hollien (eds.), *Current Issues in the Phonetic Sciences* (pp. 320–41), Amsterdam: John Benjamins.

Oller, D. K. (2000) *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum.

Ortega-Llebario, M. & Prieto, P. (2005) Disentangling stress from accent: Production patterns of the stress contrast in deaccented syllables. *Proceedings of PAPI (Phonetics and Phonology in Iberia) 2005*, Bellaterra, 23 pp.

den Os, E. (1988) Rhythm and tempo of Dutch and Italian: A contrastive study. Doctoral Dissertation, Rijksuniversiteit, Utrecht.

Osser, H. & Peng, F. (1964) A cross-cultural study of speech rate, *Language and Speech*, 7, 120–5.

Ostry, D. J. & Munhall, K. G. (1985) Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, 77, 640–8.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993) Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258–78.

Otake, T., Hatano, G., & Yoneyama, K. (1996) Speech segmentation by Japanese listeners. In T. Otake & A. Cutler (eds.), *Phonological Structure and Language Processing: Cross-Linguistic Studies* (pp. 183–201). Berlin: Mouton.

O'Shaughnessy, D. (1981) A study of French vowel and consonant durations. *Journal of Phonetics*, 9, 385–406.

O'Shaughnessy, D. (1984) A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America*, 76, 1664–72.

Padgett, J. & Tabain, M. (2005) Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica*, 62, 14–54.

Pamies-Bertran, A. (1999) Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design*, 2, 103–30.

Parmenter, C. E. & Trevino, S. N. (1936) Relative durations of stressed to unstressed vowels. *American Speech*, 10, 129–33.

Pasdeloup, V. (1990) Modèle de règles rythmiques du francais appliquées à la synthèse de la parole. Doctoral dissertation, Université de Provence.

Pasdeloup, V. (1992) A prosodic model for French text-to-speech synthesis: A psycholinguistic approach. In G. Bailly, C. Benoit, & T. R. Sawalis (eds.), *Talking Machines: Theories, Models and Designs* (pp. 335–48). Amsterdam: North-Holland.

Peng, S.-H. (1997) Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics*, 25, 371–400.

Peterson, G. E. & Lehiste, I. (1960) Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 32, 693–703.

Pfitzinger, H. R. (1999) Local speech rate perception in German speech In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences*, Berkeley, CA, University of California, 893–6.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986) Speaking clearly for the hard of hearing, II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29, 434–46.

Pickett, E. R., Blumstein, S. E., & Burton, M. W. (1999) Effects of speaking rate on the singleton/geminate consonant contrast in Italian. *Phonetica*, 56, 135–57.

Pierrehumbert, J. B. (1980) The phonology and phonetics of English intonation. Doctoral dissertation, Massachusetts Institute of Technology. (Distributed 1987 by Indiana University Linguistics Club.)

Pierrehumbert, J. & Talkin, D. (1992) Lenition of /h/ and glottal stop. In G. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 90–117). Cambridge: Cambridge University Press.

Pike, K. L. (1946) *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Pisoni, D. & Remez, R. (2005) *Handbook of Speech Perception*. Oxford: Blackwell.

Pointon, G. (1980) Is Spanish really syllable-timed? *Journal of Phonetics*, 8, 293–304.

Pompino-Marschall, B. (1989) On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, 17, 175–92.

Pompino-Marschall, B., Piroth, H. G., Hoole, P., Tilk, K., & Tillmann, H. G. (1982) Does the closed syllable determine the perception of "momentary tempo"? *Phonetica*, 39, 358–67.

Port, R. F. (1981) Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, 69, 262–74.

Port, R. F. (2003) Meter and speech. *Journal of Phonetics*, 31, 599–611.

Port, R. F., Al-Ani, S., & Maeda, S. (1980) Temporal compensation and universal phonetics. *Phonetica*, 37, 235–52.

Port, R. F., Dalby, J., & O'Dell, M. (1987) Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, 81, 1574–85.

Post, B. (2000) *Tonal and Phrasal Structures in French Intonation*. The Hague: Thesus.

Potisuk, S., Gandour, J., & Harper, M. P. (1996) Acoustic correlates of stress in Thai. *Phonetica*, 53, 200–20.

Price, P. J., Ostendorf, S., Shattuck-Hufnagel, S., & Fong, C. (1991) The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90, 2956–70.

Quené, H. (1992) Durational cues for word segmentation in Dutch. *Journal of Phonetics*, 20, 331–50.

Quené, H. (2007) On the just noticeable difference for tempo in spontaneous speech. *Journal of Phonetics*, 35, 353–62.

Quené, H. (2008) Multilevel modelling of between-speaker and within-speaker variation in spontaneous tempo. *Journal of the Acoustical Society of America*, 123, 1104–13.

Rakerd, B., Sennet, W., & Fowler, C. A. (1987) Domain-final lengthening and foot-level shortening in spoken English. *Phonetica*, 44, 147–55.

Ramus, F. (2002) Acoustic correlates of linguistic rhythm: Perspectives. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 115–20.

Ramus, F., Dupoux, E., & Mehler, J. (2003) The psychological reality of rhythm classes: Perceptual studies. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 337–42). Barcelona/Australia: Causal Productions.

Ramus, F. & Mehler, J. (1999) Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105, 512–21.

Ramus, F., Nespor, M., & Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265–92.

Remijsen, B. (2002) Lexically contrastive stress accent and lexical tone in Ma'ya. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology 7* (pp. 585–614). Berlin: Mouton de Gruyter.

Remijsen, B. & Heuven, V. van (2005) Stress, tone, and discourse prominence in the Curaçao dialect of Papiamentu. *Phonology*, 22, 205–35.

Rietveld, T., Kerkhoff, J., & Gussenhoven, G. (2004) Word prosodic structure and vowel duration in Dutch. *Journal of Phonetics*, 32, 349–71.

Rigault, A. (1962) Rôle de la fréquence, de l'intensité, et de la durée vocalique dans la perception de l'accent en français. *Proceedings of the Fourth International Congress of Phonetic Sciences*, Helsinki, 1961. The Hague: Mouton.

Riley, M. D. (1992) Tree-based modelling of segmental duration. In G. Bailly, C. Benoit, & T. R. Sawalis (eds.), *Talking Machines: Theories, Models and Designs* (pp. 265–74). Amsterdam: North-Holland.

Roach, P. (1982) On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal (ed.), *Linguistic Controversie: Essays in Linguistic Theory and Practice in Honour of F. R. Palmer* (pp. 73–9). London: Edward Arnold.

Roach, P. (1998) Some languages are spoken more quickly than others. In L. Bauer & P. Trudgill (eds.), *Language Myths* (pp. 150–8). Harmondsworth: Penguin.

Rossi, M., Di Cristo, A., Hirst, D., Martin, Ph., & Nishinuma, Y. (1981) *L'intonation: de l'acoustique à la sémantique*. Paris: Klincksieck.

Sagisaka, Y. (1992) On the modeling of segmental duration control. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production, and Linguistic Structure* (pp. 451–5). Tokyo: Ohmsa.

Sagisaka, Y., Campbell, N., & Higuchi, N. (1997) *Computing Prosody: Computational Models for Processing Spontaneous Speech*. New York: Springer.

Saltzman, E. & Munhall, K. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1, 333–82.

Santen, J. P. van (1992a) Contextual effects on vowel duration. *Speech Communication*, 11, 513–46.

Santen, J. P. van (1992b) Deriving text-to-speech durations from natural speech. In G. Bailly, C. Benoit, & T. R. Sawalis (eds.), *Talking Machines: Theories, Models and Designs* (pp. 275–86). Amsterdam: North-Holland.

Santen, J. P. van & Olive, P. (1990) The analysis of contextual effects on segmental duration. *Computer, Speech and Language*, 4, 359–91.

Santen, J. P. van & Shih, C. (2000) Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America*, 107, 1012–26.

Santen, J. P. van, Sproat, R., Olive, J., & Hirschberg, J. (1996) *Progress in Speech Synthesis*. New York: Springer.

Schötz, S. (2006) Perception, analysis, and synthesis of speaker age. *Travaux de l'Institut de Linguistique de Lund*, 47, 1–183.

Schulman, R. (1989) Articulatory dynamics and loud and normal speech. *Journal of the Acoustical Society of America*, 85, 295–312.

Scott, D. R. (1982) Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, 71, 996–1007.

Scott, S. (1998) The point of P-Centres. *Psychological Research*, 61, 4–11.

Scott, D. R., Isard, S., & de Boysson-Bardies, B. (1985) Perceptual isochrony in English and in French. *Journal of Phonetics*, 13, 155–62.

Selkirk, E. (1984) *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Shaiman, S. (2001) Kinematics of compensatory vowel shortening: The effect of speaking rate and coda composition on intra and inter-articulator timing. *Journal of Phonetics*, 29, 89–107.

Shattuck-Hufnagel, S. & Turk, A. E. (1996) A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25, 193–247.

Shriberg, E. (2001) To "err" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31, 153–69.

Sluijter, A. M. C. (1995) *Phonetic correlates of stress and accent*. Doctoral dissertation, Leiden University, The Netherlands.

Sluijter, A. M. C. & Heuven, V. J. van (1995) Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, 52, 71–89.

Sluijter, A. M. C. & Heuven, V. J. van (1996) Spectral balance as an acoustic correlate of the linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471–85.

Smiljanic, R. & Bradlow, A. (2008) Stability across temporal contrasts across speaking styles: English and Croatian. *Journal of Phonetics*, 36, 91–113.

Smith, B. L. (2002) Effects of speaking rate on temporal patterns of English. *Phonetica*, 59, 232–44.

Smith, C. (2004) Topic transitions and durational prosody in reading aloud: Production and modeling. *Speech Communication*, 42, 247–70.

Son, R. J. J. H. van & Pols, L. C. W. (1992) Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America*, 92, 121–7.

Stetson, R. (1951) *Motor Phonetics: A Study of Speech Movements in Action*, 2nd edn. Amsterdam: North-Holland.

Stone, M. (1981) Evidence of a rhythm pattern in speech production: Observations of jaw movement. *Journal of Phonetics*, 9, 109–20.

Strangert, E. (1985) Swedish speech rhythm in a cross-language perspective. *Acta Universitatis Umensis*, Umeå Studies in the Humanities, 69.

Strangert, E. (1991) Pausing in texts read aloud. *Proceedings of the 12th International Congress of Phonetic sciences*, Aix-en-Provence, 238–241.

Summers, W. V. (1987) Effects of stress and final consonant voicing on vowel production: Articulatory and acoustic analysis. *Journal of the Acoustical Society of America*, 82, 847–63.

Suomi, K. (2007) On the tonal and temporal domains of accent in Finnish. *Journal of Phonetics*, 35, 45–55.

Suomi, K., Toivanen, J., & Ylitalo, R. (2003) Durational and tonal correlates of accent in Finnish. *Journal of Phonetics*, 31, 113–38.

Suomi, K. & Ylitalo, R. (2003) Syllable weight and segmental durations in Finnish. *PHONUM*, 9, 37–40.

Swerts, M. (1997) Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101, 514–21.

Szakay, A. (2006) Rhythm and pitch as markers of ethnicity in New Zealand English. In P. Warren & C. Watson (eds.), *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, Auckland, 421–6.

Tabain, M. (2003) Effects of prosodic boundary on /aC/ sequences: Articulatory results. *Journal of the Acoustical Society of America*, 113, 2834–49.

Tabain, M. & Perrier, P. (2005) Articulation and acoustics of /i/ in preboundary position in French. *Journal of Phonetics*, 33, 77–100.

Tabain, M. & Perrier, P. (2007) An articulatory and acoustic study of /u/ in preboundary position in French: The interaction of compensatory articulation, neutralization avoidance, and featural enhancement. *Journal of Phonetics*, 35, 135–61.

Taff, A., Rozelle, L., Cho, T., Ladefoged, P., Dirks, M., & Wegelin, J. (2001) Phonetic structures of Aleut. *Journal of Phonetics*, 29, 231–71.

Tajima, K. (1998) Speech rhythm in English and Japanese: Experiments in speech cycling. Doctoral dissertation, Indiana University.

Tajima, K., Zawaydeh, B. A., & Kitahara, M. (1999) A comparative study of speech rhythm in Arabic, English, and Japanese. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 285–8). Berkeley, CA: University of California.

Takeda, K., Sagisaka, Y., & Kuwabara, H. (1989) On sentence-level factors governing segmental duration in Japanese. *Journal of the Acoustical Society of America*, 86, 2081–7.

Tauroza, S. & Allison, D. (1990) Speech rates in British English. *Applied Linguistics*, 11, 90–105.

Touati, P. (1987) *Structures prosodiques du suédois et du français*. Lund: Lund University Press.

Traunmüller, H. & Krull, D. (2003) The effect of local speaking rate on the perception of quantity in Estonian. *Phonetica*, 60, 187–207.

Trouvain, J. (2004) Tempo variation in speech production: Implications for speech synthesis. Doctoral dissertation, Saarland University. *Phonus* 8, Phonetics, Saarbrücken.

Trouvain, J., Koreman, J., Erriquez, A., & Braun, B. (2001) Articulation rate measures and their relations to phone classification of spontaneous and read German speech. *Proceedings of ISCA Workshop: Adaptation Methods for Speech Recognition*, Sophia Antipolis (France), 155–8.

Tsao, Y.-C. & Weismer, G. (1997) Inter-speaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research*, 40, 858–66.

Tsao, Y.-C., Weismer, G., & Iqbal, K. (2006) The effect of inter-talker speaking rate variation on acoustic vowel space. *Journal of the Acoustical Society of America*, 119, 1074–82.

Tuller, B., Harris, K., & Kelso, J. A. S. (1982) Stress and rate: Differential transformations of articulation. *Journal of the Acoustical Society of America*, 71, 1534–43.

Turk, A. E. & Sawusch, J. (1997) The domain of accentual lengthening in American English. *Journal of Phonetics*, 25, 25–41.

Turk, A. E. & Shattuck-Hufnagel, S. (2000) Word-related duration patterns in English. *Journal of Phonetics*, 28, 397–440.

Turk, A. E. & Shattuck-Hufnagel, S. (2007) Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35, 445–72.

Turk, A. E. & White, L. (1999) Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27, 171–206.

Uchanski, R., Choi, S., Braida, L., Reed, C., & Durlach, N. (1996) Speaking clearly for the hard of hearing, IV: Further studies of the role of speaking rate.

Journal of Speech and Hearing Research, 39, 494–509.

Uchanski, R., Geers, A., & Protopapas, A. (2002) Intelligibility of modified speech for young listeners with normal and impaired hearing. *Journal of Speech, Language, and Hearing Research*, 45, 1027–38.

Ueyama, M. (1996) Phrase final lengthening and stress-timed shortening in the speech of native speakers and Japanese learners of English. *Proceedings of ICSLP96*, Philadelphia, 610–13.

Uldall, E. T. (1971) Isochronous stresses in RP. In L. L. Hammerich, R. Jakobson, & E. Zwirner (eds.), *Forms and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jorgensen* (pp. 205–10). Copenhagen: Akademisk Forlag.

Umeda, N. (1975) Vowel duration in American English. *Journal of the Acoustical Society of America*, 58, 434–45.

Umeda, N. (1977) Consonant duration in American English. *Journal of the Acoustical Society of America,* 61, 846–88.

Vaissière, J. (1974) On French prosody. *Research Laboratory of Electronics, MIT (QPR)*, 114, 212–23.

Vaissière, J. (1983) Language independent prosodic features. In A. Cutler & D. R. Ladd (eds.), *Prosody: Models and Measurements*, (pp. 53–65). Heidelberg: Springer.

Vaissière, J. (1991) Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord, & R. Carlson (eds.), *Music, Language, Speech and Brain* (pp. 108–20). Basinstoke: Macmillan.

Vanderslice, R. & Ladefoged, P. (1972) Binary suprasegmental features and transformational word-accentuation rules. *Language*, 48, 819–39.

Vatikiotis-Bateson, E. (1988) *Linguistic Structure and Articulatory Dynamics*. Bloomington, IN: Indiana University Linguistics Club.

Vatikiotis-Bateson, E. & Kelso, J. A. S. (1993) Rhythm type and articulatory dynamics in English, French and Japanese, *Journal of Phonetics*, 21, 231–65.

Vayra, M. Avasani, C., & Fowler, C. A. (1983) Patterns of temporal compression in spoken Italian. *Proceedings of the 10th International Congress of Phonetic Sciences*, Utrecht, 541–6.

Vayra, M., Avesani, C., & Fowler, C. A. (1999) On the phonetic basis of vowel–consonant coordination in Italian: A study of stress and "compensatory shortening," In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 495–8). Berkeley, CA: University of California.

Vayra, M. & Fowler, C. (1992) Declination of supralaryngeal gestures in spoken Italian. *Phonetica*, 49, 48–60.

Venditti, J. & Santen, J. P. van (1998) Modelling segmental durations for Japanese text-to-speech synthesis. *Third ESCA Workshop on Speech Synthesis (SSW3-1998)*, 31–6.

Verhoeven, J., de Pauw, G., & Kloots, H. (2004) Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47, 297–308.

Vogel, I., Bunnell, H. T., & Hoskins, S. (1995) The phonology and phonetics of the rhythm rule. In B. Connell (ed.), *Papers in Laboratory Phonology IV* (pp. 111–27). Cambridge: Cambridge University Press.

Volskaya, N. & Stepanova, S. (2004) On the temporal component of intonational phrasing. *Ninth International Conference on Speech and Computer (SPECOM 2004)*, St Petersburg, 4 pp.

Warner, N. & Arai, T. (2001) The role of the mora in the timing of spontaneous Japanese speech. *Journal of the Acoustical Society of America*, 109, 1144–56.

Warren, P. (1997) Timing patterns in New Zealand English Rhythm. *Te Reo*, 41, 80–93.

Welby, P. (2006) French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, 34, 343–71.

Wenk, B. J. & Wioland, F. (1982) Is French really syllable-timed? *Journal of Phonetics*, 10, 193–216.

Wennerstrom, A. & Siegel, A. F. (2003) Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36, 77–107.

White, L. (2002) *English speech timing: A domain and locus approach*. Unpublished doctoral dissertation, University of Edinburgh.

White, L. & Mattys, S. L. (2007a) Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró, & M.-J. Solé (eds.), *Segmental and Prosodic Issues in Romance Phonology* (pp. 237–57). Amsterdam: John Benjamins.

White, L. & Mattys, S. L. (2007b) Calibrating rhythm: First and second language studies. *Journal of Phonetics*, 35, 501–22.

Whiteside, S. (1996) Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association*, 26, 23–40.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992) Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91, 1707–17.

Williams, B. & Hiller, S. (1994) The question of randomness in English foot timing: A control experiment. *Journal of Phonetics*, 22, 423–39.

Wilson, T. & Zimmerman, D. (1986) The structure of silence between turns in two-party conversation. *Discourse Processes*, 9, 375–90.

Woodrow, H. (1951) Time perception. In S. S. Stevens (ed.), *Handbook of Experimental Psychology* (pp. 1224–36). New York: Wiley.

Yang, L. (2007) Duration, pauses, and the temporal structure of Mandarin spontaneous speech. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1289–92.

Zawaydeh, B. A. & de Jong, K. J. (1999) Stress, phonological focus, quantity, and voicing effects on vowel duration in Ammani Arabic. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 451–4). Berkeley, CA: University of California.

Zawaydeh, B. A., Tajima K., & Kitahara, M. (2002) Discovering Arabic rhythm through a speech cycling task. In *Perspectives on Arabic Linguistics*, vols. 13–14 (pp. 39–58). Amsterdam: John Benjamins.

Zellner, B. (1994) Pauses and the temporal structure of speech. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 41–62). Chichester: John Wiley.

Zellner, B. (1996) Structures temporelles et structures prosodiques en français lu. Revue Française de Lingusitique Appliquée, 1, 1–17.

Zlatousova, L. V. (1975) Rhythmic structure types in Russian speech. In G. Fant & M. A. A. Tatham (eds.), *Auditory Analysis and Perception of Speech* (pp. 477–83). London: Academic Press.

Zuraic, W. & Sereno, J. (2007) English lexical stress cues in native English and non-native Arabic speakers. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 829–32.

Zvonik, E. & Cummins, F. (2002) Pause duration and variability in read texts. In *Proceedings of the 2002 International Conference on* spoken language processing *(ICSLP '02)*, Denver, 1109–12.

Zvonik, E. & Cummins, F. (2003) The effect of surrounding phrase lengths on pause duration. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 777–80.

16 Tone and Intonation

MARY E. BECKMAN AND
JENNIFER J. VENDITTI

1 Introduction

The technical terms *tone* and *intonation* refer to patterned variation in voiced source pitch that serves to contrast and to organize words and larger utterances.[1] In this most general statement of their meanings, they are synonyms. However, the terms are differentiated in typical usage by applying them to different aspects of these linguistic uses of pitch. This differentiation is exemplified by the two parts of the sixth definition for the entry for "tone" in the *Concise Oxford English Dictionary* (*COED*; 11th edition): "(in some languages, such as Chinese) a particular pitch pattern on a syllable used to make semantic distinctions" and "(in some languages, such as English) intonation on a word or phrase used to add functional meaning." This sixth definition is tagged as the meanings for a technical term in phonetics, and its second part subsumes the term "intonation," which is defined in its own entry as "the rise and fall of the voice in speaking." In the *COED* entry, then, the primary sense of tone as a technical term in phonetics is a localized melodic event (a note or glissando) occurring over the span of a syllable, whereas tone *qua* intonation is a pattern of glissandi distributed over a longer span. Also, tone in the primary sense invokes a system of contrastive pitch patterns that act as minimal word-differentiating elements, comparable to the inventory of vowels or consonants of a language, whereas tone *qua* intonation invokes other functions, such as mirroring the syntactic structure of an utterance or indicating its pragmatic role in the larger discourse context. These two sets of characteristics make for a multidimensional taxonomy of phonetic form in relationship to linguistic function. This much is uncontroversial.

A third aspect of the *COED* definition is more controversial. The two parts of the definition refer to two different sets of languages, reflecting the claim in many broad-stroke surveys such as Hyman (2006) that particular values along the dimensions of form and function tend to coincide in ways that are conducive to a one-dimensional classification of language types, with "some languages, such as Chinese," at one end and "some languages, such as English," at the other.

Careful descriptions of specific languages at every point along the purported continuum, on the other hand, typically use the terms together in ways that defy the typology. For example, in many descriptions of specific Chinese dialects, such as Chang (1958), "tone" is used to refer to the localized melodic events (notes or glissandi) that contrast one-syllable words in citation-form utterances, but "intonation" is also used: to designate notes and glissandi that occur at phrase edges (rather than on designated syllables) with functions other than that of lexical contrast (e.g., marking interrogative speech acts), and to refer to longer-term modulations of the implicit melodic scale that defines the relationship of contrast among different notes and between two different rising or two different falling glissandi. Conversely, in many accounts of the English intonation system, such as Halliday (1967) and Pierrehumbert (1980), "tone" is used to refer to glissandi or notes that are localized to linguistically significant positions, such as the stressed syllables of some words and the edges of phrases.

In this chapter, we will elaborate on all three aspects of this definition of tone in the *COED*. We begin by reviewing the various ways in which phoneticians have represented the pitch patterns that they observe in the laboratory and the field (section 2). We then describe how research in the phonetic sciences over the decades since the development of such crucial analytic tools as the source-filter theory of vowel production (Fant, 1960; Harrington, this volume) has contributed to the evolution of a less superficial taxonomy of the forms (section 3) and functions (section 4) of spoken language melody. In these descriptions we will use many examples of how the taxonomy applies to varieties of Chinese and to English and other varieties of Germanic, so that we can close by evaluating the typological claim that differences observed within and between these two language groups can be collapsed into a single dimension of variation ranging from "languages, such as Chinese," at one end to "languages, such as English," at the other (section 5).

2 The Representation of Tone and Intonation

2.1 *Phonetic representations*

Because of the well-behaved psychophysical relationship between the percept of pitch and the fundamental frequency (f_0) of the periodic voice source, the choice among phonetic measures that one could use to represent tone and intonation patterns is fairly straightforward. Fry (1968) describes several early methods for estimating f_0 values over shorter or longer stretches of speech. These included measuring the durations of successive periods identified in oscillographic records (as in Denes & Milton-Williams, 1962) and tracing the frequency of some higher harmonic visible in narrow-band spectrograms (as in Lehiste & Ivić, 1963). More recently, the ready availability of computer programs for estimating f_0 using autocorrelation-based algorithms in free signal-analysis packages such as Praat (Boersma, 1993) and WaveSurfer (Sjölander & Beskow, 2000) has made the f_0

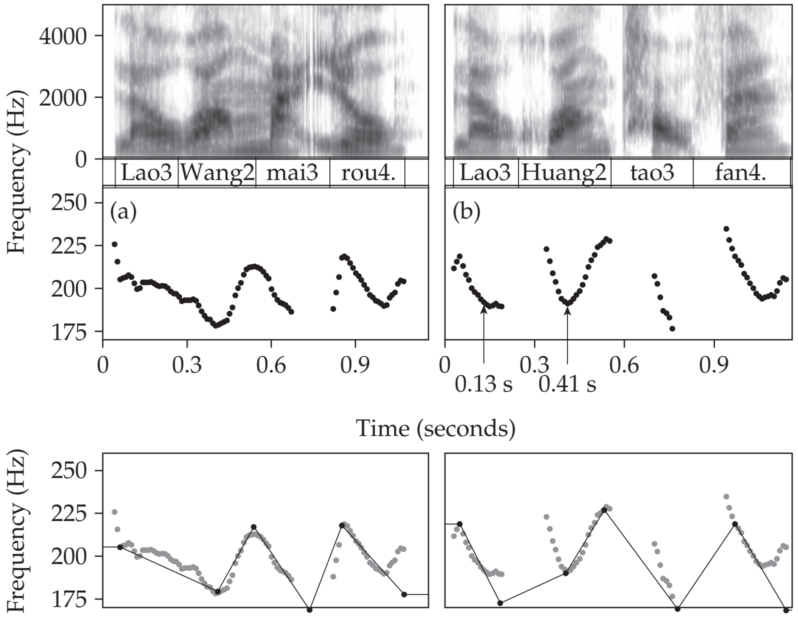


Figure 16.1 Spectrograms and f_0 contours for utterances of (a) *Lǎo Wáng mǎi rùo* [laʊ^{21}.waŋ^{35}.mai^{21}.ʐoʊ^{51}] "Old Wang buys meat" and (b) *Lǎo Huáng táo fàn* [laʊ^{21}.xwaŋ^{35}.taʊ^{21}.fan^{51}] "Old Huang begs for food", with a close-copy stylization (solid line) overlaid on each f_0 contour in the bottom row of panels.

contour of a recorded utterance an even more obvious choice as a first-pass phonetic representation of its intonation or tone pattern. The obviousness of this choice is reflected in the name by which the f_0 contour is often called. In both Praat and WaveSurfer, the f_0 estimates that are returned by the autocorrelation algorithm are called "pitch" values, and many phoneticians use the phrase "pitch track" in referring to a time plot of a sequence of estimated f_0 values over some interval of recorded speech.

It is important to remember, however, that the f_0 contour is only a very rough first-pass phonetic representation of the melodic pattern of an utterance, for at least three reasons. The first is that f_0 is not reliably estimated in stretches of speech where less regular source qualities such as creaky voice are in play. The failure of standard algorithms for estimating f_0 in such regions makes the "pitch track" a poor phonetic representation for melodic events that harness such a "nonmodal" voice quality as a cue, as illustrated in Figure 16.1. The f_0 contours in the middle panels of the figure were calculated over utterances of two sentences of Putonghua (PRC Standard Chinese) produced by an adult female speaker from Songyuan City in Jilin Province. The creaky voice in the third syllable of each utterance is a cue to the very low pitch target that characterizes this tone, as shown by Gårding et al. (1986), among others.

The second reason is essentially the same as the first, but applies to regions where the f_0 is well defined. When we see the concentrations of energy in a 600–800 Hz band at 0.27 seconds in each of the spectrograms in Figure 16.1, we can read this setting of the filter resonances in terms of the combined labial and dorsovelar constriction gestures for the initial [w] of the surname *Wáng* and for its devoiced allophone after the initial [x] in the surname *Huáng*. By contrast, when we see the f_0 value of 190 Hz at both 0.13 and 0.41 seconds in the right-hand utterance (i.e., at the points midway through the [a] nucleus in the two vowels before and after the [xw] of *Huáng* indicated by the arrows underneath the f_0 contour), we cannot know what combination of pulmonary and laryngeal gestures produced this setting of the glottal source period. There is a perceptibly lower tone target for the vowel at 0.13 seconds as compared to the target for the vowel at 0.41 seconds, and this percept of different targets matches transcriptions using Chao's (1930) tone letters. Typically, the tone in *Lăo* in this phrase-medial context is transcribed as a drop or as a dipping down to the bottom of the tonal space (i.e., [laʊ^{21}]~[laʊ^{214}]), whereas the tone in *Wáng* is transcribed as a rise from a middle region (i.e., [waŋ^{35}]).[2] Further support for positing such a difference in the target minima for these two tones comes from electromyographic studies (Sagart et al., 1986; Hallé, 1994), which show consistently high activity of the sternohyoid for the very low-pitched target in the low or dipping tone of *Lăo* but less reliable involvement of this muscle at the beginning of the rising tone of *Huáng*. The percept of a lower minimum target for [laʊ^{21}] than for [xwaŋ^{35}] in Figure 16.1 suggests that we hear the speaker's intent to produce the creaky voice quality that cues the lower tone target even when the glottal source wave is regular enough that the f_0 tracking algorithm does not fail.

The third reason that the "pitch track" can only be a first-pass phonetic representation of the melody of an utterance is the existence of so-called *micro-prosodic* effects, whereby the aerodynamics of producing contrastive properties of consonants can cause systematic variation in f_0 on consonants and neighboring vowels that is related to the percept of the consonants rather than to the percept of the utterance tune. These effects also are illustrated in Figure 16.1. The syntactic structure and the sequence of lexical tones are identical between the two sentences. The substantial differences in f_0 shape for the second, third, and fourth syllables are due to the different manners of articulation for the syllable-initial consonants. When native speakers listen to utterances such as these, they parse these micro-prosodic effects for what they are, and perceive the intended tone sequence that is common to both despite the marked differences in f_0 shape (see, e.g., Reinholt Peterson, 1986, Silverman, 1986).

These three limitations of the f_0 contour can be overcome to some extent by the use of analysis-by-synthesis techniques such as the "close-copy stylization" method pioneered by researchers at the Institute for Perception Research (IPO) in Eindhoven (see, e.g., Cohen and 't Hart, 1967; 't Hart & Collier, 1975; de Pijper, 1983). A *close-copy stylization* is defined as a synthetic approximation to the melody of the utterance which meets two criteria: "it should be perceptually indistinguishable from the original, and it should contain the minimum number of straight-line segments with which this perceptual equality can be achieved" (Nooteboom, 1997,

p. 646). This kind of downsampling of the f_0 contour is very easy to do today, because of the re-synthesis utilities based on LPC analysis or PSOLA (Atal & Hanauer, 1971; Moulines & Charpentier, 1989; also see Carlson & Granström, this volume) that have been implemented in many free signal-analysis packages. The bottom row of panels in Figure 16.1, for example, shows a close-copy stylization of each of the f_0 contours in the figure, created using the implementation of PSOLA re-synthesis in Praat (Boersma & Weenink, 2007).

2.2 *Analysis by synthesis*

Making such a close-copy stylization is a first step in developing and testing a phonological representation of the intonation pattern of an utterance in the models of the British English and Dutch intonation systems described in de Pijper (1983) and 't Hart et al. (1990). Both of these models pick out some of the line segments in a close-copy stylization as corresponding to phonologically significant events. Two types of phonologically significant event are identified: prominence-lending movements that are anchored to stressed syllables, and juncture-marking movements that occur at the edges of phrases. The phonological significance of a line segment is determined by the criterion of "perceptual equivalence" (as opposed to the "perceptual equality" of the close-copy stylization). Two stylized contours are equivalent if listeners perceive them to have the same utterance melody. In this framework for modeling utterance melody in languages such as English and Dutch, cataloguing the recurring patterns of sequences of line segments in "perceptually equivalent" melodies is analogous to cataloguing the inventory of contrasting vowels or consonants in transcriptions of words and phrases elicited in the field. That is, determining the patterns of melodic equivalence and dissimilarity across a sufficiently large and varied corpus of utterances in several iterations of analysis and perceptual testing of the re-synthesized utterances should yield the set of "melodically distinct pitch movements" for the language variety. The model parameters that contrast these line segments then are homologous to other distinctive feature sets for the language, such as its contrasting vowel heights or frication source places.

In de Pijper's model of British English, for example, there are eight melodically distinct glissandi which are parameterized in terms of their direction (rise versus fall), their slope (steep versus shallow), and the pitch levels between which they move (e.g., a half rise from lower to middle differs from a full rise from lower to upper levels). Some of these melodic elements are illustrated in the two panels of Figure 16.2. The top panel shows the original f_0 contour and a close-copy stylization of a two-phrase utterance produced by a young male speaker of British English. The lower panel repeats the close-copy stylization and overlays an approximation to the re-synthesized contour that would be generated by de Pijper's model. The intonation pattern in each phrase is the variant of what the IPO school has called the "hat pattern" depicted in figure 5.6b in de Pijper (1983). There is a steep prominence-lending half rise early in each phrase (on the first syllable of *royal* in the first phrase, and on *came* in the second) followed by a steep prominence-lending full fall near its end (on the first syllable of *messenger* in the

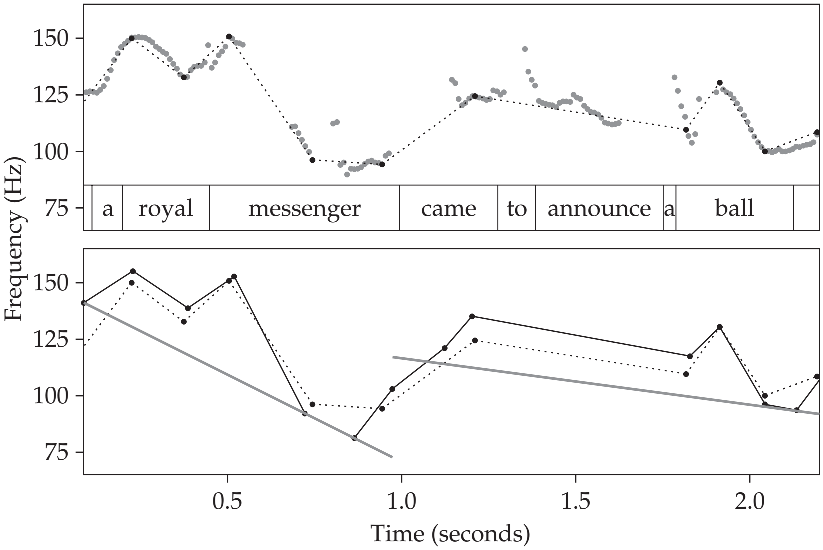


Figure 16.2 Extract from a reading of the *Cinders* passage in the IViE corpus (Grabe, 2004) with a two-accent "hat pattern" on each syntactic phrase. Dotted lines are a close-copy stylization overlaid on the original f_0 (top) and on approximations to de Pijper's (1983) model of the melodic elements in this pattern (bottom).

first phrase and on *ball* in the second). The only other phonologically significant line segments are the two "continuation rises" over the last part of *messenger* just before the interphrase boundary and over the second half of *ball*.

In addition to these parameters that specify local rises and falls, there are two other essential components of the model. One is the parameter set specifying the timing of each rise or fall relative to the segments of the syllable or at the phrase edge that licenses it. In de Pijper's model of English, for example, an early steep fall that is prominence-lending starts early enough to be completed at the onset of the vowel in the prominent syllable, a neutral fall starts 30 ms after the vowel onset, and a late fall does not start until after the end of the syllable.

The other essential component of the model is the "declination line" that is the implicit lower level for the melodically distinct rises and falls and that describes the f_0 over sections in between the melodically distinct movements. In complex utterances such as the two-phrase extract in Figure 16.2, there will be as many declination lines as there are phrases, with "reset" at the phrase boundary. The local declination line is steeper for shorter phrases and shallower for longer ones, as suggested by the two overlaid grey lines in the figure.

2.3 Phonological representations

We have presented de Pijper's model in detail because this kind of analysis by synthesis has proved to be a valuable tool for going from a database of phonetic representations of a good sample of utterances to an adequately formalized system of phonological representation for languages such as English, which do not offer the fieldworker the crutch of lexical contrast. Ladd (2008, p. 13) lauds this "IPO theory of intonational stucture" as "in many ways the first . . . serious attempt to combine an abstract phonological level of description with a detailed account of the phonetic realisation of the phonological elements." Other formal frameworks that phoneticians began to develop at about this same time also used an analysis-by-synthesis approach to decompose f_0 contours into contributions from three model components for (1) the set of localized pitch events, (2) the timing of these events relative to landmarks such as vowel onsets in prosodically relevant syllables, and (3) aspects of backdrop pitch range such as the reset points and declination slopes in the IPO model. Of course, different frameworks make different claims about the allocation of responsibility among these three components, as well as different claims about the appropriate set of parameters internal to each one. However, all fully formalized frameworks have this kind of compositional phonetics, so that the configuration of parameters of the synthesis model that generates an f_0 contour that is perceptually equivalent to each original f_0 contour for a set of utterances that share an intonation pattern can be construed as the phonological representation of that pattern. Formalizing the phonological representation of tune in terms of analysis-by-synthesis model parameters in this way gives linguists a way to compare models of a language across frameworks, or to compare models of different languages within a framework.

One point of comparison is the size of the smallest sequential element assumed. Where the IPO framework takes melodically equivalent glissandi to be the basic atomic units, many other models of English and Dutch decompose each rise or fall into a finer-grained sequence of endpoint notes. Pierrehumbert (1980) and Gussenhoven (1984), for example, both analyze the steep prominence-lending fall of the English "hat pattern" in Figure 16.2 as a transition from a higher pitch target to a lower pitch target. These targets, then, were called "high tone" and "low tone" in an explicit analogy to the use of these terms in work such as Benedict's (1948) description of Thai and Cantonese, and Ward's (1948) description of Efik and Igbo. The analogy made it possible to draw directly on the same kind of compositional phonetics that was beginning to be applied to languages such as Igbo in order to understand better the interplay between the lexically specified melodic elements that are the "tonemes" of the language and any melodic elements that are produced or parsed "post-lexically" for sequences of words or phrases in their larger discourse contexts. The seminal example of this kind of model is Bruce's (1977) description of Stockholm Swedish word tones, which inspired critical elements of Pierrehumbert's (1980) model of American English intonation patterns as well as of Pierrehumbert and Beckman's (1988) model of Tokyo Japanese word-accent patterns in sentential context.

The analogy worked in the other direction, too, making it clear, for example, that the segmentation grain for utterance melodies is a theoretically interesting question even for languages with lexically contrastive tone. Thus, where Ward (1948) and others analyze Igbo word and phrase patterns as sequences of tone levels, with high or low specified for each syllable, Clark (1978) proposes a system of "dynamic tones" so that a rise or fall is specified only at linguistically significant syllable junctures, as in the IPO framework models described above. Similarly, where Kindaichi (1957) and Haraguchi (1977) analyze Japanese word- and phrase-level melodies in terms of a succession of low or high tones specified on all moraic segments, Kawakami (1957), Hattori (1961), and Fujisaki and Sudo (1971) analyze them as combinations of underlying rises and falls. We will return to this point in section 3.1 after describing the ramifications of such differences among models for the symbol sets that are a more typical referent of the term "phonological representation."

2.4 Symbolic representations

The control parameters for producing and parsing spectral patterns of utterances often are referenced symbolically using transcription systems such as the International Phonetic Alphabet (IPA), in which each basic segmental unit is represented by a letter symbol. The second syllable of the utterance displayed in the upper-left panel of Figure 16.1, for example, can be transcribed as a sequence of three segments [w], [a], and [ŋ], whereas the second syllable of the paired utterance to its right might be transcribed as these three segments plus an initial [x]. While phoneticians disagree on the ontological status of such symbolic tags (e.g., Pierrehumbert, 1990, versus Ladefoged, 1990), there is an overwhelming consensus that this grain of syntagmatic discretization is a useful starting point for phonological analysis, and hence, that the IPA provides a useful common vocabulary for annotating recordings made in the lab or the field and for comparing models of what talkers and listeners implicitly know about how to differentiate words such as the surnames Wáng and Huáng in these sentences.

There is no comparable standard alphabet for segmenting and tagging pitch patterns. The 1989 Kiel revision of the IPA resolved a rivalry between Africanist and Sinological conventions for tagging pitch patterns in languages such as Igbo versus Cantonese, but only by including both transcription systems. In his summary of discussion by the Working Group on Labeling of Suprasegmentals at the Kiel meeting, Bruce (1989, pp. 36–7) describes the failure to achieve even such a laissez-faire resolution for tagging pitch patterns in languages such as English:

> There exists an apparent need for a direct way of symbolizing intonation in a phonetic transcription. However, the opinions diverge regarding the exact way of transcribing intonation. For a phonological transcription of intonation the symbolization is very much dependent on the language and the analysis.

Why might tone and intonation contrasts be so much less amenable to a "phonetic" transcription than are consonants and vowels? We think this is because there is no natural universal segmentation for the pitch pattern shorter than the utterance as a whole. That is, there is nothing akin to the segmentation of filter-resonance patterns afforded by the abrupt transition from a stop-like closure into a more open vocal tract in the CV-like "frame" of canonical babbling (MacNeilage & Davis, 2000) and in the "vocal motor schemes" (McCune & Vihman, 1987) that become the infant's first words. Across cultures, mothers may use a common stock of attention-getting tunes to draw very young infants into sessions of imitative turn-taking, as suggested by Papoušek et al. (1991). And this common stock of preverbal melodies may be a basis for trends such as the prevalence of raised pitch and rising terminals in yes/no questions noted by Lieberman (1967), Bolinger (1978), and Ohala (1983), among others. However, these are not universals of syntagmatic alternation within the utterance. They do not make a compelling rhythmic base for decomposing the melody of a conversation into a sequence of elements any smaller than the tunes of the alternating interlocutor turns.

This difference in preverbal rhythmic base gives the symbolic transcription of linguistically significant pitch variation a fundamentally different status from the symbolic transcription of linguistically significant spectral variation. Alphabetic transcription can be a useful pretheoretical tool for identifying events that are very likely to have phonological significance in the vocal tract filter pattern. The analogous use of symbolic transcription for voice source events makes sense only within the context of a research community in which shared expectations about tunes and their relationship to prosodic structures above the level of CV units can emerge. For example, the Africanist transcription system that is included in the IPA evolved in a community of Bantuists and of researchers working on non-Bantu languages in the West African Sprachbund, where a long history of language contact has given rise to striking commonalities in syntactic and in prosodic organization above the word. The IPO symbols for English and Dutch prominence-lending shapes and boundary pitch movements, similarly, evolved in a community of researchers who developed the symbols as shorthand for particular sets of parametric specifications in a shared analysis-by-synthesis model of the intonation systems of these two closely related language varieties.

2.5 Parametric representations in prosodic phonology

The lack of a universal preverbal rhythmic base for segmenting speech melodies at any level below the whole utterance may also explain why languages with lexical tone contrasts figured so prominently in the early development of a lineage of frameworks that includes Prosodic Phonology (Henderson, 1949) and Declarative Phonology (Coleman, 1992) on one side of the Atlantic, and Autosegmental Phonology (Goldsmith, 1976/1979), Metrical Phonology (Liberman 1975/1979), another Prosodic Phonology (Nespor & Vogel, 1986), and Articulatory Phonology (Browman & Goldstein, 1986), on the other. That is, lexical contrast typically provides a more compelling functional basis than pragmatic contrast for segmenting the melodic contour into units smaller than the whole utterance. When key ideas in this lineage were applied in describing utterance melodies in languages such as English (Pierrehumbert, 1980; Gussenhoven, 1984) and Korean (Jun, 1993/1996), they became what Ladd (1996) termed the "Autosegmental-Metrical" (AM) approach. These ideas became a basis for annotation conventions developed using the Tones and Break Indices (ToBI) framework (cf. the various language descriptions in Jun, 2005, as well as the chapter by Beckman et al. in that volume for a list of "design principles").

In reading early work in this lineage, we are often struck by the congruence between concepts that were assumed in describing, for example, "the tone-phrasing system of Kongo" (Carter, 1974) and the parameterization of intonation patterns in analysis-by-synthesis models such as de Pijper's (1983) model of British English described earlier. In particular, whether lexically specified tones are transcribed by diacritics on the vowels (as in the Africanist tradition) or by numerals after each syllable (as in the Sinological tradition), utterance melodies are typically described in terms of a convolution of two parts. One part is the concatenated sequence of pitch levels for the transcribed tones and the other part is linguistically significant modulations of what Ladd (1992) calls the *tonal space* (see section 3.3). This partitioning of tunes into tones and tonal space is evident in the Beijinghua utterances in Figure 16.3, where the rising glissando on the word *hé-zi* "box" can be transcribed in terms of a sequence of lower and higher pitch targets ([^{23}] in Chao's tone numbers or LH by Yip's, 1980, analysis) in both cases, but the high target at the end of the glissando in the right-hand utterance is higher because it is realized in an expanded tonal space. This partitioning is also congruent with the distinction in the IPO framework between the parameters that specify the melodically significant glissandi and the parameters that specify a sequence of declination lines for successive phrases.

Carter's account of "the tone-phrasing system of Kongo" is also characteristic in distinguishing between tones that are anchored to specific syllables in a word and tones that signal other phonologically significant anchoring points such as the edges of larger phrases. This distinction is congruent with the distinction between pitch movements on stressed syllables and boundary pitch

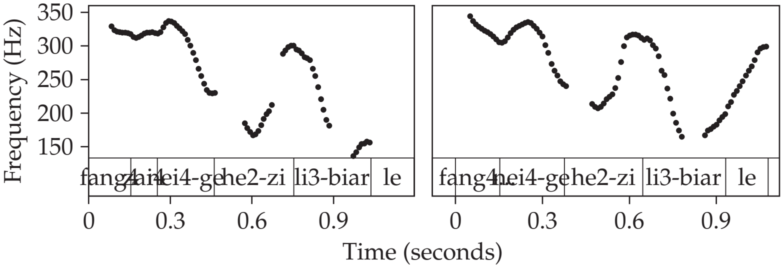


Figure 16.3 Extracted f_0 contours for the sentence *Fàng zài nèi-ge hé-zi lǐ-biar le*, produced by a female speaker of the Beijing dialect of Putonghua in staged dialogues where the context makes it a statement '(I) put (them) in that box' (left) or a question 'Had (she) put (it) in that box?' (right). (From Lee, 2005)

movements in the IPO models of English and Dutch, as well as with descriptions of some utterance-level melodic contrasts in many languages with lexically contrastive tone shapes. The declarative and interrogative utterances in Figure 16.3 illustrate the Beijinghua case. Whereas the pitch events on earlier syllables reflect the lexically specified tones, the high tone anchored at the end of the last syllable on the right is a pragmatic morpheme that signals its interrogative speech act.

In this chapter, we will call this lineage of frameworks "prosodic phonology" – using lower case to differentiate this common noun from the homophonous name of two particular frameworks within the lineage (Bazell et al., 1966; Nespor & Vogel, 1986), and also to make clear that the key ideas invoked by this term are generic. Each of these ideas was developed more or less independently in at least two frameworks within the lineage and was congruent with approaches being developed at the same time in the allied science of speech synthesis. We will adopt transcription conventions from the relevant prosodic-phonology models of tone and intonation systems when we describe f_0 contours of example utterances or discuss melodic contrasts in the language varieties from which these examples are drawn. For instance, the strings of functionally annotated high (H) and low (L) tones in (1) and (2) are possible ways to symbolize the tunes of the three utterances depicted in Figures 16.2 and 16.3.

```
(1)  (One day) a royal messenger   came to announce a ball.
                  |     |           |                  |
                L+H*  L+H*  L-H%   H*                L+H* L-H%
```

```
(2)  a.  Fàng zài nèi-ge hé-zi   lǐ-biar le.    b.  Fàng zài nèi-ge hé-zi  lǐ-biar le?
         [fã.ða     ne .ɣə  xɤ.ð   li.pə.lə]         [fã        ne.ɣə  xɤ.ðə   li.pə.lə]
           |         |        |      |                 |          |      |       |
          H+L       H+L      L+H     L     L%         H+L        H+L    L+H      L      H%
```

In both sets of transcriptions, there are linking lines. Each such line indicates that a tone or tone sequence below is associated to a designated syllable in the orthographic or phonetic transcription above. Also, in both transcriptions, the + infix conjoins tones that are anchored as a glissando around the designated syllable, and in (1), the * suffix on the L+H indicates that English contrasts two rising glissandi, L+H* versus L*+H, which differ in how they are anchored to the designated syllable. By contrast, each % affix in the transcriptions indicates a tone that is anchored at a phrase boundary rather than internally to a phrase, whereas the – suffix in (1) marks a "floating" L phrase tone that is realized somewhere between the preceding L+H* and following H% targets. While we find it useful to adopt these tagging conventions, however, we must emphasize that the symbol strings are not narrow phonetic transcriptions. Moreover, they are not even broad phonemic transcriptions until they are construed as names of meaningful configurations of parameter settings in an analysis-by-synthesis model for the speaker's dialect of British English or of Mandarin Chinese.

3 A Taxonomy of Formal Parameters

In this section, we amplify on the relationships between three key developments in prosodic phonology and the components of typical synthesis models that allow symbol strings such as the ones in (1) or (2) to be read as pieces of a broad phonemic transcription in some actual or possible formal model of the tone and intonation systems of the language.

3.1 Segmenting the melody

The first key development was the notion that the melody of an utterance can be segmented into a string of localized events – single notes or the conjoined notes in glissandi or in more complex sequences such as dipping movements – and that this segmentation is autonomous of the formal properties and functions that allow the native-speaker listener to parse the filter-resonance patterns in terms of the consonant and vowel inventories of a language. This idea gives the name of the Autosegmental Phonology framework (Goldsmith, 1976/1979), but it is not unique to that framework. It is implicit in the cataloguing of significant pitch movements in the IPO model, in the identification of turning points in the Lund model (Gårding, 1977, 1993), and in the detection of phrase and word-accent commands in the Fujisaki model (Fujisaki & Sudo, 1971; Möbius et al., 1993).

It is useful to begin this discussion of the autonomous status of tone segments by reviewing the phonetic bases for the earlier conceptualization of tone as a suprasegmental feature. A great deal of research over the last five decades highlights how a taxonomy of formal properties of vowel and consonant systems across languages emerges from the interplay between information-theoretic principles and the physiology and physics of speech. Indeed, the very fact that there are vowel and consonant systems can be related to the ways in which spoken utterances are naturally segmented by the spectral discontinuities that result when constrictions (i.e., consonant gestures) are superimposed on more open vocal tract postures (vowel gestures). As Goldstein (1989), Ohala (1992), and many others point out, even though the consonant and vowel gestures are not themselves sequenced, the acoustic patterns they produce are effectively sequenced because of two facts about the types of constriction that yield the most robust CV segmentation. First, these constrictions block the transmission of spectral information that gives the listener clues to coarticulated vowel postures behind the place where airflow is impeded. Second, some contrastive features of the most effective consonant segments, such as the place of impedance for a stop, are only audible during the acoustic intervals for coarticulated vowel postures, so that spectral properties at the edges of vowel segments must be treated as transitions between consonant states and vowel states in order to recover these "hidden" consonant features.

Although concurrent tone gestures also are "hidden" by these consonant constrictions, source and filter resonances are to a large extent independently controlled during intervals where airflow is not impeded. As Pierrehumbert (2000) points out, this independence of source and filter for vowels means that tone features are carried on a considerably more separable channel of acoustic information when compared to the "hidden" features of consonants (which, as just noted, are carried on exactly the same channel of spectral resonance patterns that carry the vowel features). Vowel segments are thus the more reliable intervals for transmitting information about concurrent tone gestures during a sequence of consonants and coproduced vowels. This is the psychophysical basis for a type of tone system in which each vowel segment in a word or phrase is the nucleus of a syllable that can be counted off in the metrical structure of the utterance by virtue of its having exactly one associated lexically specified tone. This kind of "tone syllabification" is especially characteristic of utterances in languages such as Cantonese, where most syllables are monosyllabic content words, and tone features typically preserve the syllable count even at very fast speech rates where the vowel features are swallowed up (see section 4.1).

Pike (1948, pp. 3–5) reserved the term "tonal language" for a language with this kind of tone-syllabification system, whereas many other later researchers such as Voorhoeve (1973), McCawley (1978), Goldsmith (1984, 1987), and Hyman (2001, 2006) defined "tone language" more broadly and used terms such as "restricted tone system" or "pitch accent" to differentiate the "unrestricted tone system" of Cantonese from their accounts of the underlying tone sequence in languages such as Safwa or Tonga, where words typically are longer and phonotactic constraints strongly restrict what tones can occur where within a word. One of the more fiercely contested questions in prosodic phonology today is the prevalence of tone syllabification as a basis for counting tone targets in the melodic contours of words and phrases. The question arises in part because of the alphabetic bias to model melodic contours of languages with "restricted tone systems" as a succession of "phonetic" tones for all syllables even over stretches where the observed pitch pattern on the vowels depends completely on the pitch targets for nearby "phonemic" tones.

For example, a common strategy for training models of Mandarin Chinese tone and intonation is to use databases of recorded utterances of sentences such as the two in Figure 16.1 (e.g., Lee et al., 1993; Shih & Kochanski, 2000). This strategy in effect treats the language as if it had a Cantonese-like prosodic system. By contrast, Kratochvil's (1998) corpus study suggests that the Beijing dialect at least differs prosodically from Cantonese, and that examples such as the two utterances of the sentence in Figure 16.3 are more typical. The pinyin orthographic transliteration of this sentence in (2) above shows nine "characters" or *zì*. (This Chinese term *zì*, which we will use henceforth, names a grammatical unit that Riha, 2008, terms the "morpheme-syllable" to emphasize that the orthographic unit is a salient morphosyntactic chunk corresponding to a well-formed prosodic constituent that can stand alone when it is pronounced with a fully realized tone.) However, of these nine *zì*, five (*zài*, *-ge*, *-zi*, *-biar*, and *le*) are affixes or grammatical particles with "neutral tone" – i.e., they have no lexical tone specification, as indicated for the four *zì* other than *zài* by the lack of a tone diacritic. (The locative particle *zài* is transliterated with the diacritic for the [^{51}] lexical tone because it is listed in most dictionaries under the entry for the related locative verb, but the particle and verb are not homophones; the particle has neutral tone and there is no fall in pitch even in the declarative utterance shown in the left panel of Figure 16.3, where the consonant and vowel are not lenited to nothing, as they are in the interrogative utterance on the right.) How can we account for the f_0 values in neutral tone *zì*?

Chen and Xu (2006) follow recent accounts such as Yip (1980/1990) to argue that even when there is no lexically specified tone pattern, each syllable in an utterance has a surface target tone level. The target value for a neutral tone *zì* is M (i.e., midway between the H and L targets of the four contrastive lexically specified tones) and the actual pitch value is an average of this M with the pitch value of the immediately preceding tone target. By contrast, Li (2003) follows earlier accounts such as Chao (1932, 1968) in assuming that a neutral tone *zì* has no tone target even on the surface. The pitch pattern over such a *zì* can be an extension of the pattern for the last preceding lexically specified tone (e.g., continuing the fall after the [^{51}] falling tone or realizing the optional rising tail after the [^{214}] dipping tone) or it can be just part of the transition between tone targets on either side (e.g., in (2), the last two syllables are the transition from the L tone on the root morpheme *lǐ* of *lǐ-biar* to the pragmatically specified L% or H% boundary tone at the end).

The disagreement over which type of account is better is reminiscent of the disagreements that occasionally arise in the literature on transitional elements such as the release phase of obstruents in Berber. Coleman (1998) transcribes the release as a reduced vowel, which Dell and Elmedlaoui (1998) insist is not part of the phonological inventory of the language, so that by their analysis many syllables are headed by an obstruent consonant, violating a purported universal minimum sonority constraint on syllabicity, albeit not violating most formulations of the universal as a sonority sequencing constraint. However, those disagreements arise very infrequently relative to the consensus view that consonant and vowel gestures in spoken languages tend to be configured syntagmatically in such a way that native speakers and linguists can identify a string of CV units, as in the IPA transcriptions in (2a) and (2b). By contrast, disagreements like the one that gives rise to different counts for the melodic units in these utterances are endemic across the communities of researchers working on tone and intonation. These disagreements are almost inevitable, because they reflect an inherent ambiguity in the parsing of tonal gestures. This ambiguity stems from the fact that a vowel-by-vowel segmentation of the melody is not intrinsic to the production and perception of tone gestures per se, but instead is parasitic on the CV segmentation of the spectral pattern that is intrinsic to the production and perception of obstruent gestures.

Lieberman (1967) proposes a rather different phonetic basis for segmenting the melodic contour that he describes as emerging from the interplay between syntactic structures governing the flow of information in a discourse and the coordination of respiratory and laryngeal postures to control expiratory airflow for phonation. In particular, he suggests that, absent a "marked" gesture to change laryngeal tension, the posture for sustained phonation results in a rise to high f_0 at the beginning of controlled expiration and a rapid fall in f_0 toward the end, as in the combination of prominence-lending movements in the IPO “hat pattern” in Figure 16.2. This rise and subsequent fall forms a natural unit of segmentation, which Lieberman calls the “unmarked” breath group. He also describes a “marked” breath group, which instead has a final rise produced by a localized laryngeal-tensing gesture. He claims that a comparably localized gesture to boost subglottal pressure can also make for a more extreme early rise or a rise in other positions to mark focal prominence (“emphasis” or “contrastive stress”) in the discourse context of the utterance.

Although Ohala (1970) and other later work has discredited Lieberman’s characterization of a definitive role for subglottal pressure in the production of local melodic events such as the L+H* rise when used to mark focal prominence on a particular syllable or word in English, Lieberman’s depiction of an early rise and late fall that defines the melodic contour for an “unmarked breath group” captures a fairly common aspect of phrasal melody. Safwa, Basque, Japanese, French, and many other languages use a small set of tone sequences, often involving a rise in pitch anchored near the beginning and a fall in pitch anchored later, to highlight the edges of utterances and to segment them into smaller prosodic phrases. These same prosodic phrases often seem to be the domain for specifying an expanded or compressed tonal space to express the relative prominence of the constituent as a whole (see section 3.3), and it is less implausible that this expression of phrasal prominence relationships could involve adjustments to the pulmonary expiration rate as a mechanism for overall volume control.

The utterance in Figure 16.4 illustrates this delimitative aspect of the tone-phrasing system of Japanese. There are four prosodic phrases, each of which is marked by an initial rise in pitch. This rise is analyzed in the X-JToBI tagging conventions as a sequence of a low boundary tone that is anchored strictly at the phrase edge and a high phrasal tone which is timed to follow the low tone at some loosely fixed distance which depends both on the prosodic structure of the first syllable and on distance to the next melodic event. In every phrase but the third, there is also a steep fall at a designated syllable that is marked by an apostrophe in the transliteration of the word in (3a).

```
(3)  a.  Ya’mano-wa    oyo’ideru-ga,   marude     oborete-iru       yo’o-da.
         [ja.ma.no.ɰa  o.joj.de.ɾɯ.ŋa  maɾɯ.de    o.bo.ɾe.te.i.ɾɯ   jo:.da]
           |              |                                         |
      %L  HL         L%  HL         L%    H-  L%   H-              HL  L%

     b.  Ya*mano-wa    oyo*ideru-ga,   marude     oborete-iru       yo*o-da.
          |  /          |  | |           |    |    |                 | |
          H*L           L H*L            L    H    L                 H*L

     c.  [ja.ma.no.ɰa  o.joj.de.ɾɯ.ŋa  maɾɯ.de    o.bo.ɾe.te.ɾɯ     jo:.da]
         [H  L  L  L   L HL L  L  L    L  H  H    L  H H  H  H      HL L]
```

This lexical specification for anchoring a HL tone sequence at some designated syllable differentiates "accented" words such as *oyo'ideru* 'is swimming' from "unaccented" words such as *oborete-iru* 'is drowning'.[3] Although the verb form in the second phrase is unaccented, there is a steep fall in this VP because the following evidential particle is lexically accented. Even in phrases that contain no accented words, however, there is most typically a fall, albeit often a more gradual one, which the X-JToBI conventions analyze as a transition from the high target at the end of the phrase-initial rise to the low boundary tone at the following phrase edge. Prosodic groups (*accentual phrases*) can be counted off in an utterance from the distribution of the phrasal rises and subsequent steep or gradual falls. Also, while the tone patterns differ in other dialects, with some having more complicated lexical contrasts and some having no lexical contrasts, the metrical structures that are defined by the distribution of tones relative to accentual phrase (AP) boundaries seems to be shared across dialects.

Accounts such as Kawakami (1957) and Pierrehumbert and Beckman (1988) concentrate on the way that the tone patterns mark off the salient prosodic groups to make for the pan-Japanese metrical system. In such accounts, melodic contours for utterances in the standard dialect are segmented only into those tones that are anchored relative to the phrase edges and those tones that are anchored at the designated syllables of accented words. All other parts of a contour are described as tonally "underspecified" and modeled as transitions between the nearest tones on either side, making tonal transcriptions of Tokyo Japanese utterances such as the one shown in Figure 16.4 look like transcriptions of utterances in the Autosegmental-Metrical model of the American English intonation system that was invoked in (1). The transcription in (3a) illustrates, using the X-JToBI conventions (Maekawa et al., 2002).

By contrast, in Kindaichi (1957) and Haraguchi (1977), the focus is primarily on making a spare underlying representation for the lexical contrasts between the absence versus presence of the HL sequence and (if present) among different anchoring positions within the word. These contrasts are represented by marking the designated vowel with a * to show where the HL sequence is to be inserted at the initial stage of deriving the surface pitch pattern. The pattern on other parts of the accentual phrase is modeled by derivational rules that conditionally insert L and H tones on the initial and final vowels, as in (3b), and then copy the inserted tones or the lexically specified tones onto other vowels, to produce a "fully specified" surface pattern, as illustrated in (3c). This tone-spreading account makes the intonation system of Tokyo Japanese look like Voorhoeve's (1973) picture of the "restricted tone system" of Safwa and also like Goldsmith's (1984) account of "tone and accent in Tonga" a decade later.

The difference between the 13 tone segments assumed in the transcription in (3a) and the 21 tone segments assumed in the transcription in (3c) is also parallel to the difference between specifying tones for just four of the *zì* in Li's (2003) model that yields the transcription in (2a) as compared to specifying these four plus the five M targets for the neutral tone *zì* in Chen and Xu's (2006) model. In both of these cases, one account assumes that the sequence of syllables (or

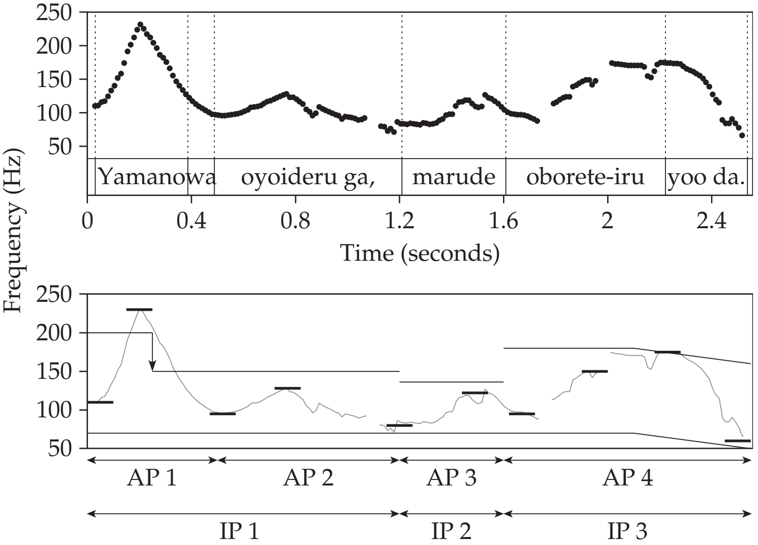


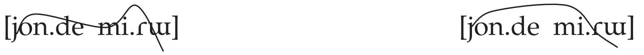
Figure 16.4 Extracted f_0 contour for an utterance of the sentence in (3) meaning 'YAMANO is swimming, but he's nearly drowning right now', produced by a male native speaker of Tokyo Japanese. The copy in the bottom panel shows overlaid lines for tone targets and tonal space settings that are described in section 3.3. The utterance is from Venditti et al. (2008).

other potential tone-bearing units) is "fully specified" for tone targets whereas the other account assumes that the nodes at this level of the prosodic hierarchy are "underspecified" for tone. These names characterize the disagreement in terms of their different assumptions about the set of localized pitch events – i.e., the first of the three synthesis model components listed in section 2.3.

Such disagreements have consequences for the depiction of the "underlying" tone specification. For example, in the fully specified account of Japanese tone patterns, the starred tone of the H*L word melody is associated to the designated mora (which is marked with a * in the lexicon) at the first stage of the derivation, and then a L is inserted on the first tone-bearing unit of an AP just in case that mora does not already bear a tone specification. This account therefore predicts that there will be no tone difference between a sequence of clauses such as *yonde mi'ru* 'call and then see' and a verb-auxiliary construction such as *yonde-mi'ru* 'try calling', since in both cases the initial vowel of *mi'ru* will already have an associated H tone at the stage of the derivation when an initial L is conditionally inserted, as in (4c). By contrast, in the underspecified account, that first L is a boundary tone that marks the edge of an AP whether or not the first syllable is accented. Thus, utterances of the two-clause sequence often would be distinct from utterances of the verb-auxiliary construction, because the two-clause sequence often will be produced as two APs, as shown in (4a).

```
(4)  a.  yonde   mi'ru 'call and then see'       yonde-mi'ru 'try calling'
                  |                                         |
      %L H-   L%  H+L L%                        %L  H-      H+L L%
```

b.



```
     c.  yonde  mi*ru 'call and then see'        yonde-mi*ru 'try calling'
          |     |   |  |                          |         |  |
          L     H   H*L                           L         H*L

     d.  [LH H   H L]                            [L H H  HL]
```

It is important to note that these differences in the analysis of the underlying forms stem from the more fundamental disagreement about the nature of surface phonetic representations. In the X-JToBI account of Tokyo Japanese (as in all ToBI framework accounts), the surface phonetic representation is the actual pitch pattern, as deduced from representations such as Figure 16.4, which shows an f_0 track calculated from a recording of a specific utterance of (3a), or as shown in (4b), which is a schematic "pitch track" summary of the many f_0 contours that we have observed for actual utterances of the phrases in (4a). In Haraguchi's account, by contrast, the surface phonetic representation is still a symbolic transcription – a sequence of discrete pitch targets associated vowel-by-vowel, as in (4d). On the surface, then, this account makes Japanese look like an unrestricted tone language. How can we decide between these two accounts?

Pierrehumbert and Beckman (1988) made the following predictions. If the fully specified account is an accurate representation of what the speaker intends to produce, then a sequence of spread L tone targets (as in the last four vowels in *oyo'ideru-ga* in (3c)) or a sequence of spread H tone targets (as in the second through sixth vowels in *oborete-iru-yo'o-da* in (3c)) should show the same pattern of actual f_0 values over the associated vowels regardless of the length of the sequence. In the underspecified account, by contrast, the f_0 contour over such stretches could fall or rise at different rates, depending on the distance between the two tone targets specified at the surface. Pierrehumbert and Beckman tested these predictions using a set of elicited utterances of three-phrase sentences in which both the accent status and the number of syllables in the words in the middle phrase were systematically varied. For unaccented medial phrases, they measured the slope of the f_0 downtrend over the interval between the peak f_0 near the beginning of the AP to the minimum f_0 at the next phrase boundary – i.e., over an interval that would be represented as a sequence of H tones in Haraguchi's account but as a mere transition from a phrasal H- to a L% boundary tone in the underspecified account. For accented phrases, they fit two slopes, differentiating the steep fall of the H+L tones at the designated syllable (which they predicted to have a fixed duration and slope) from the shallower decline over the variable-length region from the L of the accent to the L% at the end of the phrase. In both cases, the slope of the downtrend over the variable-length region up to the phrase edge was steeper for shorter intervals and shallower for longer ones, as predicted by the underspecified account.

In differentiating between the fully specified and underspecified accounts of Tokyo Japanese tone patterns, Pierrehumbert and Beckman (1988) fit very simple (straight-line) curves to the f_0 contour over both types of tonally unspecified intervals. As Pierrehumbert (1980, p. 12), van den Berg et al. (1992), Beckman and Pierrehumbert (1992), Myers (1998), Ladd and Schepman (2003), and many others point out, however, the shape of a transition over tonally unspecified regions is a research question in its own right. Moreover, it is a question that is tied up inextricably with questions about alignment or anchoring – i.e., about how the speaker synchronizes tone gestures with vowel and consonant gestures so that the listener correctly parses where the targets are anchored in relation to prosodic positions such as stressed syllables and phrase boundaries.

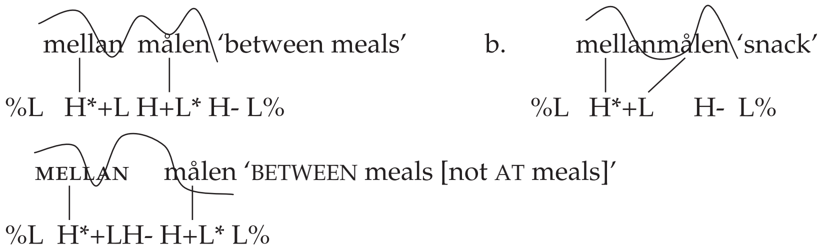
3.2 *Anchoring the tones in time*

The second key development in prosodic phonology was the idea that tonal autosegments are not suprasegmental features of the vowel segments on which they realized. Rather vowel (and consonant) segments as well as tone segments are associated to positions in a metrical structure, and this structure and the association patterns are objects of study in their own right. This idea is often associated with the Metrical Phonology framework (Liberman, 1975/1979; Liberman & Prince, 1977; Selkirk, 1981), but again, it is not unique to that framework. It is developed more fully in the treatment of coarticulation of consonant and vowel features in the Articulatory Phonology framework (see Browman & Goldstein, 1986; Byrd & Saltzman, 2003; other work reviewed by Fletcher, this volume). For tonal autosegments, this idea is implicit in the functional separation between prominence-lending pitch movements and boundary pitch movements in even the earliest IPO system models, and it corresponds to the distinction between turning points and pivots in the Lund model and to the distinction between phrase commands and accent commands in the Fujisaki model.

To show how this development was separate from the first key idea, we begin by comparing what "association" means in the two different accounts of Japanese discussed above. The fully specified "phonetic representations" in (3c) and (4d) can do away with the link lines and just list the string of H and L tones, reflecting the assumption that each tone is aligned simply to coincide with the vowel or moraic nasal segment to which it associates by rule. Beckman et al. (1983) describe a synthesis model couched in this Autosegmental Phonology framework which specifies a target f_0 value for the H or L tone midway through each vowel or moraic nasal in this way. By contrast, the underspecified surface transcriptions in (3a) and (4a) must show link lines to identify the accent tones and their designated syllables in accented words. Other tones must be annotated for their anchoring relationships. The annotation conventions differentiate the %L and L% boundary tones that anchor tightly at the phrase edge from the H- phrase tone that is only loosely aligned relative to the edge of the accentual phrase that begins with an unaccented word. Pierrehumbert and Beckman (1988) describe a synthesis program couched in the Autosegmental-Metrical framework which specifies target f_0 values at various time points that are chosen to relate the linguistically significant f_0 peaks, valleys, and inflection points ("elbows") in the phonetic representation in Figure 16.4 to the functional differences among the accent tones, the boundary tones, and the phrase tones. Although the input is a sequence of tones, the ways in which tone sequences such as the L% H- are anchored to positions such as the phrase edge makes their model much more like Kawakami's account than like Kindaichi's.

Pierrehumbert and Beckman's (1988) model of Japanese tone structure relied crucially on Bruce's seminal model of Stockholm Swedish tone patterns (Bruce, 1977, 1982, 1987, 1990). In Bruce's model, there are three types of tone which are anchored in different ways to designated constituents or positions at several levels of a hierarchy of prosodic units. The first two relevant levels are the grouping of consonant and vowel constituents into short (unstressed) and long (stressed) syllable constituents, and the grouping of unstressed syllables together with neighboring stressed syllables into word constituents. The second level is marked tonally by the *word accent*, a H+L tone sequence that is anchored to a designated strong syllable in each word. This culminative distribution of the H+L sequence means that in longer Swedish utterances, words can be counted off in the melodic contour for an utterance by recognizing the word-accent tones and their anchoring points. Above the word level, whole utterances and prosodic phrases within utterances are delimited by boundary tones such as the L% for the "terminal juncture fall" (Bruce, 1983, p. 223). Also, the melodic contour for each phrase must include a H- tone, called the *sentence accent* in Bruce (1977), the *phrase accent* in Pierrehumbert (1980), and the *focal tone* in Gussenhoven and Bruce (1999). The focal tone is realized just after the word accent of the word with "sentence stress" – i.e., a word that is in narrow focus in the discourse context or the last word in the phrase when there is broad focus over the whole phrase. All the tone types are shown in the sample transcriptions in (5). These schematic "pitch contours" are based on f_0 tracks given in Bruce (1977) and are intended to give a sense of the typical patterns of truncation and undershoot.

(5)



a. mellan målen 'between meals'
%L H*+L H+L* H- L%

b. mellanmålen 'snack'
%L H*+L H- L%

b. MELLAN målen 'BETWEEN meals [not AT meals]'
%L H*+LH- H+L* L%

As in most dialects of Swedish, the Stockholm variety has a lexical contrast between two anchoring patterns for the word accent, transcribed by Bruce (1990) as H+L* ("Accent 1") versus H*+L ("Accent 2"). In Accent 1 forms such as *anden* 'the duck', *anamma* 'accept', and *målen* 'meals' produced in contexts with one or more preceding syllables, the H+L* denotes a fall to a low pitch target within the stressed syllable that starts from a pitch peak or a high inflection point (an "elbow") about 120 ms before the low target. In Accent 2 forms such as *anden* 'the ghost', *lämna* 'leave', and *mellan* 'between' there is a peak or high elbow within the stressed syllable and a fall to a valley or a low elbow 120 ms later. A compound word such as *mellanmålen* 'snack' is marked by a H*+L (Accent 2) anchored to the designated syllable of the first component and no word accent on any later component. This accenting in compound words mirrors the typically initial stress in the native Germanic stratum of the lexicon.

Other important concepts are truncation and undershoot. When an Accent 1 word with initial stress is initial in its utterance, the leading H of the accent will be effectively "hidden" by the preceding silence, so that the underlying H+L* is truncated to be just the L* target on the designated syllable. Also, when an Accent 1 word with final stress is final in its phrase, the close succession of fall for the H+L* followed by rise to H- and fall to L% leaves very little room for the word accent to be realized. There is undershoot so that the L* is effectively a mid tone. By contrast, the trailing L of the Accent 2 fall is typically fully realized, because the designated syllable in an Accent 2 word cannot be final. Moreover, the duration of the transition from the H* target to the elbow for the trailing L is very stable. At the other extreme, the H focal accent has no very fixed constraints on its alignment other than that it occurs after the accent tones of the focalized word. In compound words, it is especially late, because the trailing L of the word accent has a secondary anchoring point at the stressed syllable of the second (or last) word in the compound. This account of the focal H- as a "floating tone" is invoked in AM-framework transcriptions by showing no line linking it to a designated syllable.

Two aspects of Bruce's work are especially noteworthy. The first was his rigorously controlled phonetic methods. He designed his materials to allow a systematic comparison of melodic contours for Accent 1 and Accent 2 words of different lengths in both final and nonfinal position and in both nonfocal and narrow focus contexts. This was necessary for him to be able to disentangle the tones that are specified by the lexicon from the tones that mark other levels of prosodic organization. He used analysis by synthesis to verify the segmentation of the melodic contour into these disparate elements and to examine their timing relative to the consonants and vowels at phrase boundaries and at the designated syllables within each phrase.

An equally important aspect of his work was his rigorously imaginative adaptation of key ideas from prosodic phonology. He did not let broad-stroke typologies dictate what analogies could be drawn between the tone patterns of Swedish and the intonational accents of English, and was among the first to grasp the implications for prosodic phonology of Bolinger's (1958) theory of pitch accent in English as interpreted by Vanderslice and Ladefoged (1972). He saw that the syllable bearing the word accent is not the only potential site for anchoring a tone target in a citation-form utterance of a Swedish word, and that tones realized at variable distances from the accented syllable in many dialects (including the Stockholm one) might reflect rhythmic organization above the word. This let him re-conceptualize the originally somewhat simplistic theory of a direct (i.e., prosodically unmediated) "association" between autonomously segmented tones and vowels as a more complex synchronization at "critical timing points" (Bruce, 1983, p. 234) that speakers and listeners control to resolve potentially conflicting demands in different phonological domains.

For the speaker, these conflicts involve "the interaction between the timing of phonatory and articulatory gestures" (Bruce, 1983, p. 222), which cannot follow an invariant rhythm because the words in a sentence can be one syllable or longer, initially accented or accented on a later syllable, in focus or subordinated to a neighboring constituent, and so on. Consonant and vowel gestures in a particular utterance of a string of words must be synchronized with each other so that the listener can parse the syllable count, hear whether each syllable is stressed or unstressed, and, if stressed, whether it is an extended vowel gesture or a coda consonant gesture that contributes the second mora in the syllable. Tone gestures also must be synchronized with the consonant and vowel gestures so that the listener can hear which stressed syllables are accented, whether an accent is H+L* or H*+L, and what word is highlighted by the focal H-. These different prosodic functions impose different demands. Realizing the word-level contrasts between short and long syllables and between H+L* and H*+L accents places stringent demands on the timing of the targets. In realizing the utterance-level contrast between focus and background, on the other hand, the exact timing of the focal H- is less relevant than achieving a particular target peak value, since the latter signals prominence relationships among words and phrases as well as among syllables within each word. Conflicts among these demands can be reconciled by adapting the tone targets (e.g., through truncation or undershoot) or by adapting the vowel and consonant targets (e.g., lengthening a final accented syllable to realize a complex sequence of word-accent tones, focal H-, and boundary tones, as suggested by Lyberg, 1981). To model the relevant interactions, the segments and tones must be observed in more contexts than citation form utterances.

In Bruce's original formulation of this "synchronization hypothesis" he differentiated between two orientations for evaluating the synchronization. From the "phonological point of view" of a "production oriented model" it is useful to specify the critical timing points for the underlying tone targets, but these may not map neatly onto the "perceptually critical" f_0 events such as rising or falling glissandi. For example, in his own perception experiments on the timing of the H+L* targets of Accent 1 versus the H*+L targets of Accent 2, Bruce found that the times of the starting and ending points traded off with the steepness of the fall in a way that suggested that subjects listened for the point of maximum velocity in the middle of the fall. However, reference to this midpoint time "is possible only in a sonorant environment" so that "from a perceptual point of view, it is probably an estimate of the timing of the entire f_0 change . . . that is decisive" (Bruce, 1983, p. 231).

Bruce's hypothesis was developed to account for the variable realizations of Swedish word accents across different sentence contexts and different dialects, but it was an important precursor to the AM model of Japanese tone structure presented in Pierrehumbert and Beckman (1988), as well as to the development of the Autosegmental-Metrical framework generally. Initially, development of the framework was addressed more to the production-oriented aims of finding "invariant" or "underlying" tone targets and their modes of association to phonologically defined positions in the hierarchy. For example, Pierrehumbert and Beckman (1988) proposed that the L% and H- tones in their model of Japanese are associated initially to the two accentual phrases on either side of the boundary that the pitch rise marks, but then that each tone is also associated secondarily at a later derivational stage to the first unaffiliated mora in the accentual phrase that begins at the boundary. They observed differences in the shape of the rise and in measured f_0 minima for what they called the "strong L%" versus the "weak L%" and they attributed these differences to a contrast between having and not having a secondary association to the first mora in the following accentual phrase. Gussenhoven and Bruce (1999) similarly propose to account for the shape of the trough in citation form utterances of compound words in Stockholm Swedish in terms of a secondary association of the trailing L of the H*+L accent. Grice et al. (2000) catalogue other examples of languages where a phrase accent can be analyzed as having a dual affiliation to both the edge of a larger prosodic domain and to some designated syllable within the domain. This focus on tone targets and their anchoring relative to the prosodic structures that the speaker controls is congruent with the production-oriented approach of the Articulatory Phonology framework (e.g., Browman & Goldstein, 1990; Byrd, 1996). For example, Xu and Liu (2007) apply a model of Putonghua lexical tone alignment to examine syllables of both Putonghua and English in order to probe for universal patterns in how an onset consonant gesture is anchored to its syllable to be coarticulated with the relevant vowel. This application suggests some of the questions that can be addressed fruitfully using production-oriented models that assume invariant underlying tone, vowel, and consonant targets for the speaker that are aligned with each other to reflect the "temporal signature of prosody" (see Fletcher, this volume, section 2).

Other recent work, however, suggests that the time is ripe to begin to reorient our models to incorporate constraints on the listener, too. For example, Arvaniti et al. (1998, 2000) show that the timing of prenuclear rising accents in Greek does not fall out from a simplistic model that designates either the L or the H as the target that is associated to the designated syllable. Rather, the L is anchored just before the syllable-initial consonant and the H is anchored to coincide with the CV boundary in the following syllable. Unlike Swedish, Greek has only five vowels, with no prosodic contrast between short and long vowels or short and long consonants. Many syllables are CV and vowels tend to be quite short. Also, whereas many Swedish words follow the common Germanic pattern of root-initial stress, the position of the stressed syllable in a Greek word is constrained only to occur on one of the last three syllables. Within this three-syllable window stress placement is "phonologically unpredictable" (Arvaniti, 1999, p. 171). Given these characteristics of the language, the observed anchoring pattern for Greek prenuclear rising accents may have emerged as a way to provide the listener with a robust "estimate of the timing of the entire f_0 change" in order to reliably parse the location of each accented syllable in an utterance.

These demands on the Greek listener are different from the demands on the listener from a language such as Dutch, where there are many more than five vowels in the inventory, vowels are typically longer, and there is also a much larger variety of typical syllable structures, including a contrast between syllables with short vowels and syllables with long vowels. Ladd et al. (2000) show that in Dutch, the timing of the end of a rising accent is not fixed in the same way as in Greek. Rather, it is later relative to the end of a syllable with a short vowel and earlier relative to the end of a syllable with a long vowel, and this difference in anchoring of the endpoints supports the vowel length contrast even for speech rates and discourse contexts where the vowel durations themselves are not robustly different.

Arvaniti et al. (2000) end their paper with a call for more research both to refine what "association" means for our models of the prosodic structures that the speaker intends to produce and to devise better methods for understanding how targets and their timing properties are realized in the speech signal that the listener parses. One promising line of research in this vein is comparative work such as Smiljanić (2006). Smiljanić looked at accent-related rises in standard Serbian and Croatian, language varieties which are mutually intelligible but which differ in whether there is a lexical contrast between word accents with an early versus a late peak. Smiljanić found that speakers of both varieties signal focal prominence on a word by manipulating the timing as well as the maximum f_0 value of the pitch rise to the accent peak. However, the timing effect is much smaller in Serbian, where the anchoring of the rise also signals the contrast between the two word accent types. We need more such comparative work in other language groups to develop our understanding of the potential role of functional load in the interaction between demands on production and demands on perception. We also need more work that does what both Bruce (1977) and Smiljanić (2006) did – namely, to observe tones in words across a good variety of sentential and discourse contexts, to see how variation in the demand for precise "horizontal" anchoring of tone targets relative to critical positions within a word interacts with variation in the demand for precise "vertical" positioning of the tone targets relative to the tonal space.

3.3 Tone scaling and the tonal space

The third key development in prosodic phonology was the idea that speakers can raise or lower and expand or compress the local tonal space as a whole and also independently scale tone targets up and down within the tonal space, to reflect both autosegmental contrast and relative metrical strength, as well as other sorts of linguistic (or "paralinguistic") relationships. While there is a broad consensus that this separation of "vertical" position into two parts is necessary, the separation is realized differently in different AM-framework models, and the linguistic nature as well as the formal status of the independence remain controversial. We will illustrate with the independence of tone-scaling and tonal-space specification parameters in an AM-framework model for Tokyo Japanese that was originally developed and tested in a synthesis system by Pierrehumbert and Beckman (1988) and subsequently modified for the X-JToBI conventions (Maekawa et al., 2002) that were developed for tagging the Corpus of Spontaneous Japanese (Maekawa, 2003). The separation of parameters in this model corresponds roughly to the specification of variable accentuation levels for turning points independent of the parameters of the tonal grid in the Lund model and to the independent specification of amplitude values for the accent commands and phrase commands in the Fujisaki model.

As noted earlier, we have adopted Ladd's (1992) term "tonal space" to talk about the effects that the IPO-framework models generate by specifying variable starting values and slopes for declination lines over different parts of an utterance. Ladd chose this term to have a framework-neutral way of referring to what Chao (1930) called the pitch "range" when he proposed his "system of tone letters" and the corresponding numerical notation that we used to indicate the lexically contrastive pitch pattern on each of the syllables in the transcriptions of the Putonghua utterances in the caption to Figure 16.1. Chao (1932, p. 124) identifies "several abstractions" that must be made to record the pitch patterns that differentiate the tone classes in any dialect of Chinese. Specifically, each pitch level must be calculated "relative to the speaker's range of voice, so that what would be a low tone for a soprano is actually higher in pitch than the high tone of a tenor." Moreover, "the range of pitch between different tones and within the limits of moving tones is also a variable quantity depending on force of articulation and force of vocalization."

The abstraction over different speakers' voices is analogous to the abstraction over different vocal tracts when computing targets in some speaker-normalized representation of the vowel formant space. The abstraction over variable "force of articulation" is analogous to the constancy of vowel-class identity across the hyperarticulation–hypoarticulation continuum (Lindblom, 1990; see review by Harrington, this volume, section 2.5). An important difference between these two spaces is that the "force of articulation" and the "force of vocalization" effects on vowel formant values are necessarily small compared to speaker effects, because maneuvers such as contracting the strap muscles to lower the larynx can change a soprano's vocal tract length by only a small amount relative to the typical length difference between her vocal tract and a tenor's. By contrast, the "force" effects on pitch values can be extremely large relative to the differences across speakers, so that the soprano's H tone in a very subdued speaking style can be much lower than the tenor's H tone in a very forceful speaking style. A phonological consequence of this difference between the phonetic spaces is that when force of articulation and force of vocalization effects on vowel formant values are phonologized as linguistically significant markers of strong versus weak positions in the prosodic hierarchy of a language, the markers typically can be described in terms of a small number of discrete prosodic constraints on what vowel targets can be specified for moras or syllables in different positions of the hierarchy.

Analogous prosodic constraints on what tone targets can be associated in different positions are fairly common across spoken languages (cf. section 4.3), but an even stronger universal is the phonologization of the control parameters for positioning tones within the tonal space and for varying the dimensions of the tonal space itself so that these can act not just as discrete markers of the set of categorical contrasts in prosodic organization, but also as gradient markers of more subtle differences in relative metrical strength as well as of other linguistic scales.

The bottom panel of Figure 16.4 illustrates the parameterization of the tonal space and of tone scaling that Pierrehumbert and Beckman (1988) built into their synthesis model for Japanese, as these parameters are understood in the version of this model that was incorporated into the X-JToBI labeling conventions on the basis of later research that is reviewed in Venditti et al. (2008). In this model, there are tonal-space or tone-scaling effects that refer to three different types of prosodic constituent – the intonation phrase (IP), the accentual phrase (AP), and the prosodic word (W).[4]

At the beginning of the first IP, the reference line that defines the bottom of the tonal space is initialized to reflect overall engagement or volume within the speaker's voice range. The reference line for the utterance in Figure 16.4, for example, is initialized at 70 Hz. This value is maintained until late in the last IP of the utterance, where the effect of "final lowering" begins to be noticeable, to signal discourse-level functions such as topic shifts or yielding of the floor to the other speaker. Final lowering is a change in the reference line time function, from having a fixed value to showing a decline over some span at the end of an IP. In the turn-final utterance shown here, for example, the effect reaches in to lower the reference line by 44 Hz per second starting at 0.45 seconds from the end of the last IP.

The IP is also the level of prosodic structure where the value for the top of the tonal space is (re)initialized. The initial topline values for the three IPs in the utterance in Figure 16.4, for example, are set at 130, 66, and 110 Hz above the reference line.

The IP is also the domain of downstop, a compression of the tonal space triggered at each lexical accent. This effect is implemented in the model by reducing the distance of the topline from the reference line by a fixed ratio. The downstep ratio in the first IP that is triggered by the accent on the first syllable of the prosodic word *Yamano*, for example, is 0.62 – compressing the tonal space to 62 percent of its original span.

Tone targets at the level of the IP, the AP, and the W are then positioned within the local tonal space that is defined by the additive effects of the initial IP topline specification, the compression at each previous downstep, and edge effects such as final lowering. Position within the tonal space first of all defines the discrete contrast between H tones (the targets that are closer to the topline) and L tones (the targets that are closer to the reference line). The level of the AP, for example, is defined by the rise from the %L or L% boundary tone to the H- phrase tone. At the level of the W, the lexical contrasts among accented and unaccented words are expressed by the presence and (if present) the location of the H+L accentual fall.

The top and bottom of the tonal space also act as a reference for continuous within-category contrasts in metrical strength. Stronger L tones are scaled lower, to be closer to the reference line, and stronger H tones are scaled higher, to be at (or even above) the topline. Some of these strength contrasts are intrinsic to the tone target type. Within an AP containing an accented W, for example, the H tone of the H+L word accent is intrinsically stronger than the H- of the phrase-initial rise; it will be higher relative to the topline, other things being equal. Other strength contrasts are extrinsic and reflect other types of linguistic structure, such as the discourse-level differentiation between given and new information. The imagined context for the performance of the utterance in Figure 16.4, for example, is a conversation between two spectators at a triathlon relay race. The other speaker has just asked whether the athlete who is swimming could be Yamano. The H of the word accent in the first AP goes above the local topline to reflect the discourse-level prominence of *Yamano* as a contrastive topic. The H of the word accent in the second AP is much lower, reflecting both the compression of the tonal space at the downstep and also the given status of the verb *oyoideru* 'is swimming' in this dialogue context.

The effects that are illustrated in Figure 16.4 are parameterized in somewhat different ways in the Fujisaki model that Hirai et al. (1997) used to analyze several large multispeaker corpora in order to develop the intonation component of CHATR, a concatenative speech synthesis system with prosody-based unit selection (Campbell & Black, 1997). For example, in the Fujisaki model, downstep is not modeled explicitly, but instead falls out from the choice of amplitude values for successive accent commands. At the same time, there are important commonalities between these two models. In particular, both models encode relative prominence relationships among tone targets using two different sets of parameters. In the Fujisaki model, there is a step function (the phrase command) to (re)initialize the backdrop tonal space at the beginning of each new IP and a matched pair of step functions (the accent command) that generates the rise to the first H target in each AP (as well as the fall at the accent or at the end of the AP if there is no accented W in the phrase), and the amplitude of each of these two commands is a continuously variable parameter. That is, the distinction between these two amplitudes corresponds roughly to the distinction between the tonal space parameters that are initialized at the level of the IP and tone-scaling parameters that are specified for the tone targets that are obligatory at the level of the AP in the AM-framework model depicted in Figure 16.4.

One critically important difference between the two models is the treatment of L-tone scaling. As noted already, prominent L tones are scaled downward toward the reference line in this AM-framework model. In the utterance in Figure 16.4, for example, the L% tone at the IP boundary after *oyoideru-ga* is lower in the local tone space than the L% tone at the mere AP boundary after *Yamano-wa*, reflecting the difference in metrical strength between those two positions in the prosodic hierarchy. In Osaka Japanese, where there is a contrast between %L-beginning and %H-beginning words, there is a similar downward scaling of this initial L tone as well as a delay in the beginning of the following rise in L-beginning unaccented words under focal prominence (Kori, 1987) and in pragmatically loaded questions (Miura & Hara, 1995). These effects would be difficult to model in the Fujisaki framework without introducing another type of (downward pulsing) accent command that can be positioned at places other than the beginning of the AP.

Another critical difference is in the treatment of effects such as final lowering. Beckman and Pierrehumbert (1986) suggest that extreme final lowering defines one end of a continuum which has (at the other end) an effect that they call "final raising" which they observed in syntactically unmarked questions. As already stated, in the AM model in Figure 16.4, this kind of edge-in effect is modeled directly as a change in the shape of the tonal space at the end of some phrasal grouping, analogous to the way that phrase-final lengthening and initial strengthening are treated in the π-gesture model of Byrd and Saltzman (2003) and other Articulatory-Phonology framework models (see review in Fletcher, this volume, section 2.2). In the Fujisaki framework, by contrast, such edge-in upward or downward slope differences cannot be modeled directly. There is a necessary downtrend across the tonal space for the whole IP, because the phrase command impulse is smoothed by a filter function and the resulting curve is convolved with the concurrent accent commands, each of which is also smoothed by a different filter function. However, since these filter shapes are intended to reflect "hard" physiological constraints (cf. Öhman, 1967; Fujisaki, 1983), they are not under the speaker's direct control. In order to vary the slope as a way of marking structural properties such as the discourse property of being turn final, the modeler must insert a phrase command with just the right amplitude at some place near the end to counter the downtrend from the damping function. The inability to model systematic slope variation directly makes the Fujisaki model fundamentally different not just from the AM-framework model of Japanese, but also from Grønnum's very different model of functionally similar effects in Standard Danish (Thorsen, 1983, 1985, 1986).

Grønnum's model, on the other hand, is fundamentally different from both the Fujisaki model and the AM model in that all aspects of the tone pattern are treated in terms of a hierarchy of trend lines, with slopes that are specified for the nested spans of the individual stress groups within individual clauses within a semantically coherent text. Factors affecting these slopes are the length of the span (e.g., the clause-level slope is very steep for clauses containing fewer stress groups and very shallow for clauses containing many stress groups) as well as the same discourse-level factors that Beckman and Pierrehumbert (1986) identify as the function of edge-in effects such as final lowering.

As Grønnum (1990, p. 199) points out, the Danish effects are formally distinct from the Japanese effects in that they are global and not localized to the phrase end. The downtrend that signals finality "does not just reach one half-second in from the end, it reaches in all the way back, across several . . . stress groups to the onset of the utterance." In Liberman and Pierrehumbert's (1984) AM-framework re-interpretations of Grønnum's results, this longer-range clause-level slope function is modeled in terms of downstep triggered locally at each successive accent. The speaker would have to be able to specify a different downstep ratio at the beginning of each IP in order to simulate the difference in slope between a final and nonfinal clause. The even longer-range slope of the text, on the other hand, is modeled in terms of the speaker's specific choices for reference line or topline initialization values for the successive phrases. Grønnum (see Thorsen, 1984, p. 307) criticizes this "local" approach as arbitrarily allocating responsibility to disparate sets of formal parameters to account for the functionally uniform hierarchy of syntactic and semantic coherence.

Grønnum (1995, p. 348) voices a related criticism in her review of the equally "local" treatment of downtrends in Möbius' (1993) Fujisaki-framework model of German. Specifically, she points to his results that the amplitude of the phrase accent command depends on the number of accents and also (in short sentences) on whether the accent is early or late. After quoting from Möbius' comparison between his more "local" approach and her "hierarchical" one, she agrees:

> That is exactly . . . the problem with Fujisaki's model as adapted by the author: it permits phrase command amplitudes to depend on accent location, and it does not supply criteria for a principled choice between several sets of phrase and accent parameters which each render an acceptable f_0 copy of an original, if such are conceivable. And that, I think, is incompatible with a model which purports to be physiologically *and* linguistically motivated.

Ladd (1992, 1993) and others point to a comparable "degrees of freedom" problem for the tone-scaling and tonal-space parameters of the AM-framework models of English and Japanese associated with Pierrehumbert and her colleagues, but it surely is a problem for Grønnum's model, too, once one goes beyond carefully scripted lab speech. Indeed, this indeterminacy will be a problem for any analysis by synthesis model that is sophisticated enough to simulate the ways in which tonal space and tone pattern interact in speech but relies exclusively on goodness of f_0 fit as a criterion for choosing among parameter settings. In short, there are very pressing research questions that need to be addressed before we have a good model of tone scaling and its relationship to tonal space control, including the overarching one that Grønnum identifies in her review of Möbius' model: What kind of criteria can be applied to distinguish among models or among different parameter settings within a model?

As noted in Grønnum's review, Möbius defends his choice of framework on the grounds that the tonal space parameters in the Fujisaki model are physiologically motivated. The basis for this claim is in Öhman's seminal model of Scandinavian "word and sentence intonation" in which he posited two distinct laryngeal gestures for word accents and sentence-level patterns, and suggested that these could be identified with independent activation of two different parts of the cricothyroid muscle (Öhman, 1967, pp. 29–30). Fujisaki (1983, pp. 53–4) follows Öhman to posit the same physiological correlates for the different damping functions that he proposed for the phrase command and the accent command. Work on the control of f_0 in speech has not supported this idea. Neither has it identified evidence of separate "gestures" for tonal-space versus tone-scaling parameters, because there is no compellingly obvious way to conceptualize the task space. In this respect, the articulation of f_0 is fundamentally different from the articulation of spectral correlates of consonant constrictions. There are some suggestive ideas in work on physiological correlates of tone and pitch accent contrasts, such as Gårding et al. (1970), Erickson (1978, 1993), Erickson et al. (1995), Beckman et al. (1995), Hallé (1994), and Sugito (2003). There is also research such as Herman (2000) and Epstein (2002), documenting perceptible differences in vowel amplitude and voice quality associated with the final lowering effect. These non-f_0 correlates perhaps could help in conceptualizing the task space for tonal-space gestures if examined at the articulatory level, as suggested in Herman et al. (1996). However, the interactions among laryngeal tension, vocal fold thickness, and pulmonary effort are complex and not completely understood. There looks to be a great deal of basic research yet to be done before physiological evidence can be brought to bear directly on the degrees of freedom question.

Another avenue of attack that may yield more immediately applicable criteria is to develop experimental paradigms for assessing whether native listeners treat tone scaling and tonal space separately, as in Herman (2000) and Gussenhoven and Rietveld (2000). Such experiments might be especially useful if paired with studies designed to pin down the meaning differences associated with minimally contrasting melodic contours where tone scaling or tonal space differences seem to act as a primary cue or as an enhancing secondary cue, as in Hirschberg and Ward (1992), Grice and Savino (1995), Venditti et al. (1998), Caspers (2000), and Lee (2000, 2005). As Gussenhoven (1999) points out, however, this avenue of research requires that we look more closely at the types of linguistic functions that are linked to different formal parameters in different languages, and think carefully about how particular experimental tasks might preclude discovery of the use of some pattern of tones, tonal anchoring, tone scaling, or tonal-space settings for a particular function. This highlights the fact that we need a better understanding of the range of linguistic functions that can be encoded in spoken language melody and of how these functions are realized in related language varieties as well as in different language families.

4 A Taxonomy of Linguistic Functions

In this section, therefore, we will briefly describe some of the functions that have been identified, beginning with the "tonemic" function of constituting a small finite set of meaningless contrasting patterns that can be combined with elements from other sets of meaningless contrasting patterns (consonant constrictions and vowel postures) to build an indefinitely large lexicon.

4.1 Tonemes and tonal morphemes

The basic "tonemic" function is most easily illustrated with utterances and words from a language such as the standard Hong Kong dialect of Cantonese. In this

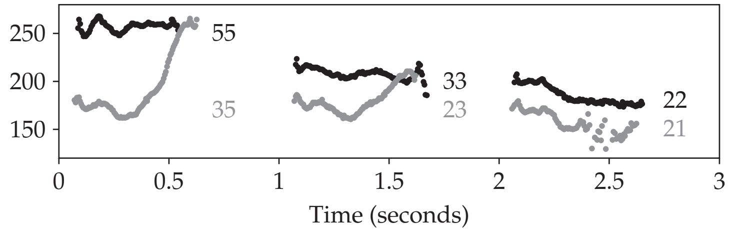


Figure 16.5 Example f_0 contours of citation intonation utterances of the level tone wordforms (black) and contour tone wordforms (gray) listed in (6a) produced by an adult female native speaker of Hong Kong Cantonese. The utterances are from Wong et al. (2005).

variety, most words are monosyllabic (that is, any given *zì* probably is a word), and every syllable is specified for one of the tone patterns exemplified by the contrasting wordforms in (6).

(6)	a.	[wɐi^{55}]	'power'	[wɐi^{33}]	'fear; pleasant'	[wɐi^{22}]	'guard'
		[wɐi^{35}]	'position'	[wɐi^{23}]	'surround'	[wɐi^{21}]	'person'
	b.	[wɐʔt^{5}]	'dense'	[wɐʔt^{3}]	'revolve'	[wɐʔt^{2}]	'kingfisher'

Figure 16.5 shows example utterances of the six wordforms with sonorant rhymes in (6a) produced as citation form sentences. The extremely low onset of the toneme that is transcribed with [^{35}] reflects a sound change in progress in the Hong Kong standard dialect (see So, 1996, who transcribes it as [^{25}], and reviews the literature on this and other recent tone changes and merges). The black and gray f_0 tracks in Figure 16.6 illustrate how the mid-level tone of the homophones meaning 'fear' and 'pleasant' is realized in two other intonational contexts. The morpheme just before [wɐi^{33}] 'pleasant' in the utterance plotted with gray in that figure also has this same mid-level tone. The pitch perturbation at the syllable boundary is a juncture-marking creaky voice quality that sets off and emphasizes the final word, which is as long as the total duration of the first four morphemes of the utterance. All of the earlier morphemes in this utterance, as well as in the utterance plotted with the black line, are shortened by a process that Wong (2006) calls "syllable fusion." When morphemes are conjoined into compound words or frequently uttered phrases, speakers can signal the particularly close juncture by weakening or deleting medial consonants and merging the two syllables' vowels. Except in the most extreme cases, however, the percept of each syllable's tone specification is preserved to maintain the syllable count. Thus in this variety of Cantonese the tone specifications are contrastive properties of syllables fully on par with such properties as the palatalized offglide in the rhymes in (6a) as opposed to the glottalized plosive coda in the rhymes in (6b). The typical shapes of words in combination with the extremely "isolating" or "analytic" nature of the grammar drives a robust segmentation of the melody into syllable-anchored tone units.

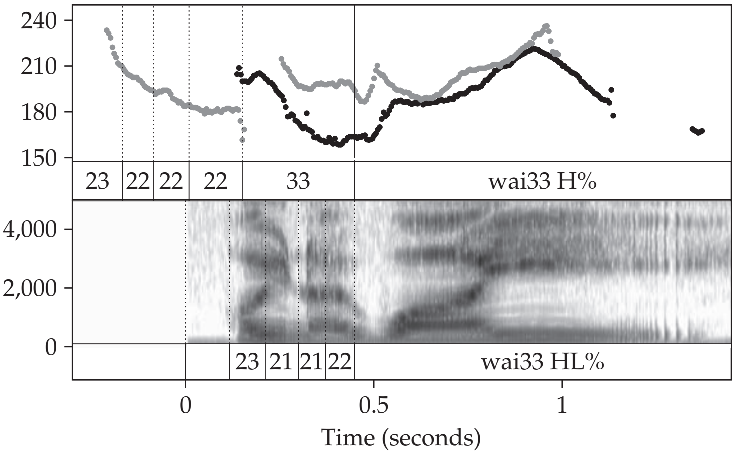


Figure 16.6 Spectrogram and f_0 contour (black) for utterance of the sentence [o^{23}jyn^{21}loi^{21}hɐi^{22}wɐi^{33}]+HL% ‘Oh I get it! The word was <fear>.’ with overlaid f_0 contour (gray) for utterance of [keoi^{23}jɐu^{22}wɒ:^{22}hɐi^{22}fɒ:i33wɐi^{33}]+H% ‘She said then that the word was <pleasant>?!’ produced by the speaker who produced the utterances in Figure 16.5. The f_0 contours for the two utterances are aligned at the onset times of the final homophonous words [wåi33] (solid cursor). The utterances are from Wong et al. (2005).

The pitch patterns on the final syllable in Figure 16.6 illustrate another way in which tones can function in lexical contrast. The [wai^{33}] syllable in each of these utterances is prolonged to be three or five times the average length of syllables earlier in the utterance. This prolongation leaves room for the realization of one or two more tone targets that are transcribed using the notational conventions described earlier for the transcriptions in (1)–(5). This “code-switching” between transcription systems follows the C-ToBI conventions proposed by Wong et al. (2005) to clearly distinguish the morphemic function of the tones transcribed as H% and HL% in these utterances from the tonemic function of the tones transcribed as [^{23}], [^{22}], [^{21}], [^{33}], and so on in (6). The meanings of these two morphemes H% and HL% are reflected in the glosses. The H% at the end of the utterance in gray makes it an incredulous echo question as indicated by the “?!” at the end of the gloss, whereas the HL% at the end of the utterance in black imparts the sense of discovery or sudden realization glossed by the ‘Oh I get it!’

Cantonese has an extremely rich set of pragmatic morphemes like these final boundary tones. Many of these sentence particles are composed of vowel and consonant phonemes as well as the tonemes, but several of them are just the toneme affixed to the final content word, as illustrated here. The H% boundary tone in the Beijinghua utterance in the right panel of Figure 16.3b, similarly, is one of two tonal morphemes among the 28 sentence particles that Chao (1968) counts. The count for any of the Mandarin dialects is somewhat easier, since the nontonal components of the sentence-final particles of Mandarin are analyzed as being neutral tone and combinations of particles are never counted separately from the particles that are simple syllables or simply tonal. By contrast, counts for Cantonese range widely (as many as 206 by Yau's, 1980, count), depending on whether polysyllabic sequences, monosyllabic particles that are potentially fused forms of polysyllabic sequences, and other complex forms are counted separately. For example, Law's (1999) count of between 35 and 40 includes sets that are traditionally described as being minimally differentiated by tone, such as the minimal pair [tse^{55}] and [tseʔk^{5}] studied by Chan (1998). Fung (2000) suggests that Cantonese sentence particles such as these can be grouped into a much smaller number of "families" of phonologically related particles that have a common core meaning. That is, she proposes that the meanings of [tse^{55}] and [tseʔk^{5}], for example, can be analyzed in terms of the core meaning of the [ts] family in combination with the meanings of the tones (which correspond to the tonal morphemes transcribed in C-ToBI as H% versus -%). Sybesma and Li (2007) analyze Fung's families in more detail, and propose that each of the 40 most common sentence particles is composed of three parts: (1) an onset morpheme that is either the default null (glottal stop) initial or one of the fully specified consonants [h, k, l, m, l~n, ts]; (2) a nucleus morpheme that is either the default vowel [e:] or one of [ɐ:, ɔ:]; and (3) a tonal morpheme that is either the default [^{3}] (tagged as a protracted neutral target :% in C-ToBI), [ʔ^{4}] (tagged as -% in C-ToBI), [^{5}] (H%), or [^{1}] (L%). By this analysis, then, the HL% transcribed for the utterance plotted in black in Figure 16.6 might be a compound of Sybesma and Li's tonal morphemes [^{5}] and [^{1}], or it might be the tonal morpheme corresponding to [^{51}], which as a toneme has merged with [^{5}] in the Hong Kong dialect (see So, 1996, among others).

The difficulty of counting the number of Cantonese sentence particles as compared to the ease of counting the nine Cantonese lexical tonemes in (6) is noteworthy. It may reflect the elusiveness of pragmatic "meaning," which is difficult to paraphrase outside of the specific contexts where a pragmatic morpheme is appropriate, as compared to the stark difference in referential meaning that lets us identify the polysemous nature of the wordform [wɐi^{33}]. It also may speak to a more basic difference between tonemes and tonal morphemes that stems from the design principle of duality of patterning (Hockett, 1960) – i.e., the principle that the lexicon of any human language is a self-diversifying system in which a small number of discretely different elements can be combined to make a large number of potentially extremely complex morphemes without losing their discrete distinctiveness (Goldstein & Fowler, 2003). Consider the analogous difference for vowel segments. It is easy to recognize that the vowel phonemes in the two syllables of the English compound *A-frame* are the same but that the vowels in the two syllables of *A-team* or *AWOL* are different. It is harder to say whether the vowel morpheme [e:] in the first syllable of each of these three words is the same or even whether *AWOL* is polymorphemic in the way that *A-frame* obviously is.

These two sources of difficulty have long confounded the analysis of the tonal morphemes of English. Is the tune in the second phrase of Figure 16.2 a sequence of four tonal morphemes, as suggested in the transcription in (1), which follows the analysis in Pierrehumbert and Hirschberg (1990)? Or is it two tone morphemes H*L H*LH to which a linking rule has applied to anchor the L of the first morpheme to the second stressed syllable, as proposed by Gussenhoven (1984)? What kinds of experiments can we use to differentiate between these two morphological analyses? Ladd (2008, ch. 4) gives an insightful description of the difficulties for English and a few other related languages, as well as a review of arguments advanced by proponents of different analyses and of the relevant experimental studies.

4.2 *Prosodic grouping*

In describing the "tonemic" function using Cantonese examples, we emphasized the monosyllabic word shapes and isolating morphology of the language, because the more general function of lexical contrast will be realized using very different segmentation and anchoring parameters in a language where words are polysyllabic or the grammar is of a more "agglutinative" or "synthetic" nature. For example, every modern Chinese dialect has a system of lexical tone contrasts that is a reflex of the same original tone categories that give rise to the Cantonese tonemes in (6), but in a Wu dialect, the tone pattern that corresponds to a toneme of a Cantonese word typically will not be realized in the same way on the cognate Wu morpheme. Words are typically at least two syllables, and very productive morphological processes (typically called "tone sandhi" – see Chen, 2000) insure that just one toneme is specified for each compound word or phrasal construction in an utterance. The Shanghai examples in (7) are from Zee and Maddieson's (1980) study, and the schematics are based on the f_0 tracks they show. The compound nouns in (7c) and (7d) are derived from the sets of four *zì* in (7a) and (7b), respectively. These examples illustrate the tone sandhi processes that relate the patterns of derived words to the citation form tone patterns of the *zì* from which they are derived. The most general description is what Chan and Ren (1989) call "Pattern Extension"; the underlying toneme of the first *zì* is the only one realized, and its component tones are extended to cover the whole word or phrase, as in (7c).

(7) a. [ɕin^{51}] 'new' [vəŋ^{14}] 'to hear' [tɕi^{34}] 'to record' [tse^{51}] 'person'
HL LH MH HL

b. [tɕiʔ^{5}] 'to unite' [h^{w}əŋ^{51}] 'matrimony' [tsəŋ^{34}] 'proof' [sɹ^{51}] 'book'
H HL MH HL

c. [ɕin.vəŋ.tɕi.tse] 'reporter'
H L

d. [tɕi.h^{w}əŋ.tsəŋ.sɹ^{51}] 'marriage license'
H L-

All of the Wu dialects use some variant of this Pattern Extension process, although details such as the typical phonological anchoring pattern may differ across tone types and across dialects. For example, Zee and Maddieson analyze the abrupt fall in (7d) in terms of a constituent-final tone that they posit for all such compound words, whatever the initial toneme, but they do not discuss the early anchoring point for the tone in some cases, such as (7d). Kennedy's (1953) description of a very similar abrupt fall in Tangxi compounds that have initial syllables with checked tone rhymes suggests an alternative analysis in which the abrupt fall is the realization over longer material of the creaky voice register that characterizes the checked tone. More recent work by Chen (2008) suggests that both analyses may be supported for variant realizations of longer compounds for at least some younger speakers. The cross-dialect differences in anchoring point can be appreciated by comparing the Shanghai falling-tone example in (7c) to the three Wuxi falling-tone examples in (8). The four examples in (9) are an alternative pattern for Wuxi compounds that begin with a falling-tone *zì*.

(8) a. [sɛ] 'three' b. [sɛ.ŋie] '3 years' c. [sɛ.dʌɯ.mɔʔ.dõ] '3 big wooden tubs'
H L H HL H HL

(9) [fi] 'fly' [fi.tɕi] 'airplane' [fi.tɕi.p^{h}iʌ] 'air ticket' [fi.tɕi.p^{h}iʌ.tɕia] 'airfare'
HL L LH L LH L LH

The transcriptions and schematics in (8) and (9) are based on the descriptions and f_0 tracks in Chan and Ren's (1989) account of the history of two different morphological processes that they identify in this dialect. They describe the Pattern Extension process in Wuxi as typically applying to number+classifier expressions, as in (8), and also to reduplicated verbs, verbs with resultative or directional complements, and reduplicated nouns in child-directed speech. The Wuxi "Pattern Substitution" process in (9), by contrast, is typically applied to verb phrases with direct objects, to reduplicated nouns in the adult lexicon, as well as to the very productive compound word formation process illustrated in (9), where [fi^1.tɕi^{14}] is 'fly machine', [fi^1.tɕi.^{1}p^{h}iʌ^{14}] is 'fly machine ticket', and [fi^1.tɕi.^{1}p^{h}iʌ^1.tɕia^{14}] is 'fly machine ticket price'. Chan and Ren relate these two Wuxi processes to a contrast that Kennedy (1953) describes for the Tangxi dialect, where the morphosyntactic difference is clearer. When the two Tangxi processes apply in combining the morphemes [tsɔ^{51}] 'to fry' and [vɛ^{24}] 'rice', the Pattern Extension process yields the compound noun [tsɔ^5.vɛ^1] 'fried rice' whereas the second type of process yields the verb phrase [tsɔ.vɛ^{24}] 'to fry rice'.

Despite the differences across the examples in (7)–(9), however, the function is essentially the same. The toneme specification is a property of the constituent as a whole, and the boundaries between successive constituents are marked by a transition to the next lexically contrastive tone pattern. The contrasting melodic contours, then, effectively group strings of syllables into coherent prosodic constituents (tone sandhi groups) that align to constituents or domains specified by other parts of the grammar. When utterances are short and decontextualized, as in the utterances examined in Zee and Maddieson (1980) and Chan and Ren (1989), the domains are described in terms of morphosyntactic relationships. When utterances are longer or produced in more elaborated discourse contexts, other types of relationship, such as the articulation of an utterance into topic and focus or given and new, come into play, as discussed by Selkirk and Shen (1990) among many others.

This same function of prosodic grouping is invoked by Carter's (1974) description of the "tone-phrasing system of Kongo" and by our description of the distribution of L% and H- tones in Japanese in section 3.1 above. It also is a critical element in Halliday's (1967, p. 9) description of English utterances as "an unbroken succession of tone groups each of which selects one or another of the five tones" as well as of Pierrehumbert's (1980, p. 19) definition of the "tune" in English as "the melody for the intonation phrase." As should be obvious from this list of languages, as well as from the differences between Cantonese and Shanghai, the ways in which melody is harnessed for the function of prosodic grouping are orthogonal to the ways in which melody is harnessed for the function of lexical contrast. Cantonese and Shanghai have inherited the same set of tone categories from their common ancestor language, but Cantonese does not have any morphological process like these "tone sandhi" processes in Shanghai and the other Wu dialects and instead uses the consonant- and vowel-focused process of syllable fusion. Thus, the surface melodies of cognate compound words and phrases make for very different tone groups in the two languages. The modern Japanese dialects offer the complementary evidence, for a double dissociation. Although the accentual phrase melodies of Japanese mark off analogous tonally delimited prosodic phrases in very similar ways in the Tokyo and Osaka dialects, the tone at the AP boundary in Tokyo is invariantly L, whereas Osaka preserves an older tonemic contrast between %L-beginning and %H-beginning words.

4.3 Metrical prominence

In accounting for the melodic differences between the disyllabic compound noun [tsɔ^{5}.vɛ^{1}] 'fried rice' and the verb phrase [tsɔ .vɛ^{24}] 'to fry rice' in Tangxi, Kennedy (1953) talked about the prosodic grouping function that the two patterns have in common, but he also described differences in the "stress pattern," with the compound noun pattern having "louder stress" on the first syllable and the verb phrase having it on the second. A related segmental difference is specific to the checked tone; the glottal coda in morphemes such as [baʔ^{3}] 'white' or [tɕ^{h}ʊk^{5}] 'to drink' can trigger gemination of a following syllable onset in the compound-noun pattern but not in the verb-phrase pattern, as in [bas^{3}.sɛ^{51}] 'white water' versus [tɕ^{h}ʊ.t^{h}saŋ^{33}] 'to drink soup'. (This condition on the gemination of there being immediately preceding stress at the word boundary is similar to the condition of word-final stress on Raddoppiamento Sintattico in many varieties of Italian – see, e.g., Vayra, 1994; D'Imperio & Gili Fivela, 2003.)

In other Wu dialects, also, the syllable in the tone sandhi group that is associated with the toneme bears segmental hallmarks that are associated with phrasal or lexical stress in other languages. For example, Zee (1990) documents a process of vowel lenition in Shanghai whereby the high vowels of the language [i, ɹ, u][5] can be devoiced or deleted in certain environments. This is essentially the same process as the "syncope" that Cedergren and Simoneau (1985) describe for [i, y, u] in Quebec French, and the "reduction" that Dauer (1980) describes for [i, u] in Greek, which Arvaniti (1994) uses as a metric of stress on pretonic syllables. As in these other two languages, devoicing in Shanghai is a variable process that depends on speech rate as well as on the identity of the neighboring consonants. It is also tonally conditioned. In Quebec French, syncope never affects vowels in the final syllable of the constituent that Cedergren et al. (1990) define as the domain of final pitch accent. This is the constituent that Jun and Fougeron (2002) call the accentual phrase, highlighting the demarcative function of the obligatory final pitch accent and the optional initial rise. In Greek, similarly, devoicing does not occur on the stressed syllable – i.e., the syllable which is associated with one of the tonal morphemes of the utterance melody. In Shanghai, too, devoicing never affects the first syllable in the tone sandhi group – i.e., the syllable to which the toneme is associated phonologically. Chao (1968, pp. 31, 141) also notes high-vowel devoicing in neutral tone syllables in the Beijing dialect of Putonghua, a relationship that he describes in Chao (1932, p. 129) in terms of the notion "stress-accent" or "tonic stress":

> Stress-accent does not play any important role in most Chinese dialects. But in a few dialects, including that of Peiping, tonic stress plays such an important part that unstressed syllables not only tend to have their vowels obscured, but also lose their proper tones, and acquire a level, usually short tone, the pitch being determined by the preceding syllable.

Thus, in all four of these languages, the property of being eligible to bear an associated toneme or tonal morpheme prohibits application of a process that weakens or deletes vowels. This is true both of Beijing Mandarin and of Greek, where the location of this "tonic stress" is not predictable from the prosodic grouping into words and phrases, and of Shanghai and French, where the fixed position of the "tonic stress" serves to demarcate the tone sandhi group or accentual phrase.

By contrast, high-vowel devoicing is not constrained by the tone pattern of the accentual phrase in either Japanese or Korean. Maekawa and Kikuchi (2005) observe devoiced vowels in the Corpus of Spontaneous Japanese in many syllables that are aligned to the phrasal H- in unaccented phrases or associated to the H of the H+L lexical pitch accent. Jun and Beckman (1994) likewise document pervasive high-vowel devoicing in a corpus of enacted lab speech dialogues of Korean, both in syllables that are AP-medial and in syllables that are associated to the LH or HH sequence that marks the beginning of the AP.

This reduction of vowels in the first syllable of the AP in these two languages contrasts with other segmental effects in this position. In Japanese, for example, older speakers who produce the nasal allophone of [g] in AP-medial positions (such as in the *-ga* particle in the *Yamano-wa oyoideru-ga* clause in Figure 16.6) do not produce [ŋ] in AP-initial position. Keating et al. (2003) and others show that the beginning of the AP in Korean also is a position marked by "initial strengthening" of the consonant (see Fletcher, this volume, section 2.2.4). The consonantal effects are more in line with the segmental effects of metrically strong position in languages with tonic stress. In Shanghai, for example, voiced stops are voiceless with breathy voice releases when they are onsets of syllables at the beginning of a sandhi tone group, but are voiced with short closures in tone-group medial position where the syllable does not bear a tone specification (Cao & Maddieson, 1992). That is, they show the same allophonic patterns with respect to the tone sandhi group that Jun (1993/1996) and others document for the Korean lax stops in AP-initial versus AP-medial position. There are similar effects in syllables with tonic stress versus syllables with neutral tone in the Beijing dialect. For example, the voicing of the [ts] in the second syllable of *fàng zài* in the left panel of Figure 16.3 is a cue to the neutral tone status of the syllable. One possible way to characterize the different treatments of the vowel in Japanese and Korean as compared to the other four languages, then, is to say that the vowels are less important than the consonants in defining the syllables and the rhythms of anchoring points for tones in these two languages. Another way to characterize the difference is to say that Japanese and Korean emphasize edges at all levels of metrical structure, from the consonant-focused definition of the syllable to the primarily demarcative use of tonemes and tonal morphemes, whereas Beijing Mandarin, like English, emphasizes heads.

This difference has ramifications for the realization of focal prominence. In Korean and Japanese, focal prominence is realized primarily by an expansion of the tonal space to enhance the demarcative rise at the beginning of the first AP coupled with a post-focal erasure of AP boundaries (see, e.g., Venditti et al., 1996). Other prominence-enhancing mechanisms include the choice of IP-final boundary tones such as the H% tone of Tokyo Japanese (see, e.g., Venditti et al., 2008). In Beijing Mandarin, English, and Swedish, by contrast, focal prominence instead singles out a syllable with tonic stress and then either reduces or deletes the tones associated to following stressed syllables (see Jin, 1996; Xu & Wang, 2001, among many others, for Mandarin; ch. 6 of Ladd, 2008, for a review of the literature on English; and Bruce, 1977, 1982, among many others, for Swedish). Chapter 7 of Ladd (2008) gives a particularly insightful discussion of this difference between edge-focus and head-focus strategies. He also suggests a common underlying unity. The syllable with tonic stress in languages such as English and Swedish plays a culminative role in marking words and larger morphosyntactic constituents, as illustrated by the tone pattern that marks the compound word in (5). The word that is the domain of the focal H- in Stockholm Swedish, similarly, plays a culminative role in marking intonation phrases and their alignment with the domains of focus in the information structure of the sentence. Pierrehumbert (1980) posits a similar "phrase accent" for English, as in the transcription in (1). Gussenhoven (1984), by contrast, treats the rise-fall-rise over *messenger* and *ball* as a single H*LH tonal morpheme. By either analysis, however, the word that contains the stressed syllable to which the H* tone is associated plays a similarly culminative role vis-à-vis the focus domain in English. Ladd (2008, p. 278) suggests that both the culminative function and the demarcative function can be viewed as ways of identifying levels of grouping in the metrical hierarchy of a language:

> If . . . we see prosodic phrasing as the ultimate basis of sentence stress, we may see that the correct way to pose questions about universals of the prosody-focus link is not "Why is the main accent in this sentence on word X rather than on word Y?" but rather "Why is this sentence divided up into phrases the way it is?"

A further advantage of thinking of "sentence stress" in this way is that focus and other aspects of information structure at the level of the sentence then become the local expression of the same types of discourse structure relationships that are encoded in such effects as final lowering, as discussed in Nakatani (1997) and Venditti (2000).

5 Is a Typology Needed?

As noted earlier, Pike (1948) reserved the term "tonal language" for languages like Cantonese, but later researchers define tone language more broadly. For example, Hyman (2006, p. 229) gives a "working definition of tone" which turns out instead to be a definition of a language type: "A language with tone is one in which an indication of pitch enters into the lexical realisation of at least some morphemes." He rejects a distinction between "pitch accent system" and "tone system" (corresponding to Voorhoeve's distinction between "restricted" and "unrestricted" tone systems described above, and Pike's earlier distinction between a true "tonal language" and a language with a "word-pitch system"). He points out that languages cited as examples of the "pitch accent" type are a varied lot, including languages as different as Tokyo Japanese (where "accent" does not imply metrical prominence and the majority of native words are unaccented) and Stockholm Swedish (where every word has at least one syllable with tonic stress and a compound word has exactly two). Therefore "culminativity" does not seem such a useful metric for typological grouping. While he rejects the idea of this third type, however, he maintains that a tenable distinction can still be made between a "tone language" prototype and a "stress-accent language" prototype. His criteria for setting up this distinction require that he treat stress as a "suprasegmental" property on par with H tone, rather than as a structural property on par with syllabicity. That is, he proposes that the prototype stress-accent system is one in which "every word has at least one stress accent" and "the stress-bearing unit is necessarily the syllable."

Kubozono (2001, p. 264) agrees that the languages that are singled out as "pitch-accent" languages are not anything like a homogenous type, but he points out that there are as "many subtypes of tone and stress systems as the subtypes of pitch-accent systems." As our examples from three Chinese languages and as many Germanic languages in the previous three sections should make clear, we concur with this assessment. We suspect that the appearance of prototypes comes from looking too closely at just one or two of the functions in which tone participates, as well as from being thoroughly immersed in the consensus assumptions of specialists in just one or two Sprachbund regions. For example, Hyman's tone language prototype looks to us like a description of features which are frequently encountered together in prosodic systems in the Bantu and West African Sprachbunds, where tone patterns often function as tonemes, whereas his stress-accent language prototype looks to us like a description of features that are frequently encountered together in West Germanic languages such as English, where tone patterns typically function as tonal morphemes that are associated to syllables with tonic stress.

In this context, we are also struck by a comment that Gussenhoven (2004, pp. 46–7) makes about Hyman's definition of tone. He first points to languages where no word pairs are distinguished by having different tone patterns, but where tone nonetheless enters into the specification of words, either by consistently being associated at some position such as the first or last syllable in every word (so as to demarcate words in utterances) or by being associated with particular phrasal boundaries in a way that makes them morphological markers. He then suggests:

> . . . a case can be made for lexically specified tone in intonation-only languages. Clearly, pitch accents and boundary tones constitute an "intonational lexicon" from which speakers make semantically and pragmatically appropriate choices for every accented syllable and intonational phrase (Liberman 1975). Additionally, intonation-only languages may have tonal specifications in the "segmental" lexicon for particles which invariably appear with a particular intonation contour, like Dutch sentence-final [hɛ], which expresses an appeal for agreement, as in *Leuk, hè'?* "Nice, isn't it?"

It was this salient remark that prompted us to group the tonemic function and the morphemic function together in section 4.1.

Although we have singled out Hyman (2006) as representative of the broad-stroke typologies, we could have cited any of many other papers that assume that other prosodic differences naturally fall out from the difference between using tone "to make semantic distinctions" and using it "to add functional meaning." For example, Hyman's definition is virtually identical to that of Welmers (1973). The difference that is deemed critical in these broad-stroke typologies is a distinction between the tonemic function of lexical contrast and everything else – between languages such as Cantonese, where many of the tones in the melody of a typical utterance are tonemes that combine productively with the consonant and vowel phonemes of the language to make a large and expandable set of morphemes, and languages such as English, where the tones are pragmatic morphemes chosen from a small and relatively closed set. This is a useful distinction, because it predicts that there would be sharp differences in native speakers' and linguists' metalinguistic awareness of the tone count, as suggested in section 4.1. However, contra Hyman (2001, 2006) we do not see that it correlates neatly with all of the other distinctions that could be made on the basis of the functions outlined in sections 4.2 and 4.3. Appreciating the difference in ease of counting that falls out from the fact that a L+H that is anchored to a stressed syllable in Putonghua is a toneme whereas a L+H* that is anchored to a stressed syllable in English is a pragmatic morpheme does not preclude these two languages from being far more like each other in many other respects than either is to a language such as Japanese. Trying to reduce the multidimensional taxonomic space outlined in sections 3 and 4 to a continuum from a "tone language" prototype such as Putonghua (or any other variety of Chinese) to a "stress-accent language" prototype such as English (or any other variety of Germanic) makes as much sense as reducing the space to a one-dimensional continuum from a prototypical head-focus language (such as Beijing Mandarin or English) to a prototypical edge-focus language (such as Japanese). Or (since we are phoneticians and not syntacticians) it makes as much sense as reducing the space to a one-dimensional continuum from a prototypical vowel language (such as Shanghai or Swedish) to a prototypical consonant language (such as Korean).

In summary, we have managed to keep this introduction to tone and intonation within the space of a chapter by concentrating primarily on Chinese and Germanic (the languages mentioned specifically in the *COED* definition) along with Japanese (a language that is often cited as intermediate in type). We have tried to cover a small part of the large literatures on these three languages in a way that hints at the enormous variety of systems possible within even this small handful. However, we must emphasize that there are thousands of languages in the world, and we have in-depth descriptions of the tone and intonation systems of fewer than two dozen of them. Until we have a much more thorough taxonomy, along with more extensive comparative work within and between languages, any typology is bound to be premature.

NOTES

[1] At many points in this chapter, we will use the term *prosody* to refer to aspects of the organizational structure of utterances that are critical for understanding tone and intonation, but which have more to do with the rhythm of segments. We refer the reader to Fletcher (this volume) for a definition of this term and a review of the relevant work.

[2] Chao's tone letters locate notes and glissando turning points in the local tonal space in terms of five points numbered from 1 for the bottom to 5 for the top of the tone space.

[3] The traditional use of the term "accent" both for pragmatic tonal morphemes in English and for lexically specified tone patterns in Japanese is the source of frequent confusion among scholars of both languages. Further confusion is caused by the fact that the Japanese word *akusento*, which 'accent' translates, here refers to the entire configuration of tones for the level of prosodic grouping that is called the accentual phrase, including both the lexically specified pitch fall at the designated syllable and the "post-lexically" specified pitch rise at the beginning of the accentual phrase. See Venditti et al. (2008) for an explication of the differences between the two phenomena.

4 While we focus on the tonal aspects of the definition here, each of these levels of prosodic grouping is also marked by segmental effects such as differing degrees of "initial strengthening" and "phrase-final lengthening" (see Fletcher, this volume).

5 Following Zee and Maddieson (1980), we transcribe the apical vowel here and in the examples with [ɹ] rather than with the non-IPA symbol used in the Sinological literature.

REFERENCES

Arvaniti, A. (1994) Acoustic features of Greek rhythmic structure. *Journal of Phonetics*, 22, 239–68.

Arvaniti, A. (1999) Standard Modern Greek. *Journal of the International Phonetic Association*, 19, 167–72.

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998) Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics*, 26, 3–25.

Arvaniti, A., Ladd, D. R., & Mennen, I. (2000) What is a starred tone? Evidence from Greek. In M. Broe & J. B. Pierrehumbert (eds.), *Papers in Laboratory Phonology 5* (pp. 119–31). Cambridge: Cambridge University Press.

Atal, B. & Hanauer, S. (1971) Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 55, 1304–12.

Bazell, C. E., Catford, J. C., Halliday, M. A. K., & Robins, R. H. (eds.) (1966) *In Memory of J. R. Firth*. London: Longmans, Green, & Co.

Beckman, M. E., Erickson, D., Honda, K., Hirai, H., & Niimi, S. (1995) Physiological correlates of global and local pitch range variation in the production of high tones in English. *Proceedings of the 13th International Congress of Phonetics Sciences*, 2, 638–41.

Beckman, M. E., Hertz, S. R., & Fujimura, O. (1983) SRS pitch rules for Japanese. *Working Papers of the Cornell Phonetics Laboratory*, 1, 1–16.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005) The original ToBI system and the evolution of the ToBI framewordk. In S.-A. (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 9–54). Oxford: Oxford University Press.

Beckman, M. E. & Pierrehumbert, J. B. (1986) Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 225–309.

Beckman, M. E. & Pierrehumbert, J. B. (1992) (Strategies and tactics for thinking about F0 variation.) Comments on chapters 13 and 14. In G. J. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology 2* (pp. 387–97). Cambridge: Cambridge University Press.

Benedict, P. K. (1948) Tonal systems in Southeast Asia. *Journal of the American Oriental Society*, 64, 184–91.

Berg, R. van den, Gussenhoven, C., & Rietveld, T. (1992) Downstep in Dutch: Implications for a model. In G. J. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology 2* (pp. 335–58). Cambridge: Cambridge University Press.

Boersma, P. (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences, Amsterdam*, 17, 97–110.

Boersma, P. & Weenink, D. (2007) *Praat: Doing phonetics by computer (Version 4.6.05)* (computer program). Retrieved June 3, 2007, from www.praat.org.

Bolinger, D. (1958) A theory of pitch accent in English. *Word*, 14, 109–49.

Bolinger, D. (1978) Intonation across languages. In J. Greenberg (ed.), *Universals of Human Language*, vol. 2: *Phonology* (pp. 471–524). Stanford, CA: Stanford University Press.

Browman, C. P. & Goldstein, L. (1986) Towards an articulatory phonology. *Phonology Yearbook*, 3, 219–52.

Browman, C. P. & Goldstein, L. (1990) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (eds.), *Papers in Laboratory Phonology 1* (pp. 341–76). Cambridge: Cambridge University Press.

Bruce, G. (1977) *Swedish Word Accents in Sentence Perspective*. Lund: Gleerup.

Bruce, G. (1982) Developing the Swedish intonation model. *Working Papers, Lund University, Department of Linguistics*, 23, 51–116.

Bruce, G. (1983) Accentuation and timing in Swedish. *Folia Linguistica*, 17, 221–38.

Bruce, G. (1987) How floating is focal accent? In K. Gregersen & H. Basbøll (eds.), *Nordic Prosody IV* (pp. 41–9). Odense: Odense University Press.

Bruce, G. (1989) Report from the IPA working group on suprasegmental categories. *Working Papers, Lund University, Department of Linguistics*, 35, 15–40.

Bruce, G. (1990) Alignment and composition of tonal accents: Comments on Silverman and Pierrehumbert's paper. *Papers in Laboratory Phonology 1* (pp. 107–14). Cambridge: Cambridge University Press.

Byrd, D. (1996) A phase window framework for articulatory timing. *Phonology*, 13, 139–69.

Byrd, D. & Saltzman, E. (2003) The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31, 149–80.

Campbell, N. & Black, A. W. (1997) Prosody and the selection of source units for concatenative synthesis. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, & J. Hirschberg (eds.), *Progress in Speech Synthesis* (pp. 279–82). New York: Springer.

Cao, J. & Maddieson, I. (1992) An exploration of phonation types in Wu dialects of Chinese. *Journal of Phonetics*, 20, 77–92.

Carter, H. (1974) Negative structures in the syntactic tone-phrasing system of Kongo. *Bulletin of the School of Oriental and African Studies, University of London*, 37, 29–40.

Caspers, J. (2000) Experiments on the meaning of four types of single-accent intonation patterns in Dutch. *Language and Speech*, 43, 127–61.

Cedergren, H., Perreault, H., Poiré, F., & Rousseau, P. (1990) L'accentuation québécoise: Une approche tonale. *Revue québécoise de linguistique*, 19, 25–38.

Cedergren, H. & Simoneau, L. (1985) La chute des voyelles hautes en français de Montreal: "As-tu entendu la belle syncope?" In M. Lemieux & H. Cedergren (eds.), *Les tendentes dynamiques du français parlé à Montréal* (pp. 57–144). Québec: Bibliothèque nacionale du Québec.

Chan, M. K. M. (1998) Sentence particles *je* and *jek* in Cantonese and their distribution across gender and sentence types. In S. Wertheim, A. Bailey, & M. Corston-Oliver (eds.), *Engendering Communication: Proceedings of the Fifth Berkeley Women and Language Conference* (pp. 117–28). Berkely, CA: Berkeley Women and Language Group.

Chan, M. K. M. & Ren, H. (1989) Wuxi tone sandhi: From last to first syllable dominance. *Acta Linguistica Hafniensia*, 21, 35–64.

Chang, N. T. (1958) Tone and intonation in the Chengtu dialect (Szechuan, China). *Phonetica*, 2, 59–85.

Chao, Y. R. (1930) A system of tone-letters. *Le Maître Phonétique*, 30, 24–7.

Chao, Y. R. (1932) Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology, Academia Sinica*, 4, 121–34.

Chao, Y. R. (1968) *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, M. Y. (2000) *Tone Sandhi: Patterns across Chinese Dialects.* Cambridge: Cambridge University Press.

Chen, Y. (2008) Revisiting the phonetics and phonology of Shanghai tone sandhi. *Proceedings of the Fourth Conference on Speech Prosody, 6–8 May 2008, Campinas, Brazil*, 253–56.

Chen, Y. & Xu, Y. (2006) Production of weak elements in speech: Evidence from F0 patterns of neutral tone in Standard Chinese. *Phonetica*, 63, 47–75.

Clark, M. M. (1978) A dynamic treatment of tone with special attention to the tonal system of Igbo. Doctoral dissertation, University of New Hampshire.

Cohen, A. & Hart, J. 't (1967) On the anatomy of intonation. *Lingua*, 19, 177–92.

Coleman, J. (1992) The phonetic interpretation of headed phonological structures containing overlapping constituents. *Phonology*, 9, 1–44.

Coleman, J. (1998) Declarative syllabification in Tashlhit Berber. In J. Durand & B. Laks (eds.), *Current Trends in Phonology: Models and Methods*, vol. 1 (pp. 177–218). Salford, UK: European Studies Research Institute, University of Salford.

Concise Oxford English Dictionary (2004) 11th edn., ed. C. Soanes & A. Stevenson. Oxford: Oxford University Press.

Dauer, R. M. (1980) The reduction of unstressed high vowels in Modern Greek. *Journal of the International Phonetic Association*, 10, 17–27.

Dell, F. & Elmedlaoui, M. (1998) Nonsyllabic transitional vocoid in Imdlawn Tashlhiyt. In J. Durand & B. Laks (eds.), *Current Trends in Phonology: Models and Methods*, vol. 1 (pp. 217–44). Salford, UK: European Studies Research Institute, University of Salford.

Denes, P. & Milton-Williams, J. (1962) Further studies in intonation. *Language and Speech*, 5, 1–14.

D'Imperio, M. & Gili Fivela, B. (2003) How many levels of phrasing? Evidence from two varieties of Italian. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in Laboratory Phonology 6* (pp. 130–44). Cambridge: Cambridge University Press.

Epstein, M. A. (2002) Voice quality and prosody in English. Doctoral dissertation, University of California, Los Angeles.

Erickson, D. (1978) A physiological analysis of the tones of Thai. Doctoral dissertation, University of Connecticut.

Erickson, D. (1993) Laryngeal muscle activity in connection with Thai tones. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, University of Tokyo*, 27, 135–49.

Erickson, D., Honda, K., Hirai, H., & Beckman, M. E. (1995) The production of low tones in English intonation. *Journal of Phonetics*, 23, 179–88.

Fant, G. (1960) *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fry, D. B. (1968) Prosodic phenomena. In B. Malmerg (ed.), *Manual of Phonetics* (pp. 364–410). Amsterdam: North-Holland.

Fujisaki, H. (1983) Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage (ed.), *The Production of Speech* (pp. 39–55). New York: Springer.

Fujisaki, H. & Sudo, H. (1971) Synthesis by rule of prosodic features of connected Japanese. *Proceedings of the 7th International Congress on Acoustics (ICA)*, Budapest, 133–6.

Fung, R. S.-Y. (2000) Final particles in Standard Cantonese: Semantic extension and pragmatic inference. Doctoral dissertation, Ohio State University.

Gårding, E. (1977) The importance of turning points for the pitch patterns of Swedish accents. In L. M. Hyman (ed.), *Studies in Stress and Accent*, Southern California Occasional Papers in Linguistics 4 (pp. 27–35). Los Angeles, CA: University of Southern California.

Gårding, E. (1993) On parameters and principles in intonation analysis. *Working Papers, Lund University, Department of Linguistics*, 40, 25–47.

Gårding, E., Fujimura, O., & Hirose, H. (1970) Laryngeal control of Swedish word tones: A preliminary report on an EMG study. *Annual Report of the Research Institute on Logopedics and Phoniatrics*, 4, 45–54.

Gårding, E., Kratochvil, P., Svantesson, J.-O., & Zhang, J. (1986) Tone 3 and tone 4 identification in modern standard Chinese. *Language and Speech*, 29, 281–93.

Goldsmith, J. A. (1976/1979) Autosegmental phonology. Doctoral dissertation, MIT. [(Published 1979, New York: Garland Publishing.)

Goldsmith, J. A. (1984) Tone and accent in Tonga. In G. N. Clements & J. Goldsmith (eds.), *Autosegmental Studies in Bantu Tone* (pp. 19–51). Dordrecht: Foris.

Goldsmith, J. A. (1987) Tone and accent, and getting the two together. In J. Aske, N. Beery, L. Michaelis, & H. Filip (eds.), *Proceedings of the 13th Annual Meeting of the Berkeley Linguistics Society* (pp. 88–104). Berkeley, CA: Berkeley Linguistics Society.

Goldstein, L. (1989) On the domain of Quantal Theory. *Journal of Phonetics*, 17, 91–7.

Goldstein, L. & Fowler, C. A. (2003) Articulatory phonology: A phonology for public use. In N. O. Schiller & A. S. Meyer (eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 159–207). Berlin: Mouton de Gruyter.

Grabe, E. (2004) Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles & J. Peters (eds.), *Regional Variation in Intonation* (pp. 9–31). Tübingen: Niemeyer.

Grice, M., Ladd, D. R., & Arvaniti, A. (2000) On the place of phrase accents in intonational phonology. *Phonology*, 17, 143–85.

Grice, M. & Savino, M. (1995) Low tone versus "sag" in Bari Italian intonation: A perceptual experiment. In *Proceedings of the 13th International Congress of Phonetic Sciences*, 4, 658–61.

Grønnum, N. (1990) Prosodic parameters in a variety of regional Danish standard languages, with a view towards Swedish and German. *Phonetica*, 47, 182–214.

Grønnum, N. (1995) Review of Möbius (1993). *Zeitschrift für Dialektologie und Linguistik*, 62, 346–8.

Gussenhoven, C. (1984) *On the Grammar and Semantics of Sentence Accents*. Dordrecht: Foris.

Gussenhoven, C. (1999) Discreteness and gradience in intonational contrasts. *Language and Speech*, 42, 283–305.

Gussenhoven, C. (2004) *The phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C. & Bruce, G. (1999) Word prosody and intonation. In H. van der Hulst (ed.), *Word Prosodic Systems in the Languages of Europe* (pp. 233–71). Berlin: Mouton de Gruyter.

Gussenhoven, C. & Rietveld, T. (2000) The behavior of H* and L* under variations in pitch range in Dutch rising contours. *Language and Speech*, 43, 183–203.

Hallé, P. (1994) Evidence for tone-specific activity of the sternohyoid muscle in modern Standard Chinese. *Language and Speech*, 37, 103–23.

Halliday, M. A. K. (1967) *Intonation and Grammar in British English*. The Hague: Mouton.

Haraguchi, S. (1977) *The Tone Pattern of Japanese: An Autosegmental Theory of Tonology*. Tokyo: Kaitakusha.

Hart, J. 't & Collier, R. (1975) Integrating different levels of intonation analysis. *Journal of Phonetics*, 1, 309–27.

Hart, J. 't, Collier, R., & Cohen, A. (1990) *A Perceptual Study of Intonation: An*

Experimental Phonetic Approach to Speech Melody. Cambridge: Cambridge University Press.

Hattori, S. (1961) Prosodeme, syllable structure and laryngeal phonemes. *Studies in Descriptive and Applied Linguistics, Bulletin of the Summer Institute of Linguistics, International Christian University (Tokyo)*, 1, 1–27.

Henderson, E. J. A. (1949) Prosodies in Siamese. *Asia Major* (new series), 1, 189–215.

Herman, R. (2000) Phonetic markers of global discourse structure in English. *Journal of Phonetics*, 28, 466–93.

Herman, R., Beckman, M. E., & Honda, K. (1996) Subglottal pressure and final lowering in English. *Proceedings of the 1996 International Conference on Spoken Language Processing*, 145–48.

Hirai, T., Higuchi, N., & Sagisaka, Y. (1997) Comparison of F0 control rules derived from multiple speech databases. In Y. Sagisaka, N. Campbell, & N. Higuchi (eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 211–23). New York: Springer.

Hirschberg, J. & Ward, G. (1992) The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, 20, 241–51.

Hockett, C. F. (1960) The origin of speech. *Scientific American*, 203, 88–96.

Hyman, L. M. (2001) Privative tone in Bantu. In S. Kaji (ed.), *Cross-Linguistic Studies of Tonal Phenomena: Tonogenesis, Japanese Accentology, and Other Topics* (pp. 237–57). Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Hyman, L. M. (2006) Word-prosodic typology. *Phonology*, 23, 225–57.

Jin, S. (1996) An acoustic study of sentence stress in Mandarin Chinese. Doctoral dissertation, Ohio State University.

Jun, S.-A. (1993/1996) The phonetics and phonology of Korean prosody: Intonational phonology and prosodic structure. Doctoral dissertation, Ohio State University. (Published 1996, New York: Garland Publishing.)

Jun, S.-A. (ed.) (2005) *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, S.-A. & Beckman, M. E. (1994) Distribution of devoiced high vowels in Korean. *Proceedings of the 1994 International Conference on Spoken Language Processing*, 2, 479–82.

Jun, S.-A. & Fougeron, C. (2002) Realizations of accentual phrase in French intonation. *Probus*, 14, 147–72.

Kawakami, S. (1957) Jun akusento ni tsuite [On the tone of connected words in Japanese]. *Kokugo kenkyū*, 7, 44–60.

Keating, P., Cho, T., Fougeron C., & Hsu, C.-S. (2003) Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in laboratory phonology*, 6, 145–63. Cambridge, UK: Cambridge University Press.

Kennedy, G. A. (1953) Two tone patterns in Tangsic. *Language*, 29, 367–373.

Kindaichi, H. (1957) Nihongo akusento no hiken [English title: On the unit of Japanese accent]. *Kokugo kenkyū*, 7, 1–32.

Kori, S. (1987) The tonal behavior of Osaka Japanese: An interim report. *Ohio State University Working Papers in Linguistics*, 36, 31–61.

Kratochvil, P. (1998) Intonation in Beijing Chinese. In D. Hirst & A. Di Cristo (eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 417–30). Cambridge: Cambridge University Press.

Kubozono, H. (2001) Comments on "Privative tone in Bantu" by Larry Hyman. In S. Kaji (ed.), *Cross-Linguistic Studies of Tonal Phenomena: Tonogenesis, Japanese Accentology, and Other Topics* (pp. 259–65). Tokyo: Institute for the

Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Ladd, D. R. (1992) An introduction to intonational phonology. In G. Docherty and D. R. Ladd (eds.), *Papers in Laboratory Phonology 2* (pp. 321–34). Cambridge: Cambridge University Press.

Ladd, D. R. (1993) In defense of a metrical theory of intonational downstep. In H. van der Hulst & K. Snider (eds.), *The Phonology of Tone: The Representation of Tonal Register* (pp. 109–32). Dordrecht: Foris.

Ladd, D. R. (1996) *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R. (2008) *Intonational Phonology*, 2nd edn. Cambridge: Cambridge University Press.

Ladd, D. R., Mennen, I., & Schepman, A. (2000) Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America*, 107, 2685–96.

Ladd, D. R., & Schepman, A. (2003) “Sagging transitions” between high pitch accents in English: Experimental evidence. *Journal of Phonetics*, 31, 81–112.

Ladefoged, P. (1990) Some reflections on the IPA. *Journal of Phonetics*, 18, 335–46.

Law, S. (1990) The syntax and phonology of Cantonese sentence-final particles. Doctoral dissertation, Boston University.

Lee, L., Tseng, C., & Hsieh, C. (1993) Improved tone concatenation rules in a formant-based Chinese text-to-speech system. *IEEE Transactions on Speech and Audio Processing*, 1, 287–94.

Lee, O. J. (2000) The pragmatics and intonation of ma-particle questions in Mandarin. Masters thesis, Ohio State University.

Lee, O. J. (2005) The prosody of questions in Beijing Mandarin. Doctoral thesis, Ohio State University.

Lehiste, I. & Ivić, P. (1963) *Accent in Serbocroatian*, Michigan Slavic Materials 4. Ann Arbor, MI: University of Michigan, Department of Slavic Languages and Literatures.

Li, Z. (2003) The phonetics and phonology of tone mapping in a constraint-based approach. Doctoral dissertation, MIT.

Liberman, M. Y. (1975/1979) The intonational system of English. Doctoral dissertation, MIT. (Published 1979, New York: Garland Publishing.)

Liberman, M. Y. & Pierrehumbert, J. B. (1984) Intonational invariance under changes in pitch range and length. In M. Aranoff & R. T. Oehrle (eds.), *Language Sound Structure* (pp. 157–233). Cambridge, MA: MIT Press.

Liberman, M. Y. & Prince, A. (1977) On stress and linguistic rhythm, *Linguistic Inquiry*, 8, 249–336.

Lieberman, P. (1967) *Intonation, Perception, and Language*. Cambridge, MA: MIT Press.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling* (pp. 403–39). Dordrecht: Kluwer.

Lyberg, B. (1981) Some consequences of a model for segment duration based on F0 dependence. *Journal of Phonetics*, 9, 97–103.

Maekawa, K. (2003) Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, Tokyo, 7–12.

Maekawa, K. & Kikuchi, H. (2005) Corpus-based analysis of vowel devoicing in spontaneous Japanese: An interim report. In J. van de Weijer, K. Najo, & T. Nishihara (eds.), *Voicing in Japanese* (pp. 205–28). Berlin: Mouton de Gruyter.

Maekawa, K., Kikuchi, H., Igarashi, Y., & Venditti, J. J. (2002) X-JToBI: An extended J_ToBI for spontaneous speech. *Proceedings of the 7th International Conference on Spoken Language Processing, Boulder, Colorado*, 1545–8.

MacNeilage, P. F. & Davis, B. L. (2000) Deriving speech from nonspeech:

A view from ontogeny. *Phonetica*, 57, 284–96.

McCawley, J. D. (1978) What is a tone language? In V. A. Fromkin (ed.), *Tone: A Linguistic Survey* (pp. 113–31). New York: Academic Press.

McCune, L. & Vihman, M. M. (1987) Vocal motor schemes. *Papers and Reports on Child Language Development*, 26, 72–9.

Miura, I. & Hara, N. (1995) Production and perception of rhetorical questions in Osaka Japanese. *Journal of Phonetics*, 23, 291–303.

Möbius, B. (1993) *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläfen*. Tübingen: Niemeyer.

Möbius, B., Pätzold, M., & Hess, W. (1993) Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication*, 13, 53–61.

Moulines, E. & Charpentier, F. (1989) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 453–67.

Myers, S. (1998) Surface underspecification of tone in Chichewa. *Phonology*, 15, 367–91.

Nakatani, C. H. (1997) The computational processing of intonational prominence: A functional prosody perspective. Doctoral dissertation, Harvard University.

Nespor, M. & Vogel, I. (1986) *Prosodic Phonology*. Dordrecht: Foris.

Nooteboom, S. (1997) The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Sciences*, 1st edn. (pp. 640–73). Oxford: Blackwell.

Ohala, J. J. (1970) *Working Papers in Phonetics, 15: Aspects of the Control and Production of Speech*. Department of Linguistics, University of California, Los Angeles. (http://repositories.cdlib.org/uclaling/wpp/No15)

Ohala, J. J. (1983) Cross-language use of pitch: An ethological view. *Phonetica*, 40, 1–18.

Ohala, J. J. (1992) The segment: Primitive or derived? In G. J. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology 2* (pp. 166–89). Cambridge: Cambridge University Press.

Öhman, S. E. G. (1967) Word and sentence intonation: A quantitative model. *Royal Institute of Technology, Speech Transmission Laboratory Quarterly Progress and Status Report*, 2–3, 20–54.

Papoušek, H., Papoušek, M., & Symmes, D. (1991) The meanings of melodies in motherese in tone and stress languages. *Infant Behavior and Development*, 14, 415–40.

Pierrehumbert, J. B. (1980) The phonology and phonetics of English intonation. Doctoral dissertation, MIT.

Pierrehumbert, J. B. (1990) Phonological and phonetic representation. *Journal of Phonetics*, 18, 375–94.

Pierrehumbert, J. B. (2000) The phonetic grounding of phonology. *Bulletin de la Communication Parlée*, 5, 7–23.

Pierrehumbert, J. B. & Beckman, M. E. (1988) *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J. B. & Hirschberg, J. (1990) The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (eds.), *Intentions in Communication* (pp. 271–311). Cambridge, MA: MIT Press.

Pijper, J. R. de (1983) *Modelling British English intonation.* Dordrecht: Foris.

Pike, K. L. (1948) *Tone Languages: A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion*. Ann Arbor: University of Michigan Press.

Reinholt Peterson, N. (1986) Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations. *Phonetica*, 43, 31–42.

Riha, H. (2008) Lettered words and roman letter characters: A study of alphabetic writing in Chinese newswires. Doctoral dissertation, Ohio State University.

Sagart, L., Hallé, P., Boysson-Bardies, B. de, & Arabia-Guidet, C. (1986) Tone production in modern Standard Chinese: An electromyographic investigation. *Cahiers de linguistique Asie Orientale*, 15, 205–21.

Selkirk, E. O. (1981) On the nature of phonological organization. In T. Myers, J. Laver, & J. Anderson (eds.), *The Cognitive Representation of Speech* (pp. 379–88). Amsterdam: North-Holland.

Selkirk, E. O. & Shen, T. (1990) Prosodic domains in Shanghai Chinese. In S. Inkelas & D. Zec (eds.), *The Phonology–Syntax Connection* (pp. 313–37). Chicago: University of Chicago Press.

Shih, C. & Kochanski, G. (2000) Chinese tone modeling with Stem-ML. *Proceedings of the Sixth International Conference on Spoken Language Processing*, 2, pp. 67–70.

Silverman, K. (1986) F0 cues depend on intonation: The case of the rise after voiced stops. *Phonetica*, 43, 76–92.

Sjölander, K. & Beskow, J. (2000) WaveSurfer: An open-source speech tool. *Proceedings of the Sixth International Conference on Spoken Language Processing*, 4, pp. 464–7.

Smiljanić, R. (2006) Early vs. late focus: Pitch-peak alignment in two dialects of Serbian and Croatian. In L. Goldstein, D. H. Whalen, & C. T. Best (eds.), *Laboratory Phonology 8*, 394–518.

So, L. K. H. (1996) Tonal changes in Hong Kong Cantonese. *Current Issues in Language and Society*, 3, 186–9.

Sugito, M. (2003) Timing relationships between prosodic and segmental control in Osaka Japanese word accent. *Phonetica*, 60, 1–16.

Sybsema, R. & Li, B. (2007) The dissection and structural mapping of Cantonese sentence final particles. *Lingua*, 117, 1739–83.

Thorsen, N. G. (1983) Two issues in the prosody of Standard Danish: The lack of sentence accent and the representation of sentence intonation. In A. Cutler & D. R. Ladd (eds.), *Prosody: Models and Measurements* (pp. 27–38). Berlin: Springer.

Thorsen, N. G. (1984) Intonation and text in Standard Danish, with special reference to the abstract representation of intonation. In W. U. Dressler, H. C. Luschützky, O. E. Pfeiffer, & J. R. Rennison (eds.), *Phonologica 1984: Proceedings of the Fifth International Phonology Meeting, Eisenstadt* (pp. 301–9). Cambridge: Cambridge University Press.

Thorsen, N. G. (1985) Intonation and text in Standard Danish. *Journal of the Acoustical Society of America*, 77, 1205–16.

Thorsen, N. G. (1986) Sentence intonation in textual context: Supplementary data. *Journal of the Acoustical Society of America*, 80, 1041–7.

Vanderslice, R. & Ladefoged, P. (1972) Binary suprasegmental features and transformational word-accentuation rules. *Language*, 48, 819–38.

Vayra, M. (1994) Phonetic explanations in phonology: Laryngealization as the case for glottal stops in Italian word-final stressed syllables. In W. U. Dressler, M. Prinzhorn, & J. R. Rennison, *Phonologica 1992: Proceedings of the 7th International Phonology Meeting* (pp. 275–93). Torino: Rosenberg & Sellier.

Venditti, J. J. (2000) Discourse structure and attentional salience effects on Japanese intonation. Doctoral dissertation, Ohio State University.

Venditti, J. J., Jun, S.-A., & Beckman, M. E. (1996) Prosodic cues to syntactic and other linguistic structures in Japanese, Korean, and English. In J. Morgan & K. Demuth (eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition* (pp. 287–311). Mahwah, NJ: Lawrence Erlbaum.

Venditti, J. J., Maeda, K., & van Santen, J. P. H. (1998) Modeling Japanese boundary pitch movements for speech synthesis. In M. Edgington (ed.), *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 317–22.

Venditti, J. J., Maekawa, K., & Beckman, M. E. (2008) Prominence marking in the Japanese intonation system. In S. Miyagawa & M. Saito (eds.), *Handbook of Japanese Linguistics* (pp. 458–514). Oxford: Oxford University Press.

Voorhoeve, J. (1973) Safwa as a restricted tone system. *Studies in African Linguistics*, 4, 1–21.

Ward, I. C. (1948) Verbal tone patterns in West African languages. *Bulletin of the School of Oriental and African Studies, University of London*, 12, 831–7.

Welmers, W. (1973) *African Language Structures*. Berkeley, CA: University of California Press.

Wong, W.-Y. P. (2006) Syllable fusion in Hong Kong Cantonese connected speech. Doctoral dissertation, Ohio State University.

Wong, W.-Y. P., Chan, M. K.-M., & Beckman, M. E. (2005) An Autosegmental-Metrical analysis and prosodic annotation conventions for Cantonese. In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 271–300). Oxford: Oxford University Press.

Xu, Y. & Liu, F. (2007) Determining the temporal interval of segments with the help of F_0 contours. *Journal of Phonetics*, 35, 398–420.

Xu, Y. & Wang, E. (2001) Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33, 319–37.

Yau, S. (1980) Sentential connotations in Cantonese. *Fangyan*, 80, 35–52.

Yip, M. J. W. (1980/1990) The tonal phonology of Chinese. Doctoral dissertation, MIT. [(Published 1990, New York: Garland Publishing.])

Zee, E. (1990) Vowel devoicing in Shanghai. *Chinese University of Hong Kong Papers in Linguistics*, 2, 69–104.

Zee, E. & Maddieson, I. (1980) Tones and tone sandhi in Shanghai: Phonetic evidence and phonological analysis. *Glossa*, 14, 45–88.

17 The Relation between Phonetics and Phonology

JOHN J. OHALA

1 Introduction

The question "What is the relation between phonetics and phonology?", like any question with historical and philosophical implications, cannot be answered objectively. Whenever an answer to this question is given it is inevitably subjective, grounded in the experience, beliefs, and prejudices of the answerer. This is no less true of my discussion here than of the views I survey from the past and the present. Nevertheless, I will try to make explicit the basis of my opinions so that those who have other views will be in a better position to evaluate my position and, perhaps, be persuaded by it.

To begin we must establish the defining characteristics of phonetics and phonology. Among the defects that I perceive in many prior attempts to define phonetics or phonology is specifying a field in a way which owes too much to modern theories and methods and which therefore implies that the field didn't even exist N years ago. To correct this defect it is necessary to take a broad historical view and to attempt to establish the recurring themes of the fields and aims of their practitioners. I believe that the defining characteristics of a discipline are not its methods and not its theories – the answers to questions – but rather the questions themselves. The methods used to get answers to questions and the candidate answers to the questions show periodic change; the questions themselves are remarkably long-lived and stem from the ordinary experience and puzzlings people have about spoken language. Ancient myths, oral and written, from many diverse cultures show, by the candidate answers given, that the following are some of the persistent questions people have had about spoken language:

Q1 How did speech originate? Why do humans use speech but not other species? What is the relation between human speech and animals' cries? Is vocal communication possible across species?

Q2 Why is there such diversity in the form of speech, i.e., between different linguistic communities and between generations of the same linguistic community?

Q3 What is the physical structure of speech? How is it made? How is it perceived?
Q4 How is speech represented in the brain?
Q5 How does one learn to pronounce one's first language, a second language?
Q6 How can we communicate effectively with speech under adverse conditions, over great distances, with high background noise? How can we effectively "freeze" ephemeral speech so that we can store spoken messages for later recall?
Q7 How can we ameliorate speech communication deficits (e.g., cleft palate, stuttering, lisping, deafness)?
Q8 How does meaning come to be associated with the sounds of language?

Although I would maintain that these are among the perennial questions about spoken language, it may happen that at any given time in history or in specific communities one or more may be the focus of inquiry with others neglected. For example, in the eighteenth century there was more focus on Q1 and Q2; in the twentieth century, greater effort was spent on Q4. Only some of these questions have received widely accepted answers: Q2 was partially answered by the development of the comparative method in the nineteenth century; the accumulation of efforts devoted to Q3 over three centuries and more have given us an understanding of the physical structure of speech sufficient to make machines which speak and understand speech. Similarly, the invention of writing, the telephone, the hearing aid, surgical repair of cleft palate, etc. have provided some answers to Q6 and Q7. Q1 has yet to be satisfactorily answered and, indeed, the curiosity we have about all these aspects of speech will never be completely satisfied. It is the never-ending search for answers to these questions that unites everyone in the field, from von Kempelen to von Humboldt, from Grimm to Greenberg, from Helmholtz to Halle. Methods and theories come and go; the questions remain.

2 Some History

With this base it is now appropriate to survey opinions on the relation of phonetics and phonology. Sommerstein (1977, p. 1) states:

> Phonology is a branch of linguistics; phonetics is often considered not to be. Phonetics deals with the capabilities of the human articulatory and auditory systems with respect to the sounds and prosodic features available for use in language, and with the acoustic characteristics of these sounds and features themselves . . .
>
> Phonology, in a sense, begins where phonetics leaves off. It is concerned with the ways in which the sounds and prosodic features defined by phonetics are actually used in natural languages . . .
>
> . . . there have been two main views on [the object of phonological inquiry] . . . "What phonic features (a) serve in the language under investigation, or (b) are capable

> of serving in natural language, to distinguish one utterance from another?" [this is classical phonology] . . . [The other asks] "What are the principles determining the pronunciation of the words, phrases and sentences of a language; and to what extent are these principles derivable from more general principles determining the organization . . . of all human languages?" [= generative phonology].

From these and similar quotes from other texts one could characterize phonetics as concerned with discovering and describing the vocal sounds utilizable by humans, and studying articulation, acoustics, and perception; in other words approximately the domain of inquiry of Q3 above. Phonology is said to be concerned with how the sounds used in language pattern or function, how they are represented and used in the mental grammar of speakers; approximately Q1 and Q4 and perhaps 5. Phonetics deals with concrete, physical manifestations of speech sounds; phonology with abstract, psychological manifestations – indeed, more generally, with the nature of human language and the genetic endowment which makes it possible. Phonetics is characterized as using the methods of the natural sciences; phonology as using the methods of the social sciences or perhaps of the humanities. (See also Kenstowicz & Kisseberth, 1979, p. 1; Hawkins, 1984, p. 7; Lass, 1984, p. 1; Clark & Yallop, 1990, pp. 1–2.)

Curiously, in few of the texts I surveyed (admittedly not a thorough search) is there any mention of sound change, i.e., Q2, when the definition of phonology and its differentiation with phonetics is specified. Perhaps phonological change is assumed to be covered by the psychological focus if one assumes that "sound change is grammar change" (Kiparsky, 1968). In any event, it is widely recognized that there is an intimate relation – some would say an identity – between the sound patterns which figure in synchronic, supposedly mental, grammars and the sound changes that occur in successive generations (Halle, 1962; Kiparsky, 1968; King, 1969). It is safe to assume that an understanding of diachronic variation is essential to an understanding of the many sound patterns which occupy mainstream phonologists today: vowel harmony, spirantization, epenthesis, deletion, diphthongization, etc.

I would now like to contrast the contemporary view of the relation between phonetics and phonology with earlier attitudes on the matter, i.e., up to three centuries ago. Given that both "phonetics" and "phonology" did not necessarily exist as separately recognized disciplines in earlier centuries, I will instead look at the relation between the domains of study that are classified today as phonetic or phonological.

2.1 Amman to Rousselot

Johann Conrad Amman (1669–1724), like his contemporaries John Wallis and William Holder, delved into the study of the physical nature of speech because of his interest in teaching the deaf to speak (Amman, 1694). But in pursuing this he made many original observations and analyses that would today be considered "phonological": He proposed an elementary, binary, hierarchical system of phonetic features that still merits attention. In his system manner features dominated place features (see Miller & Nicely, 1955, who found that manner distinctions are more resistant to confusion than place distinctions). He considered his system as a "natural" hierarchy, i.e., in accord with nature, among other reasons, because substitutions of sounds in, for example, pathological speech, involve similar sounds at the lowest level of the hierarchy, not the highest; thus an alveolar "semi-vowel" like *l* is substituted for another, *r*, or one nasal for another whereas vowels and consonants which are differentiated at the highest level, are rarely if ever substituted one for the other. Amman (like Wallis before him) also made some elementary phonological observations, e.g., "If any word terminates in **n** and the following word begins with **b** or **p** . . . then in pronouncing the **n** we unconsciously change it, for the sake of euphony, into **m** . . ."

Charles de Brosses (1709–77) wanted to do for language what Descartes and Newton had done for the physical universe: derive it from first principles. The principal thesis of his work *Traité de la formation méchanique des langues, et de principes physiques de l'étymologie* (1765) was that the phonetic properties of words originally shared certain features of the things they designated (thus, he pointed out, words for *lip* often contain labial sounds, words for *nose* often contain nasal sounds, etc.). (See also Court de Gebelin, 1776.) Thus he was concerned with Q1 and Q8, above. In pursuing this argument he found it useful to do a completely original anatomical–physiological analysis of speech sounds. I say "original" because it was so far below what was known at his time that it is evident he must have done it without consulting contemporary sources. But it is not the sophistication of his analysis or the lack of it but rather the integration of what we now consider phonetics with phonology that is noteworthy. He invented two phonetic notations which enabled him to show the phonetic similarity of cognate words in diverse languages in spite of their having quite different spelling.

Erasmus Darwin (1731–1802), grandfather of Charles Darwin, in his work *The Temple of Nature* (1803) attempted to explain the origin of human society as well as language and speech. This work includes brief reports of his efforts at speech synthesis and what may be the first recorded instance of an instrumental phonetic study on a live, intact, speaker: To determine the place of articulation of vowels, he writes, "I rolled up some tin foil into cylinders about the size of my finger; and speaking the vowels separately through them [that is, inserting the cylinders into his mouth], found by the impressions made on them [i.e., where they were dented], in what part of the mouth each of the vowels was formed . . ." (p. 119). He also proposes 13 unary features for differentiating all human sounds (including the Welsh [ɬ]): three basic places of articulation, oral resonance,[1] nasal resonance, voiceless frication, voiced frication, etc.

Robert Willis (1800–75), a Cambridge professor of mechanics ("engineering," to use the modern equivalent), in his 1830 work "On the vowel sounds," specified quantitatively a single characteristic vocal tract resonance for each vowel and claimed that their principal articulatory determinant was vocal tract length. He remarked that with some refinement of his study he should be able to provide "philologists with a correct measure for the shades of differences in the pronunciation of the vowels by different nations." In other words, he envisioned a universal, quantitative, acoustically based specification of vowel quality (a goal that unfortunately still eludes us). His "single resonance" theory of vowels, though superseded by Grassmann's (1854) and Helmholtz' (1863) subsequent work (the basis for the modern acoustic theory of vowels), bears a resemblance to modern auditorily based theories of vowel quality, e.g., Fant and Risberg's F2-prime (Fant & Risberg, 1963; and Hermansky's PLP (perceptual linear predictive) transform, 1990).

T. Hewitt Key (1799–1875), at first professor of Latin and later professor of comparative philology at London University (now University College) and at one time professor at the newly formed University of Virginia, attempted to apply Willis' theories to problems of sound change. In his paper "On vowel-assimilation, especially in relation to Professor Willis' experiment on vowel-sounds" (1855), he proposed explanations for vowel harmony and umlaut by invoking Willis' notion that vocal tract length is the main articulatory determinant of their quality. Although his explanations would not be judged worthy by modern standards, his account is an admirable attempt to integrate what he knew about acoustic phonetics and the traditional problems of historical phonology. His article also contains some memorable and still pertinent admonitions:

> [Some scholars of language] have allowed themselves . . . to be led astray by paying more attention to the symbols of sound than to sounds themselves . . . Scholars seldom unite the love of classical and scientific pursuits; and a paper [i.e., Willis'] of the highest value for philology might well fail to meet with all the attention it deserves from the students of language . . .

Rudolf von Raumer (1815–76) in his paper "Die sprachgeschichtliche Umwandlung und die naturgeschichtliche Bestimmung der Laute" (1856 published in von Raumer 1863 and trans. W. P. Lehmann, 1967) strongly advocated the integration of the latest phonetic research with historical phonological studies. He wrote:

> Through the discoveries of historical linguistic investigation, the significance of phonetics has been placed in a new light. The more the importance of phonetics becomes recognized, the more apparent becomes the need to understand as clearly and precisely as possible its subject matter, namely the sounds themselves.

He attempted to give a more physiologically realistic interpretation to the first and second Germanic sound shifts in the light of philological and phonetic principles (based largely on Brücke, 1856).

Karl Verner (1846–96), one of the giants of nineteenth-century linguistics whose paper "Eine Ausnahme der ersten Lautverschiebung" (1875) was a prime inspiration for the Neo-grammarian revolution, in his later years plunged into phonetic studies of accent in order to understand better how it could influence sound change. He was one of the first in Denmark to obtain an Edison phonograph. He then constructed an elaborate and quite sophisticated optical instrument which permitted him to enlarge the tiny grooves it traced on the metal foil when recording speech such that he could measure their waveforms and analyze them mathematically. Unfortunately all this effort did not produce significant results and was never published, except posthumously in his Collected Letters (Verner, 1913; see also Jespersen, 1933, and Fischer-Jørgensen, 1979).

Abbé Pierre-Jean Rousselot (1846–1924), often called the father of experimental phonetics, introduced into phonetics the physiological methods of E. J. Marey, physician, pioneer in the study of locomotion, and the one who perfected the kymograph (with his invention, "Marey's capsule"). Rousselot attempted to do for phonetics what Helmholtz had done for hearing and vision: reduce their function to known physiological principles. Indicative of his view of the broad integrative character of the phonetic sciences are two of his major works: first, his dissertation (1891) which was an attempt to give an instrumental phonetic account of the sound changes which shaped the dialect spoken in his home town;[2] and, second, the application of phonetics to the communication problems of the deaf (1903).

Many other examples of a similar sort could be given (see also Rapp, 1836; Bindseil, 1838; Jacobi, 1843; Weymouth, 1856; Techmer, 1880; Ellis, 1877; Sievers, 1881; Passy, 1890; Grandgent, 1896; Sweet, 1877; Panconcelli-Calzia, 1904; Grammont, 1933). These all testify to the fact that there was no bar to the integration of what we would now label phonetics and phonology – a quite different attitude to what exists in the latter part of the twentieth century.

2.2 Strains between phonetics and phonology

To be sure, there were some signs of tension between phonetics and phonology and Key's remarks quoted above reflect that. In addition there are the following remarks of Roudet (1910; a student of Rousselot's) and Spargo (1931; the translator of Holger Pedersen's history of nineteenth-century linguistics, 1924), which by their intensity suggest that there were already some bitter feelings between the two fields.

> . . . la phonétique fournit a l'étude théorique des langues anciennes et modernes une base indispensable, faute de laquelle une foule de faits linguistiques demeurent inintelligibles, faute de laquelle toute une part de la grammaire historique se réduit à un pur psittacisme, à une collection de formules verbales à peu près dénuées de signification réele. (Roudet, 1910, p. v)
>
> . . . one important feature of [Pedersen's work] . . . is the striking rôle assigned to the study of phonetics in increasing our knowledge of linguistics. It is shown clearly that every important advance during the last century and a quarter was made by a scholar who attacked his problem from the phonetic side. Surely this fact has its importance for the future of linguistic study, and suggests that the indifference to phonetics in many of the graduate schools in the United States is an evil presage for future progress. (Spargo, 1931; in the preface, dated 1930, to his translation of Holger Pedersen's *Sprogvidenskaben i det nittende aarhundrede*, 1924)

But it would seem that the strained relations arose out of a difference in temperament or background: the bulk of those doing historical phonology had little or no training in or taste for phonetic work.[3] Rousselot, in the introduction to his *Principes* (1897–1908) offers a somewhat kinder view (p. 1):

> . . . les procédes des sciences expérimentales sont assez étrangers aux linguistes. Une sorte de terreur superstitieuse s'emparé eux dès qu'il s'agit de toucher au mécanisme le plus simple. II fallait donc leur montrer que la difficulté est moindre qu'ils ne se la figurent et leur faire entrevoir le champ immense que l'experimentation ouvre devant eux.

The picture that emerges from these brief vignettes from the history of phonetics and phonology up to the early decades of the twentieth century is that there was no hardening of the division between phonology and phonetics (or speech technology and speech pathology). Those who studied speech pursued their research in whatever way was comfortable for them, depending on their training: medical, mathematical, physical, or philological, but with many unhesitating excursions into new methodological territory.

2.3 *The split between phonetics and phonology*

What precipitated the apparent split between phonetics and phonology later on in the twentieth century? It is generally recognized that the division occurred due to the rise of structuralism, taught initially by Ferdinand de Saussure (1857–1913) and Jan Baudouin de Courtenay (1845–1929) but fully developed in phonology by the Prague School. In his 1939 work *Grundzüge der Phonologie* (trans. 1969 by C. Baltaxe), N. S. Trubetzkoy (1890–1938), a leader of the Prague School, distinguished

> . . . the study of sound pertaining to the act of speech [phonetics] which is concerned with concrete physical phenomena, [and] would have to use the methods of the natural sciences, while the study of sound pertaining to the system of language [phonology] would use only the methods of linguistics, or the humanities, or the social sciences. (p. 4)

In this way phonetics was placed outside of linguistics proper and phonology was conceived of as an autonomous discipline. The emphasis on system or the relationship between speech sounds and their function, rather than on the substance of those sounds, represented a new concern and one which seemed at the same time to open up new frontiers for phonological study and to liberate the study of speech sounds from physical phonetics and all the burdens of the methodology of the natural sciences.

Without a doubt phonology has a rich inheritance from Trubetzkoy and the school which he helped to develop. In fact, some of Trubetzkoy's phonological generalizations were based on intuitive phonetic grounds (though *he* felt he had to apologize and explain at some length how this didn't imply that he thought precise phonetic correlates of sound contrasts mattered). But Trubetzkoy's conception of phonetics was something of a stereotype:

> La phonétique actuelle se propose d'étudier les facteurs materiels des sons de la parole humaine: soit les vibrations de l'air qui leur correspondent, soit les positions et les mouvements des organes qui les produisent . . . Le phonéticien est necessairement atomiste ou individualiste . . . Chaque son de la parole humaine ne peut être étudié qu'isolement, hors de tout rapport avec les autres sons de la même langue. (1933, pp. 232–3)

A similar stereotype applied to astronomy would characterize its proper activity as merely looking at and cataloguing stars. No mention would be made of cosmology, astrophysical theory, etc., i.e., attempts to generalize about the birth, development, and death of stars, the formation of galaxies, the origin of the universe. This is the fallacy of equating the immediate, visible object of study with the ultimate object of study. Though the immediate object of study in phonetics (and in the psychological study of speech) may be the sounds of speech observed at various stages in the speech chain, the ultimate objects of study are the underlying *causes* of speech sound behavior, where "behavior" includes the same broad domain that Johann Amman studied three centuries ago, how laterals are produced, the assimilation of nasals to the place of articulation of following stops, the patterns of substitution of one speech sound for another, the organization of speech sounds.

A few qualifications must be added to the above historical interpretation of the origin of the split between phonetics and phonology. First, Trubetzkoy was not alone in his attitude. As the earlier quote from Spargo reveals, North American linguists had formed much the same opinion independently. Sapir's emphasis on the psychological aspect of speech sounds (Sapir, 1925) also led to a depreciation of phonetics within linguistics:

> Mechanical and other detached methods of studying the phonetic elements of speech are, of course, of considerable value, but they have sometimes the undesirable effect of obscuring the essential facts of speech-sound psychology.

Second, there was also some opposition to Trubetzkoy's divorce of phonetics from phonology. Gyula Laziczius (1948 [1966]), a member of the Prague School, insisted on the essentially linguistic concerns of phonetics. Eberhard Zwirner (Zwirner & Zwirner, 1936) emphasized that a proper experimental phonetic study of speech sounds must take into account the sounds' linguistic function. In fact, although Zwirner's views were not very influential in the development of the field, modern linguistic phonetics independently developed the same operating principles (see also Fischer-Jørgensen, 1985). Furthermore, the British school of linguistics did not separate phonology and phonetics and, indeed, were much later than many other schools in adopting two separate names for the joint activity.

Third, it must be acknowledged that some of the new interests of phonology were more in the psychological domain (this was especially true of Baudouin de

Courtenay's and Sapir's conception of phonology) and were not the typical focus of phonetic studies at that time. However, this situation has changed considerably today where there is substantial overlap between phonological, phonetic, and psychological studies of speech (e.g., Fowler, 1981; Cutler & Norris, 1988; Lahiri & Marslen-Wilson, 1991; Ohala & Ohala, 1995). Even so, although expressing an interest in the psychological aspect of speech, phonologists since Trubetzkoy's and Sapir's time have shown little initiative in adopting or developing rigorous psychological methods of studying sound patterns in language. The consequences of this neglect could be profound: many of the sound patterns in language claimed to be part of the native speaker's psychological endowment may simply be the residue of past sound changes which themselves came about primarily due to phonetic factors (Ohala, 1992a). Other aspects of speakers' awareness of sound patterns in their language may stem from their knowledge of their language's orthography (Jaeger, 1984; Read et al., 1986; Wang & Derwing, 1986; Derwing, 1992; Morais & Kolinsky, 1994) which, being a cultural artefact, can hardly count as knowledge required of a competent native speaker.

It could also be claimed that there is a sense in which all phonological work does in fact incorporate some phonetics insofar as it uses terms such as "obstruent," "voice," etc. However, I would like to differentiate between two forms of phonetics (see also Ohala, 1990a): one I call "taxonomic" phonetics and the other "scientific" phonetics. Taxonomic phonetics has provided us with traditional phonetic terms and symbols used to describe and classify speech sounds and has remained largely unchanged since the formation of the International Phonetic Association a century ago. Scientific phonetics, on the other hand, continues to change. It constantly expands its horizons; it develops new data, concepts, and methods; it rejects or revises earlier beliefs shown to be deficient, and, to the extent that the surviving beliefs or theories have congruence with the universe, it has practical payoff, e.g., in language teaching, speech pathology, and speech technology. Of course, it also has payoff in phonology: e.g., how would we be able to make sense of the inherent tendency of obstruents to become or remain voiceless if Husson's (1950) neurochronaxic theory of vocal cord vibration had not been effectively refuted? While autonomous phonology embraces taxonomic phonetics, for the most part it excludes scientific phonetics. A good bit of what is called and taught as "phonetics" in many universities – if it is taught at all – is exclusively taxonomic phonetics. Scientific phonetics is the intellectually most exciting form of the field – and one of the most successful and rigorous within linguistics (if one allows, of course, that it is part of linguistics). It addresses issues of fundamental importance for phonology. (See below.)

And it was not just the domain of inquiry that phonology left behind after its divorce from phonetics; it also abandoned phonetics' manner of bringing evidence to bear on theoretical claims. Over the decades the phonetic sciences had established a respectable degree of accountability in the way that generalizations and theories are proposed and defended. The degree of accountability in the field has been improved and tightened. As a result there is a relatively continuous and cumulative tradition on which to develop and refine both methods and theories.

To give just one example, and one which has far-reaching implications for phonology and for the behavioral sciences in general: careful phonetic studies spanning a century have demonstrated, the tremendous amount of variation – essentially infinite in character – that exists in the speech signal (Ohala, 1989). This synchronic variation parallels to a great extent documented diachronic variation which, in turn, gives rise to sound patterns studied by modern phonology: morphophonemic variations, phonotactic patterns, universal and language-particular patterns in languages' segment inventories, and allophonic variation. In addition, some patterns of variation parallel the phonological variation in language acquisition (first and second), as well as listeners' misperceptions over the telephone. Understanding variation in one of these domains has the potential to explain it in the other domains.

3 Philosophy

Anderson (1981) presents a useful scheme for discussing the relation of phonology with other disciplinary domains, e.g., phonetics, psychology, ethology, social and cultural factors, etc. Given Figure 17.1, where the thick-line circle represents the domain of "Language" (where, presumably, Phonology belongs) and the thinner-line circles intersecting it represent other disciplinary domains to which one may refer to explain specific aspects of language, the question may be stated: Is there any area within Language which remains outside the intersection of these circles? Are there any phenomena that are uniquely linguistic and which "cannot be explained as special cases of other systems"? Anderson endorses the Chomskyan position that language is a uniquely human

> mental organ . . . which is not reducible to features of other kinds (at least, within the limits of present knowledge in such areas as neurology, brain chemistry, the genetic control of development, etc.). It is exactly this area . . . that ought to occupy the central concern of linguists if they wish to arrive at an adequate conception of the essential and special nature of human Language.

There is, of course, no question that there will always be some things about language (or any other domain) that we will at the time be unable to explain by reference to physics, psychology, etc. These things should not be ignored but should be described. But to enshrine the things we are ignorant about as the central concern of linguistics is to misplace one's priorities. If one believes that there are irreducible phenomena in language then there will be no motivation to seek explanations for them. Indeed, left on its own, autonomous phonology endlessly recycles much the same data, trying out different labels and descriptive devices on it (markedness, abstract underlying forms, ordered rules, alpha-variables, atomic rules, upside-down rules, charm, optimality, a staggering variety of conditions and principles), all of which are attributed quite facilely to the new theoretical *deus ex machina*, "universal grammar." But it does this without

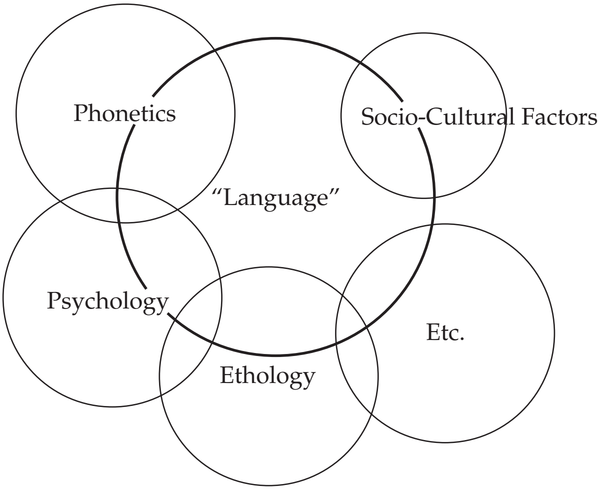


Figure 17.1 Venn diagram illustrating the overlapping domains of Linguistics and other disciplines. (After Anderson, 1981)

achieving any greater insight into the mechanisms of speech. If one is committed to seek explanations for phenomena, not simply to relabel them, then there is a chance that the area of ignorance becomes smaller with time and as a spinoff will have practical benefits in speech technology, language teaching, and in communication disorders. Such a reductionist research strategy should not be misinterpreted, as it often is, as a requirement that every phenomenon in language *must* be immediately reducible to principles from other disciplinary domains. This is unreasonable; one might as well proclaim "let ignorance be abolished!" Rather, this is a strategy of what may be called "opportunistic reductionism": when an opportunity presents itself to explain something linguistic, that opportunity should be pursued and evaluated. (See also the exchanges between Lass, 1980, and Ohala, 1987; Pierrehumbert, 1990, and Ohala, 1990b; Pierrehumbert, 1991, and Ohala & Ohala, 1991.)

There is another sense in which Anderson's Venn diagram confuses the central issue of the relation between phonology and phonetics (and psychology, ethology, socio-cultural factors, etc.). When a given phenomenon is explained (reduced), it does not imply that one discipline "owns" those facts and therefore shrinks the "turf" of another discipline – in the case of linguistics, what Anderson refers to as the domain of "Language per se." When anthropologists cite linguistic evidence of name taboos from linguistics as support for posited societal structure, who owns the notion that there are hierarchical relations between individuals in a community? When neurologists discover localization of specific linguistic functions in the brain such as pronominalization, who owns the notion of modularity of these linguistic functions? If universals of sentence prosody (as well as size sound symbolism, facial expressions involving the mouth, and sexual dimorphism of the human vocal anatomy) are argued to be governed by the same ethological principles that determine the shape of other species's agonistic vocalizations (Ohala, 1984, 1994), which discipline owns the explanans? When Bantu specialists trace the introduction of the word for "iron" in the various Bantu languages and paleontologists date the spread of iron smelting throughout sub-Saharan Africa, who owns the resulting picture of pre-historic Bantu migrations and the resulting spread of technology? Neither linguistics nor other disciplines "lose ground" by such a cross-disciplinary union of data, methods, and theories. If anything, such a marriage increases the scope of all disciplines in the partnership. It is this idea, I think, that underlies the old, hopefully not out-dated, notion of the *unity of science.* (See below.)

Anderson qualified his view on the autonomy of linguistics by acknowledging that it applied "within the limits of present knowledge." But he does not make clear *how* the present limitations of knowledge are to be overcome. Shall linguists wait for those in other disciplines to provide answers to their questions or shall linguists themselves take the initiative?

4 The Integration of Phonetics and Phonology

As argued by Laziczius and Zwirner, virtually all phonetic studies embrace and are guided by the phonological notion that speech sounds are part of a system and that the primary function of their physical make-up is to contrast with each other – both paradigmatically and syntagmatically. Thus many phonetic studies attempt to tease out the cues differentiating phonologically-specified contrasts using a corpus of minimal pairs or n-tuples. (See, e.g., Lisker & Abramson, 1964, 1970; Ladefoged, 1963; Lehiste, 1967.) The common practice within phonetics of making a given measurement (e.g., vowel duration, formant frequency) on multiple tokens and reporting the means of these measurements is evidence that phonetics seeks some sort of pronunciation norm which is more abstract than any given speech token.

But the integration of phonetics and phonology is evident in other ways as well. I will briefly mention some of the traditional questions of phonology that can benefit from phonetic studies.

4.1 How is the pronunciation of words and other posited units of language represented and processed in the head of the speaker?

The first, most candid, answer to this question is: we don't know. Even such a fundamental issue as whether phoneme-sized segments are employed – or employed at *all* stages of encoding and decoding – has not been settled satisfactorily. There is an abundance of candidate answers given to the question of how speech is represented mentally, but until they have been properly evaluated they are just speculations. Within phonology the basic criterion applied in evaluating claims about mental representation of language is *simplicity* and such related notions as *naturalness* and *elegance*. But these are quite subjective and we learn from the history of science that the workings of the universe do not always coincide with pre-conceived human preferences.[4]

Insofar as phonetic studies – or psychological studies (there is not always a clear distinction) – can shed light on the structure and processing of speech sounds in the mind of the speaker at some stages before the activation of muscle contractions and in the mind of the listener at some stages after the acoustic signal is transduced into an auditory signal, they may help us to discover other aspects of speech representation in the brain. Representative studies in this area (from among hundreds) include Stetson (1928), Kozhevnikov and Chistovich (1965), Lisker et al. (1962), Lieberman (1967), Ohala (1981, 1992b), Cutler et al. (1986), Maddieson (1989), Krakow (1993), Tuller and Kelso (1994).

4.2 *How can we explain the occurrence of common cross-language sound patterns?*

At least since the work of Passy (1890) and Rousselot (1891), parallels have been noted between synchronic, non-distinctive, variation in pronunciation, which can be discovered in fine-grained instrumental study of speech, and diachronic variation discovered via reconstruction or by the direct evidence in ancient texts. Moreover, the synchronic variation in many cases is understandable by reference to known physical phonetic principles. From this one may conclude that (1) many sound changes arise first as non-distinctive synchronic variation and (2) that it is physical principles that determine the direction of this variability, including articulation (the topological geometry of the vocal organs as well as their inertia and elasticity), aerodynamics, how given vocal tract configurations give rise to sound, and auditory principles. A cognitive element, e.g., how listeners may err in "parsing" the events in the speech signal, is also important (Ohala, 1992a, 1993a). Although speaker-specific and culture-specific psychological or cultural factors play some role in sound change (certainly in the actual triggering of sound changes), phonetic factors are the most important factors and those most amenable to experimental study in determining cross-language universals or tendencies for sound patterning, i.e., patterns in phoneme inventories, in phonotactics, as well as in morphophonemic or allophonic variation.

Though the physical constraints shaping speech sound behavior are universal, their influence on languages is probabilistic, not absolute, because there are often ways that they can be overcome. Similarly, gravity is universal but individuals are capable of walking upright; occasionally, however, they lose their balance and stumble and then gravity asserts itself and they fall.

I will briefly present two examples of phonetically explained sound patterns (see also Ohala, 1983, 1985, 1989, 1990d, 1992a, 1993b, 1994, 1995, 1997a, 1997b, 1997c; Ohala & Busà, 1995; Ohala & Lorentz, 1977; Kawasaki, 1986, 1992; Ohala & Ohala, 1991; Wright, 1986).

4.2.1 The "bias" against voiced obstruents As is well known, there is a distinct "bias" against voiced obstruents in languages. Some languages, like Mandarin and Korean, have only voiceless stops and others, like English, which have both voiced and voiceless, show a lesser frequency of occurrence of voiced stops in running speech. Voicing in fricatives is even more infrequent than in stops. This pattern arises for the following reasons. Simplifying somewhat, vocal cord vibration has two requirements: first, the vocal cords must be lightly adducted, i.e., neither pressed against one another nor too far from the midline; and, second, there must be sufficient air flowing between the vocal cords. Assuming the first requirement is met, one of the principal factors influencing the second is the state of the supraglottal cavity. Obstruents, by definition, block the flow of air out of the vocal tract. During an obstruent the air accumulates in the air space between the point of constriction and the glottis; air pressure thus increases. Eventually the air pressure above the glottis will rise to approach that below the glottis. When the pressure differential across the glottis falls below a certain value (estimated at 1 to 2 cm H_2O), the air flow will drop below the level necessary to maintain voicing. Vocal cord vibration will then stop. (See Ohala, 1983, 1990c, 1994.)

This is the principal reason for the bias against voiced obstruents. But there are many extensions and further elaborations of this principle.

The longer a stop closure is held, the more likely this constraint is to manifest itself. Thus voiced geminate stops often become devoiced, see Table 17.1. This aerodynamic constraint can be overcome (within limits) by enlarging the oral cavity during the obstruent closure in order to make more room for the accumulating air. Some enlargement happens passively due to the natural "give" or compliance of the vocal tract walls to impinging air pressure, but even more enlargement can be done actively by lowering the tongue and jaw, letting the cheeks bulge out, raising the velum, lowering the larynx, etc. This factor must be responsible for the fact that the voiced implosives in Sindhi developed from geminate voiced stops, see Table 17.2. To maintain voicing during the long (geminate) stop closure the oral cavity volume was increased, including by lowering the larynx, and a sound change occurred when listeners took the cues for this active cavity enlargement as purposeful.

However, the option of maintaining voicing by enlarging the oral cavity is less effective the further back the supraglottal closure is made because there is less

Table 17.1 Geminate devoicing (Klingenheben, 1927)

Original ("ursprüng")	*Libanon-Neusyrischen*	
naggīb	nakkīb	*trocken*
m^{e}daggel	mdukkel	*Lügner*
šaddar	šattar	*schickte*
zabben	zappen	*verkaufte*

Table 17.2 Development of implosives in Sindhi (Varyani, 1974)

Prakrit	*Sindhi*	
*pabba	paɓuɳi	*lotus plant fruit*
gaddaha	gaɗ̣ahu	*donkey*
-(g)gam̊ʈ^{h}i	ɠaɳɖ^{h}i	*knot*
b^{h}agga	b^{h}aːɠu	*fate*

Table 17.3 Stop inventories showing absence of voiced velars

Thai	p	t	k
	p^{h}	t^{h}	k^{h}
	b	d	
Chontal	p	t	k
	b	d	
	p′	t′	k′

surface area to yield to the impinging pressure and because there are few options for cavity enlargement. Thus voiced uvular and velar stops, [ɢ], [g], therefore, are vulnerable; they may lose their voicing, their stop character, or both. This no doubt underlies the frequent absence of these sounds in languages which otherwise have one or more voiceless uvular or velar stops (see Table 17.3).

Southern (Nobiin) Nubian exhibits a morphophonemic pattern where both the influence of geminates and the influence of place of articulation are manifested. See Table 17.4. Here an inflectional process meaning 'and' adds the suffix [ɔn] to a noun stem and geminates the final consonant. But if this final consonant is voiced, the geminate that results is voiceless, unless it is articulated at the furthest forward place: labial.

Statistics show that the bias against voicing in obstruents is even stronger in fricatives than in stops (Ohala, 1983). Although this may at first glance seem

Table 17.4 Morphophonemic variation in Nobiin Nubian (Bell, 1971; Ohala & Riordan, 1979)

Noun stem	*Stem + 'and'*	
fab	fobːɔn	*father*
sɛgɛd	sɛgɛtːɔon	*scorpion*
kad͡ʒ	kat͡ʃ:ɔn	*donkey*
mʊg	mʊkːɔn	*dog*

puzzling because the fricatives, unlike stops, do involve some venting of the air accumulating behind the point of constriction, other factors are involved. Optimal voicing, as mentioned above, requires maximizing the $\Delta P_{transglottal} = P_{subglottal} - P_{oral}$. Optimal frication, on the other hand, requires maximizing $\Delta \mathrm{P}_{transoral} = P_{oral} - P_{atmosphere}$. $P_{subglottal}$ and $P_{atmosphere}$ offer little or no opportunities for systematic, rapid, control. Therefore P_{oral} is the only parameter that can be controlled in order to optimize voicing and frication during voiced fricatives. But the one constraint would require keeping P_{oral} as low as possible and the other keeping it as high as possible. Obviously, it is not possible to do both simultaneously. Thus to the extent that voiced fricatives have good frication, they are liable to be devoiced (and this is true of the sibilant fricatives [z, ʒ]) and to the extent that they maintain their voicing, they are liable to have little or no frication (and this is true of the "weak" fricatives such as [β, v, ð, ɣ]; see Pickett, 1980, p. 155).

4.2.2 When nasal, labial velars behave like velars The labial velar consonants [w, ʍ, k͡p, ɡ͡b, ŋ͡m] have two simultaneous constrictions, one labial and one velar. Nevertheless, when these sounds become nasal or have a nasal assimilating in place to them, insofar as the resulting nasal is other than labial velar, they pattern like velars, rarely or never as labials (see Table 17.5). In fact, this pattern appears even in languages where in other cases the labial velars pattern like labials, e.g.,

Table 17.5 Labial velars pattern as velars in assimilating nasals (Ohala & Lorentz, 1977)

Tswana passive verb formation (Cole, 1955)

Verb root + passive suffix		*Passive verb stem*	
-bala	+ wa	-balwa	*to read*
BUT:			
-roma	+ wa	-roŋwa	*to send*
-akaɳa	+ wa	-akaŋːwa	*to think*

Kpelle definite formation (Welmers, 1962)

Indefinite	*Definite*	
ɓɔ́ɔ	ʹmɔ́ɔi	*wax*
lúu	ʹnúui	*fog, mist*
ɣîla	ʹŋilaï	*dog*
wée	ʹŋwéei	*white clay*

Table 17.6 Labial velars patterning with labials and velars. Tenango Otomi (Blight & Pike, 1976)

/h/ >	[Φ]	/ __ w
/n/ >	[ŋ]	/ __ w

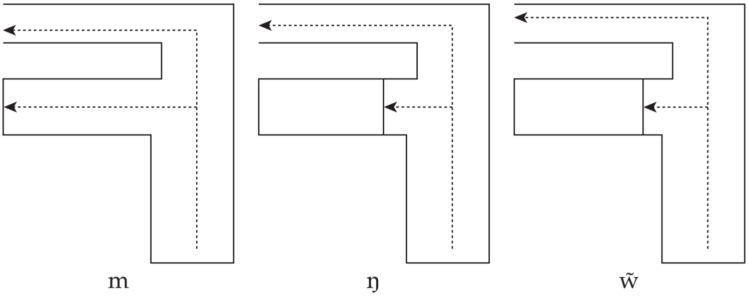


Figure 17.2 Schematic representation of the vocal tract during the production of [m], [ŋ], and [w̃]. The portion of the vocal tract contributing these sounds' resonances are shown with the dashed lines and arrows. See text for details.

when interacting with or becoming obstruents, (see Table 17.6). This pattern occurs due to the factors determining the acoustic differences between nasals. All nasal consonants have resonances from the pharyngeal-nasal air spaces; what differentiates one nasal from another is the length of the side cavity, the oral cavity, branching off of the pharyngeal-nasal passage. For this purpose it is the length of the oral cavity *measured from the pharynx forward* that matters (see Figure 17.2). In the case of a labial velar nasal, the effective length of this side cavity is that measured from the pharynx to the first, the velar, constriction. Thus labial velar nasals will tend to sound like simple velar nasals and listeners may interpret them thus.

As for why, when the labial velar approximant [w] becomes voiceless, it often becomes a labial or labio-dental, two principles can be cited. First, an approximant can become an obstruent, i.e., give rise to turbulent noise, not due to any change in the supraglottal configuration but because air moves through the constriction with greater volume and thus with greater velocity. But in principle for [w] (or [ʍ]) there should be two more or less equal noise sources at the labial and the velar constrictions. But frication noise is inherently a high frequency sound and the downstream air space in the labial velar [w] constitutes a low-pass filter to the noise produced at the innermost, the velar, constriction. The frication noise at the velar place will be attenuated. Thus the noise generated at the outermost, the labial, constriction will dominate.

4.3 Phonetics and phonological theory

To the extent that phonetic explanations such as the above are judged successful, they present challenges to phonological theory.

4.3.1 The relevance of language structure to sound change First, it has been common since Prague school work on diachronic phonology in the 1930s (Jakobson, 1931 [1972]) and subsequent work influenced by it, to propose that language structure, i.e., the system of contrasts, both paradigmatic and syntagmatic, plays an important role in motivating sound change. For example, asymmetries in the segment inventory are claimed to motivate on the one hand "filling of the gaps" – so-called "pull chains" – or, on the other hand, modification or elimination of segments to relieve the pressure due to crowding in one part of the segment inventory – so-called "push chains" (Martinet, 1968). But the similar behavior of given speech sound types in different languages, briefly reviewed above, occurs in languages that have very different structure. One is left with the impression that the physical structure of the given speech sound is more of a determinant of its diachronic fate than are the character and patterning of its sister phones. The role of a language's sound structure in diachronic phonology deserves more careful study (see Ohala & Busà, 1995).

4.3.2 Sound change is phonetically natural; should grammars be, too? Generative phonology assumes that speakers construct a grammar of their language that is simple and that simplicity correlates with the generality and naturalness of the grammatical rules (Halle, 1962; Chomsky & Halle, 1968, ch. 9). Natural and general rules thus have a preferred status in the grammar. In the past few decades there has been a continuing procession of devices and representations aimed at showing the generality and naturalness of the phonological processes embodied by rules, i.e., where the natural and general behavior of speech sounds falls out from the representation itself. Among these are features, marking conventions, autosegmental phonology, and feature geometry. But these representations fail in the vast majority of cases to represent the naturalness of phonological processes. For example, feature geometry, widely considered the most elaborate and phonetically oriented phonological representation, cannot reflect the naturalness of the processes discussed above:

1 why obstruents inhibit voicing
2 why place of articulation and the duration of a stop closure further modulate this inhibition
3 why implosives might develop out of geminate voiced stops
4 why there is a stronger bias against voiced fricatives than voiced stops
5 why labial velars tend to pattern as velars when nasal or interacting with nasals
6 why approximants become obstruents when devoiced
7 why labial velars tend to pattern as labials when becoming or interacting with obstruents

In defense of feature geometry it might be acknowledged that it is incapable of representing the naturalness of these processes since it does not incorporate phonetic principles based on aerodynamics and acoustics and such principles underlie (1)–(7). But if another, more elaborated, version of feature geometry were developed where the dependency relations of voicing on place and closure duration, etc., were incorporated then this defect could be corrected. I submit that such an elaboration if done would be identical to the phonetic models we have already (Fant, 1960; Stevens, 1971; Ohala, 1976, 1990c; Westbury & Keating, 1985; Scully, 1990, etc.). But phonology has shown considerable reluctance to adopt the continuous, physical models of speech and perhaps for good reason: it seems unlikely that the native speakers' grammars include physical principles such as Boyle's Law and the like. Among other things, grammars are thought to operate on discrete entities, not continuous parameters.

So there is an inherent problem: Grammars are supposed to give priority to natural phonological rules but the ultimate embodiment of naturalness would require rules and representations that are psychologically implausible. The resolution of this problem may require re-thinking one of the fundamental assumptions of modern phonology: that phonetic naturalness plays any role in speakers' grammars. Do speakers, in fact, recognize the difference between phonetically natural and phonetically unnatural patterning of speech sounds? There is no substantial body of evidence suggesting that they do. The fact that phonetically natural sound patterns can be found in languages does not necessarily mean that language users are aware of them. Many "natural" patterns exist in language (Ohala, 1992c) – indeed, in the universe as a whole – that may escape the attention of the individual even though he or she knows in detail the individual objects or events which manifest the pattern. Were it otherwise, there would be much less history of science; every pattern and regularity of nature would be instantly evident to everyone who observed it. The phonological grammars in speakers' heads – i.e., the rules and representations that underlie native speakers' mastery of their language – may be coded using unanalyzed phoneme-like units and large look-up tables. The phonological concord evident in vowel harmony, for example, could be handled this way, though it may seem inelegant to the linguist who is aware of the general pattern that underlies it. However, a historical and phonetic account of how natural sound patterns arise in languages – also a proper concern of the phonologist – should involve as much physical phonetics as necessary to make a convincing explanatory scenario.

In conclusion, I personally advocate a characterization of phonology as the discipline that occupies itself with the questions listed at the start of this chapter and that seeks answers to the questions by employing the methods, data, and theories from phonetics (as well as psychology, social science (including history), ethology, etc.). Inherent in this view is that phonology should not be conducted as an autonomous discipline but rather should embrace any means that will help it to get the answers it seeks.

NOTES

Portions of this paper have appeared earlier in Ohala (1991).

1 Darwin used different terminology; I am “translating: his terms into their approximate modern sense.

2 In the introduction to his *Principes de phonétique expérimentale* (1897–1908), Rousselot declared that the synchronic study of speech was intertwined with the study of the development of speech sounds in the past (p. 2).

3 There are exceptions, of course: von Raumer, Grandgent, Passy. On the other side, those who received their formative training in phonetics, there were many who were well versed in traditional historical phonology, e.g., Rousselot (a student of the Romance philologist Gaston Paris).

4 The Ptolemaic school and even the great Copernicus labored under the assumption that planetary orbits consisted of circles or one or more circular epicycles – in part because the circle was regarded as the perfect geometrical shape: “since the movement of the heavenly bodies ought to be the least impeded and most facile, the circle among plane figures offers the easiest path of motion . . . ; likewise that, since of different figures having equal perimeters those having the more angles are the greater, the circle is the greatest plane figure . . . and the heavens are greater than any other body” (Ptolemy, *The Almagest*, from Hutchins, 1952). It wasn’t until the work of Kepler at the end of the sixteenth century that astronomers accepted that orbits were ellipses. With great reluctance it was realized that nature didn’t necessarily share Man’s notions of what was simple and natural.

REFERENCES

Amman, J. C. (1694) *The Talking Deaf Man: Or, a Method Proposed Whereby He Who is Born Deaf May Learn to Speak.* London: Hawkins.

Anderson, S. R. (1981) Why phonology isn’t “natural.” *Linguistic Inquiry*, 12, 493–539.

Bell, H. (1971) The phonology of Nobiin Nubian. *African Language Review*, 9, 115–59.

Bindseil, H. E. (1838) *Abhandlungen zur allgemeinen vergleichenden Sprachlehre.* Hamburg: F. Perthes.

Blight, R. C. & Pike, E. V. (1976) The phonology of Tenango Otomi. *International Journal of American Linguistics*, 42, 51–7.

Brosses, C. de (1765) *Traité de la formation méchanique des langues, et de principes physiques de étymologie.* Paris: Chez Saillant, Vincent, Desaint.

Brücke, E. (1856) *Grundzuge der Physiologie und Systematik der Sprachlaute*. Wien.

Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York: Harper & Row.

Clark, J. & Yallop, C. (1990) *An Introduction to Phonetics and Phonology.* Oxford: Blackwell.

Cole, D. T. (1955) *An Introduction to Tswana Grammar*. London: Longmans, Green.

Court de Gebelin (1776) *Histoire naturelle de la parole ou pris de I’origine du language et de la grammaire universelle.* Paris.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986) The syllable’s differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385–40.

Cutler, A. & Norris, D. (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–21.

Darwin, E. (1803) *The Temple of Nature*. London.

Derwing, B. L. (1992) A "pause-break" task for eliciting syllable boundary judgments from literate and illiterate speakers: preliminary results for five diverse languages. *Language and Speech*, 35, 219–35.

Ellis, A. J. (1877) *Speech in Song, being the Singer's Pronouncing Primer of the Principal European Languages for which Vocal Music is Usually Composed.* London: Novello, Ewer & Co.

Fant, G. (1960) *Acoustic Theory of Speech Production.* The Hague: Mouton.

Fant, G. & Risberg, A. (1963) Auditory matching of vowels with two formant synthetic sounds. *STL-QPSR* (*Speech Transmission Laboratory – Quarterly Progress and Status Report*, Royal Institute of Technology, Stockholm), 4, 7–11.

Fischer-Jørgensen, E. (1979) A sketch of the history of phonetics in Denmark until the beginning of the 20th century. *Annual Report of Institute of Phonetics, University of Copenhagen*, 13, 135–69.

Fischer-Jørgensen, E. (1985) Review of *Grundlagen der phonometrischen Linguistik* (3rd edn.) by E. Zwirner and K. Zwirner. *Phonetica*, 42, 198–213.

Fowler, C. A. (1981) Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, 46, 127–49.

Grammont, M. (1933) *Traité de phonétique*. Paris: Librairie Delagrave.

Grandgent, C. H. (1896) Warmpth. *Publications of the Modern Language Association 11* (new series 4), 63–75.

Grassmann, H. (1854) *Leitfaden der Akustik*. Programm des Stettiner Gymnasiums.

Halle, M. (1962) Phonology in generative grammar. *Word*, 18, 54–72.

Hawkins, P. (1984) *Introducing Phonology*. London: Hutchinson.

Helmholtz, H. L. F. von (1863) *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Braunschweig: Vieweg und Sohn.

Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87, 1738–52.

Husson, R. (1950) *Etude des phenomènes physiologiques et acoustiques fondamentaux de la voix chantée*. Thèse de la Faculté de Science, Paris, Revue Scientifique, 88, 67–112, 131–46, 217–35.

Hutchins, R. M. (ed.) (1952) *Great Books of the Western World, 16: Ptolemy, Copernicus, Kepler*. Chicago: Encyclopaedia Britannica.

Jacobi, W. A. T. (1843) *Beiträge zur deutschen Grammatik*. Berlin: T. Trautwein.

Jaeger, J. J. (1984) Assessing the psychological status of the vowel shift rule. *Journal of Psycholinguistic Research*, 13, 13–36.

Jakobson, R. (1931 [1972]) Prinzipien der historischen Phonologie, *Travaux du Cercle Linguistique de Prague,* 1931, 247–67. Translated by A. R. Keiler, 1972, as Principles of historical phonology, In A. R. Keiler (ed.), *A Reader in Historical and Comparative Linguistics* (pp. 121–38). New York: Holt, Rinehart & Winston.

Jespersen, O. (1933) Karl Verner. *In Linguistica: Selected Papers in English, French, and German* (pp. 12–23). Copenhagen: Levin & Munksgaard.

Kawasaki, H. (1986) Phonetic explanation for phonological universals: The case of distinctive vowel nasalization. In J. J. Ohala & J. J. Jaeger (eds.), *Experimental phonology* (pp. 81–103). Orlando, FL: Academic Press.

Kawasaki-Fukumori, H. (1992) An acoustical basis for universal phonotactic constraints. *Language and Speech*, 35, 73–86.

Kenstowicz, M. & Kisseberth, C. (1979) *Generative Phonology: Description and Theory*. New York: Academic Press.

Key, T. H. (1855) On vowel-assimilation, especially in relation to Professor Willis' experiment on vowel-sounds. *Transactions of the Philological Society (London)*, 5, 191–204.

King, R. D. (1969) *Historical Linguistics and Generative Grammar*. Englewood Cliffs, NJ: Prentice-Hall.

Kiparsky, P. (1968) Linguistic universals and linguistic change. In E. Bach & R. T. Harms (eds.), *Universals in Linguistic Theory* (pp. 170–202). New York: Holt, Rinehart & Winston.

Klingenheben, A. (1927) Stimmtonverlust bei Geminaten. *In Festschrift Meinhof* (pp. 134–45). Hamburg: Kommissionsverlag von L. Friederichsen & Co.

Kozhevnikov, V. A. & Chistovich, L. A. (1965) *Speech: Articulation and Perception* (trans. US Department of Commerce, Clearing House for Federal Scientific and Technical Information, No. 30, 543). Washington, DC: Joint Publications Research Service.

Krakow, R. A. (1993) Nonsegmental influences on velum movement patterns: Syllables, sentence, stress, and speaking rate. In M. K. Huffman & R. A. Krakow (eds.), *Nasals, Nasalization, and the Velum* (pp. 87–116). San Diego: Academic Press.

Ladefoged, P. (1963) Some physiological parameters in speech. *Language and Speech*, 6, 109–19.

Lahiri, A. & Marslen-Wilson, W. (1991) The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, 38, 245–94.

Lass, R. (1980) *On Explaining Language Change*. Cambridge: Cambridge University Press.

Lass, R. (1984) *Phonology: An Introduction to Basic Concepts*. Cambridge/New York: Cambridge University Press.

Laziczius, G. (1948) Phonétique et phonologie. *Lingua*, 1, 293–302. (Reprinted in *Selected Writings of Gyula Laziczius* (pp. 95–104), 1966. The Hague: Mouton.)

Lehiste, I. (ed.) (1967) *Readings in Acoustic Phonetics.* Cambridge: MIT Press.

Lehmann, W. P. (1967) *A Reader in Nineteenth-Century Historical Indo-European Linguistics.* Bloomington, IN: Indiana University Press.

Lieberman, P. (1967) *Intonation, Perception, and Language.* Cambridge, MA: MIT Press.

Lisker, L. & Abramson, A. S. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422. (Reprinted in R. D. Kent, B. S. Atal, & J. L. Miller (1991) *Papers in Speech Communication: Speech Production* (pp. 671–88). Woodbury, NY: Acoustical Society of America.)

Lisker, L. & Abramson, A. (1970) The voicing dimension: Some experiments in comparative phonetics. In B. Hala, M. Romportl, & P. Janota (eds.), *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 563–7). Prague: Czechoslovak.

Lisker, L., Cooper, F. S., & Liberman, A. M. (1962) The uses of experiment in language description. *Word*, 18, 82–106.

Maddieson, I. (1989) Linguo-labials. In R. Harlow & R. Hooper (eds.), *Proceedings of the 5th International Congress of Austronesian Linguistics: Oceanic Languages* (pp. 349–75). Auckland: Linguistic Society of New Zealand.

Martinet, A. (1968) Phonetics and linguistic evolution. In B. Malmberg (ed.), *Manual of Phonetics* (pp. 464–87). Amsterdam: North-Holland.

Miller, G. A. & Nicely, P. E. (1955) Analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27, 338–53.

Morais, J. & Kolinsky, R. (1994) Perception and awareness in phonological processing: The case of the phoneme. *Cognition*, 50, 287–97.

Ohala, J. J. (1976) A model of speech aerodynamics. *Report of the Phonology Laboratory* (Berkeley), 1, 93–107.

Ohala, J. J. (1981) Speech timing as a tool in phonology. *Phonetica*, 38, 204–12.

Ohala, J. J. (1983) The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (ed.), *The Production of Speech* (pp. 189–216). New York: Springer.

Ohala, J. J. (1984) An etiological perspective on common cross-language utilization of f_0 of voice. *Phonetica*, 41, 1–16.

Ohala, J. J. (1985) Around flat. In V. A. Fromkin (ed.), *Phonetic linguistics: Essays in Honor of Peter Ladefoged* (pp. 223–41). Orlando, FL: Academic Press.

Ohala, J. J. (1987) Explanation in phonology: Opinions and examples. In W. U. Dressier, H. C. Luschiitzky, O. E. Pfeiffer, & J. R. Rennison (eds.), *Phonologica 1984* (pp. 215–25). Cambridge: Cambridge University Press.

Ohala, J. J. (1989) Sound change is drawn from a pool of synchronic variation. In L. E. Breivik & E. H. Jahr (eds.), *Language Change: Contributions to the Study of Its Causes* (pp. 173–98). Berlin: Mouton de Gruyter.

Ohala, J. J. (1990a) There is no interface between phonetics and phonology. *Journal of Phonetics*, 18, 153–71.

Ohala, J. J. (1990b) A response to Pierrehumbert's commentary. In J. Kingston & M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 280–2). Cambridge: Cambridge University Press.

Ohala, J. J. (1990c) Respiratory activity in speech. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 23–53). Dordrecht: Kluwer.

Ohala, J. J. (1990d) The phonetics and phonology of aspects of assimilation. In J. Kingston & M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 258–75). Cambridge: Cambridge University Press.

Ohala, J. J. (1991) The integration of phonetics and phonology. *Proceedings of the 12th International Congress of Phonetic Sciences*, 1, 1–16. Aix-en-Provence, 19–24 August.

Ohala, J. J. (1992a) What's cognitive, what's not, in sound change. In G. Kellermann & M. D. Morrissey (eds.), *Diachrony within Synchrony: Language History and Cognition* (pp. 309–55). Frankfurt: Peter Lang.

Ohala, J. J. (1992b) The segment: Primitive or derived? In G. J. Docherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 166–83). Cambridge: Cambridge University Press.

Ohala, J. J. (1992c) The costs and benefits of phonological analysis. In P. Downing, S. D. Lima, & M. Noonan (eds.), *The Linguistics of Literacy* (pp. 211–37). Amsterdam/Philadelphia: John Benjamins.

Ohala, J. J. (1993a) Sound change as nature's speech perception experiment. *Speech Communication*, 13, 155–61.

Ohala, J. J. (1993b) The perceptual basis of some sound patterns. In B. A. Connel & A. Arvaniti (eds.), *Papers in Laboratory Phonology TV: Phonology and Phonetic Evidence*. Cambridge: Cambridge University Press.

Ohala, J. J. (1994) The frequency code underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. J. Ohala (eds.), *Sound Symbolism* (pp. 325–47). Cambridge: Cambridge University Press.

Ohala, J. J. (1995) A probable case of clicks influencing the sound patterns of some European languages. *Phonetica*, 52, 160–70.

Ohala, J. J. (1997a) Phonetics in phonology. *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)*, 11–15 August, 45–50. (Also published 1999 in the Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm 4: Selected Papers from SICOL '97*, pp. 105–13.)

Ohala, J. J. (1997b) Aerodynamics of phonology. *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)*, 11–15 August, 92–7.

Ohala, J. J. (1997c) Emergent stops *Proceedings of the 4th Seoul International Conference on Linguistics (SICOL)*, 11–15 August, 84–91.

Ohala, J. J. & Busà, M. G. (1995) Nasal loss before voiceless fricatives: A perceptually-based sound change. In C. A. Fowler (ed.), *Rivista di Linguistica* (Special issue on The Phonetic Basis of Sound Change), 7, 125–44.

Ohala, J. J. & Lorentz, J. (1977) The story of [w]: An exercise in the phonetic explanation for sound patterns. *Berkeley Linguistic Society, Proceeding of Annual Meeting*, 3, 577–99.

Ohala, J. J. & Ohala, M. (1991) Reply to commentators. *Phonetica*, 48, 271–4.

Ohala, J. J. & Ohala, M. (1995) Speech perception and lexical representation: The role of vowel nasalization in Hindi and English. In B. A. Connel & A. Arvaniti (eds.), *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence* (pp. 41–60). Cambridge: Cambridge University Press.

Ohala, J. J. & Riordan, C. (1979) Passive vocal tract enlargement during voiced stops. In J. J. Wolf & D. H. Klatt (eds.), *Speech Communication Papers* (pp. 89–92). New York: Acoustical Society of America.

Panconcelli-Calzia, G. (1904) De la nasalité en Italien. Doctoral dissertation. Institut de Laryngologie et Orthopédie, Université de Paris.

Passy, P. (1890) *Etude sur les changements phonétiques et leur caractères generaux*. Paris: Librairie Firmin-Didot.

Pedersen, H. (1924) *Sprogvidenskaben i det nittende aarhundrede*. Copenhagen: Gylden-dalske Boghandel Nordisk Forlag.

Pickett, J. M. (1980) *The Sounds of Speech Communication.* Baltimore, MD: University Park Press.

Pierrehumbert, J. (1990) On the value of reductionism and formal explicitness in phonological models: Comments on Ohala's paper. In J. Kingston & M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 276–9). Cambridge: Cambridge University Press.

Pierrehumbert, J. (1991) The whole theory of sound structure. *Phonetica*, 48, 223–32.

Raumer, R. von (1863) *Gesammelte sprachwissenschaftliche Schriften*. Frankfurt: Heyder & Zimmer.

Rapp, K. M. (1836) *Versuch einer Physiologie der Sprache nebst historischer Entwicklung der abendländischen Idiome nach physiologischen Grundsätzen*. Stuttgart und Tubingen: Cotta.

Read, C., Zhang, Y., Nie, H., & Ding, B. (1986) The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, 24, 31–44.

Roudet, L. (1910) *Eléments de phonétique générale*. Paris: Librairie Universitaire.

Rousselot, P. J. (1891) *Les modifications phonétiques du langages*. Paris: H. Welter.

Rousselot, P. J. (1897–1908) *Principes de phonétique expérimentale*, parts 1 and 2. Paris: H. Welter.

Rousselot, P. J. (1903) *Phonétique expérimentale et surdité*. Paris.

Sapir, E. (1925) Sound Patterns in Language. *Language*, 1, 37–51.

Scully, C. (1990) Articulatory synthesis. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 151–86). Dordrecht: Kluwer.

Sievers, E. (1881) *Grundziige der Phonetik*. Leipzig: Breitkopf & Hartel.

Sommerstein, A. H. (1977) *Modern Phonology*. Baltimore: University Park Press.

Spargo, J. W. (1931) Translator's preface. In H. Pedersen, *Linguistic Science in the Nineteenth Century: Methods and Results.* Cambridge, MA: Harvard University Press.

Stetson, R. H. (1928) *Motor Phonetics: A Study of Speech Movements in Action*.

La Haye: Archives néerlandaises de phonétique experimentale, 3.

Stevens, K. N. (1971) Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, 50, 4, 1180–92.

Sweet, H. (1877) *Handbook of Phonetics*. Oxford: Clarendon. Cited by Wood, S. A. J. (1993) Crosslinguistic cineradiographic studies of the temporal coordination of speech gestures. *Working Papers* (Lund University), 40, 251–63.

Techmer, F. (1880) *Phonetik. Zur vergleichenden Physiologie der Stimme und Sprache*. Leipzig: Wilhelm Engelmann.

Trubetzkoy, N. (1933) La phonologie actuelle. *Journal de Psychologie*, 30, 227–46.

Trubetzkoy, N. (1939) Grundzüge der Phonologie. *Travaux du cercle linguistique de Prague* 7. Trans. C. A. M. Baltaxe 1969, as *Principles of Phonology*. Berkeley: University of California Press.

Tuller, B. & Kelso, J. A. S. (1994) Action theory and the production of speech. In R. E. Asher & J. M. Y. Simpson (eds.), *The Encyclopedia of Language and Linguistics*, vol. 1 (pp. 21–4). Oxford: Pergamon.

Varyani, P. L. (1974) Sources of implosives in Sindhi. *Indian Linguistics*, 35, 51–4.

Verner, K. (1875) Eine Ausnahme der ersten Lautverschiebung. *Zeitschrift für vergleichende Sprachforschung*, 23, 97–130.

Verner, K. (1913) Letters to Hugo Pipping. *Oversigt over del kongelige danske videnskabemes selskabs forhandlinger* (Copenhagen), 3, 161–211.

Wang, H. S. & Derwing, B. L. (1986) More on English vowel shift: The back vowel question. *Phonology Yearbook*, 3, 99–116.

Welmers, W. E. (1962) The phonology of Kpelle. *Journal of African Languages* 1, 69–93.

Westbury, J. R. & Keating, P. A. (1985) On the naturalness of stop consonant voicing. *Working Papers in Phonetics* (UCLA), 60, 1–19.

Weymouth, R. F. (1856) On the liquids, especially in relation to certain mutes. *Transactions of the Philological Society* (London), 18–32.

Willis, R. (1830) On the vowel sounds, and on reed organ-pipes. *Transactions of the Cambridge Philosophical Society*, 3, 229–68.

Wright, J. T. (1986) The behavior of nasalized vowels in the perceptual vowel space. In J. J. Ohala & J. J. Jaeger (eds.), *Experimental Phonology* (pp. 45–67). Orlando, FL: Academic Press.

Zwirner, E. & Zwirner, K. (1936) *Grundfragen der Phonometrie*. Berlin: Metten & Co.

18 Phonetic Notation

JOHN H. ESLING

1 Introduction

Many excellent treatises on phonetic notation exist (Sweet, 1877, 1880–1, 1971; Abercrombie, 1953, 1991, pp. 91–100; Wells, 1976, 2006; Catford, 1981; Laver, 1994, pp. 549–62; MacMahon, 1996; Kemp, 2006). The present chapter may be regarded as an update. The information presented here derives primarily from the International Phonetic Association's revision to the International Phonetic Alphabet as of 2005. To emphasize its consistency and auditory accuracy, the tradition of IPA notation is traced back to 1926. The IPA does not hold a monopoly on auditory accuracy in phonetic notation; it is represented here as a common core of standard usage that transcribers of language can universally refer to and understand. It is the notational system that can be considered the most widely relied on international standard for the phonetic transcription of dictionary entries (Rey & Rey-Debove, 1988, p. xii; Roach & Hartman, 1997, pp. viii–xv; de Wolf et al., 1998, p. vi). Within any system of phonetic notation, there are bound to be areas of ambiguity that challenge the users of the system in practical terms and that challenge the very theoretical constructs of the system itself. Therefore, this chapter will also consider aspects of transcription practice that may force theoretical issues and that may require reformulation of both the way in which certain symbols are used and the theoretical phonetic relationships that pertain between sets of symbols.

1.1 Written transcription of speech

Phonetic notation refers to the ways in which speech sounds are written down or "transcribed" visually. The ideal goal is to be able to transcribe any sound that can occur in any language of the world. Since the early twentieth century, in the European tradition, this task has been accomplished most commonly using alphabetic notation. In the nineteenth century, however, the tradition had iconic origins, specifically in the articulatorily based *Visible Speech* notation developed by Alexander Melville Bell (1867) and elaborated by Alexander Graham Bell (1872, 1906). Orthographic notation, as opposed to phonetic notation, represents the sounds of speech of the world's languages in a variety of ways depending on the cultural traditions of the particular language or language group. Orthographies are commonly categorized as ideographic, syllabic, and alphabetic (Daniels & Bright, 1996). The principal difference between orthographic representations and phonetic representations of the speech sounds of a language is that an orthography may not represent every sound that occurs, may not represent every sound uniquely, and may maintain historical representations that no longer indicate how the particular language forms are pronounced. That is, an orthographic writing system may choose to represent words, syllables, or groups of sounds of the language rather than the individual speech sounds that a phonetic form of notation typically isolates in a string of auditorily distinct articulatory maneuvers, representing them one after the other as far as the particular phonetic theory allows. There may also be several ways in a given orthography of writing what is in effect the "same" phonetic sound. This can be the result of the orthographic system's cultural persistence over time, where various ways of representing the same speech sound evolve and are maintained in orthographic usage. Changes in pronunciation over time may also not continue to be reflected in the symbolization used in the orthography. An orthography may maintain, therefore, more symbols or combinations of symbols than there are distinctive sounds in the contemporary pronunciation of a language. In addition to these basic differences, there is of course also the matter of dialect change and divergence, whereby various forms of a language diverge while still maintaining the same or a similar orthography to write the language. It is usual for many historical orthographic representations to persist for centuries as these changes and splits occur and as new language groups adopt older, neighboring, or innovative orthographies as their cultural norm. A single language can illustrate all of these traits in its orthographic repertory. For example, English shorthand can be both ideographic at the word level and phonetic at other levels. Numerals such as "8" are of course word-level representations; and there are several ways in standard English orthography of writing vowels such as /i/ ("m*ea*t," "gr*ee*n," "sc*e*n*e*") or /e/ ("m*ai*n," "m*a*n*e*," "d*ay*," "*ei*ght") or consonants such as /f/ or /k/ ("*ph*ysics," "*c*ou*gh*," "*kick*"); and sequences such as "gh" may have lost their historical value in words such as "ni*gh*t" where only a vowel remains.

Phonetic transcription systems have evolved with the intention of representing in a universal way the speech sounds that can occur in human language, irrespective of the orthographic means by which they may be written. This implies that the basis of notation is the "speech sound" rather than a phonological, morphological, or syntactic entity. Nevertheless, phonological considerations may enter into the choice of representations, distinguishing a "broad" phonemic level of transcription from the "narrow" phonetic level (Abercrombie, 1964). Universality also implies that one speech sound will have one and the same notational representation, whatever languages it occurs in. This in turn assures that it remains possible to transcribe a language variety phonetically even after sound changes occur and dialect pronunciations diverge to the point that they are no longer reflected in the popular orthography. The principles that phonetic notation should follow are outlined in the *Handbook of the International Phonetic Association*, beginning with the theoretical assumptions that underlie how speech is analyzed and repeating policy statements first developed in the late nineteenth century (IPA, 1999, pp. 3–4, 159–60, 196). The main underlying tenet is that each "distinctive sound," in the sense of a sound that is meaningfully distinct from another in a particular phonology, merits its own sign (IPA, 1999, p. 27). Some more narrow, detailed, "allophonic" differences in sound quality are intended to be shown with supplementary symbolization – with smaller letters or with diacritics (IPA, 1999, p. 28). Other allophones may nevertheless have full symbols available for their transcription, e.g. [ɱ] for /m/ before labiodental fricatives in English. Although the phonetic speech sound referents of particular symbols are quite concise, certain symbols can be used as superordinate "cover symbols" in broad transcription to represent an entire class of more narrowly transcribed sounds/symbols (Laver, 1994, p. 553). Typically, the broadly representative symbol is either presented in orthographic form, e.g., *r*, or enclosed in phonemic slashes, e.g. /r/, while the submembers of the set are referred to within square brackets, e.g. [ɹ, ɾ, r, ʀ, ʁ].

1.2 Iconic phonetic notation

Iconicity in notation is another level of consideration. It exists in ideographic systems, but it is not absent in alphabetic or in phonetic notation. Abercrombie (1991, p. 93) outlines the parallel between the development of shorthand, of phonetic notational systems, and of writing systems in general. Articulatory iconicity exists at various levels and may be at the root of most writing systems. Nevertheless, however cogent articulatory iconicity may have been for the inventor of a writing system, the contemporary learner of the system can assemble many symbols of diverse provenance to logically represent the spoken forms. In phonetic notation, systems of transcription that have remained primarily iconic, such as Bell's *Visible Speech* (1867), have their limitations and are not as efficient for the greatest number of users as a system that is primarily alphabetic, however iconic the development of those alphabetic characters may once have been. The incorporation over the years of characters that possess widely recognized culturally familiar meanings has proved much more usable. In the case of the IPA, most of these characters have come from the Greco-Roman alphabetic tradition, and to this extent the IPA has a European bias.

1.3 Alphabetic phonetic notation: The International Phonetic Alphabet

The iconic system of Bell's *Visible Speech* was modified by Henry Sweet in the 1880s and called the *Revised Organic Alphabet* (MacMahon, 1996, p. 839). Paul Passy and Daniel Jones produced a similar system in 1907, called simply the *Organic*

Alphabet (MacMahon, 1996, p. 840). By 1926, the International Phonetic Alphabet had evolved from an iconic form to an alphabetic form. The 1926 IPA chart is presented in Figure 18.1 as an early example (IPA, 1926).

Many modifications have been made over the years to the Association's official alphabet (e.g., Wells, 1976; IPA, 1989a), but the 1926 version still bears a striking similarity to the set of symbols in use in 2005. The 2005 revision to the IPA is presented in Figure 18.2. It has the basic form of the chart that grew out of the 1989 Kiel Convention to revise the IPA (IPA, 1989a), hosted by the Institute of Phonetics and Digital Speech Processing of the Christian-Albrechts-University of Kiel, and which was published after two more rounds of revisions as the 1996 IPA chart. The vowel chart in particular remains virtually intact, despite several interim alterations ventured between 1926 and the 1996 revision. Only the rounded version of the open front vowel [ɶ], missing but implicit by virtue of its position in the 1926 version, has been explicitly added. Of the four mid-central vowels in addition to schwa that were present in 1926, all continue to be recognized, and only the shape of the symbol for the half-open central rounded vowel has reversed its direction from the 1926 representation of that sound. Chronicling it as Cardinal Vowel 22, Abercrombie (1967) and Catford (1977, p. 179) already used the closed reversed epsilon shape [ɞ], which was reinstated in the IPA chart as of 1996.

The symbol shapes themselves derive from various sources, as explained in detail by MacMahon (1996, pp. 822–3). Some are traditionally Roman in origin, with values very close to their widely proliferated orthographic usage; some are extended Roman characters; some are Greek; a few are adaptations of familiar letters or punctuation marks; a few are derived from letters or diacritics used in other languages; and some are new creations. Symbols have an articulatory–auditory basis, whereby each particular sound is associated with a specific symbol, and a series of sounds with similar articulatory shapes comes to be associated with a particular set of symbols. Despite the mandate that there should be one symbol for one sound, this is not strictly the case. The symbols [t] and [d] etc. can be used for Dental, Alveolar, or Postalveolar sounds, but strictly speaking, when those sounds are Dental, they should be designated with a dental diacritic ([t̪ d̪]). Also, the ejectives are marked with an apostrophe (e.g. [p’ t’]), to represent glottal closure, and when stops are aspirated ([p^h t^h]), a superscript [^{h}] (a puff of breath) is added in narrow transcription.

A number of formal conventions govern the IPA chart. Place of articulation of the consonant sounds is presented from left to right. In the time of A. M. and A. G. Bell, when purely iconic articulatory notation was used, the chart was turned the other way round, with the lips at the right, and their iconic symbols reflected this fact. By the time of the 1926 chart, the convention had changed, with the lips being placed at the left. This was probably a typesetting expediency, reflecting a preference to set tables from left to right. Manner of articulation is now presented at the left of the chart, beginning at the top with sounds that exhibit full closure (Plosives or “Stops”) and Nasals, which also have oral closure. Trills, Taps/Flaps, and Fricatives have progressively greater opening of the articulators, and Approximants (at least central or “median” approximants) have the greatest

		Bi-labial	Labio-Dental	Dental and Alveolar	Retroflex	Palato-Alveolar	Alveolo-Palatal.	Palatal	Velar	Uvular	Pharyngal	Glottal
CONSONANTS	Plosive	p b		t d	ʈ ɖ			c ɟ	k g	q ɢ		ʔ
	Nasal	m	ɱ	n	ɳ			ɲ	ŋ	ɴ		
	Lateral			ɬ l ł	ɭ			ʎ	(ł)			
	Rolled			r ř						ʀ		
	Flapped			ɾ	ɽ					ʀ		
	Fricative	ꜰ ʋ ƕ	f v	θ ð s z ɹ	ʂ ʐ	ʃ ʒ	ɕ ʑ	ç j	(ƕ) x ǥ	χ ʁ	ħ ʕ	h ɦ
	Semi-vowel	w ɥ						j (ɥ)	(w)			
VOWELS								Front	Central	Back		
	Close	(u ʉ y)						i y	ɨ ʉ	ɯ u		
		(ʊ ʏ)						ɪ ʏ		ʊ		
	Half-close	(o ɵ ø)						e ø	ɘ ɵ	ɤ o		
									ə			
	Half-open	(ɔ ɞ œ)						ɛ œ	ɜ ɞ	ʌ ɔ		
								æ	ɐ			
	Open								a	ɑ ɒ		

(Sounds appearing twice on the chart have a double articulation, the secondary articulation being shown by the symbol in brackets.)

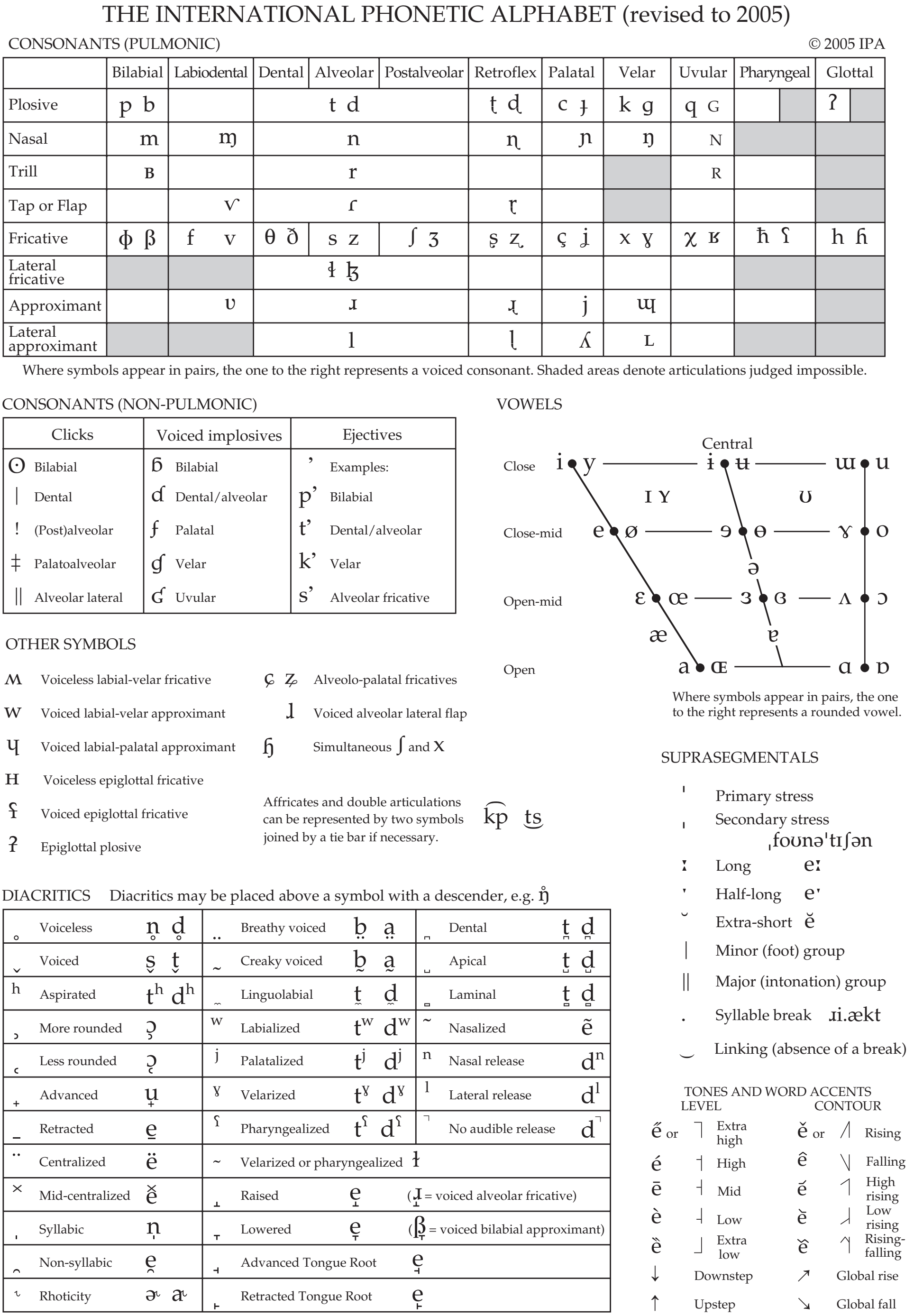
Other Sounds.—Palatalized consonants: ƫ, ȡ, etc. Velarized or pharyngalized consonants: đ, ƶ, etc. Ejective consonants (plosives with simultaneous glottal stop): p', t', etc. ɮ (fricative l). ƍ, ϱ (labialized θ, ð, or s, z). ʅ, ʓ (labialized ʃ, ʒ). ʇ, ʗ, ʖ (clicks, Zulu *c*, *q*, *x*). ɺ (a sound between r and l).

Affricates are normally represented by groups of two consonants (ts, tʃ, dʒ, etc.), but, when necessary, ligatures are used (ʦ, ʧ, ʤ, etc.), or the marks ͡ or ‿ (t͡s or t‿s, etc.).

Length, Stress, Pitch.— ː (full length). ˑ (half length). ˈ (stress, placed at beginning of the stressed syllable). ˉ (high level pitch); ˍ (low level); ˊ (high rising); ˏ (low rising); ˋ (high falling); ˎ (low falling); ˆ (rise-fall); ˇ (fall-rise). See *Ecriture Phonétique Internationale*, p. 9.

Modifiers.— ˜ nasality. ̥ breath (l̥ = breathed l). ̬ voice (s̬ = z). ʻ slight aspiration following p, t, etc. ̣ retroflexion (inversion) (ạ = a with curled up tongue: ṭ = ʈ). ˙ palatalization (ż = ʑ). ˔ tongue slightly raised. ˕ tongue slightly lowered. ˒ lips more rounded. ˓ lips more spread. ˌ syllabic consonant. ˘ consonantal vowel. ʃˢ variety of ʃ resembling s, etc.

Figure 18.1 The chart of the International Phonetic Alphabet in 1926 from *Le Maître Phonétique* (IPA, 1926). (Reproduced by permission of the International Phonetic Association)



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d			ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ	n			ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ		r					▨	ʀ		▨
Tap or Flap		ⱱ	ɾ			ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨	ɬ ɮ							▨	▨
Approximant		ʋ	ɹ			ɻ	j	ɰ			▨
Lateral approximant	▨	▨	l			ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

VOWELS

		Central	
Close	i y	ɨ ʉ	ɯ u
	ɪ ʏ		ʊ
Close-mid	e ø	ɘ ɵ	ɤ o
		ə	
Open-mid	ɛ œ	ɜ ɞ	ʌ ɔ
	æ	ɐ	
Open	a ɶ		ɑ ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative
w	Voiced labial-velar approximant
ɥ	Voiced labial-palatal approximant
ʜ	Voiceless epiglottal fricative
ʢ	Voiced epiglottal fricative
ʡ	Epiglottal plosive
ɕ ʑ	Alveolo-palatal fricatives
ɺ	Voiced alveolar lateral flap
ɧ	Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͜s

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

̥	Voiceless	n̥ d̥	̤	Breathy voiced	b̤ a̤	̪	Dental	t̪ d̪
̬	Voiced	s̬ t̬	̰	Creaky voiced	b̰ a̰	̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	̼	Linguolabial	t̼ d̼	̻	Laminal	t̻ d̻
̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	̃	Nasalized	ẽ
̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	̚	No audible release	d̚
̈	Centralized	ë	̴	Velarized or pharyngealized	ɫ			
̽	Mid-centralized	e̽	̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
̩	Syllabic	n̩	̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
̯	Non-syllabic	e̯	̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	̙	Retracted Tongue Root	e̙			

SUPRASEGMENTALS

ˈ	Primary stress	
ˌ	Secondary stress	ˌfoʊnəˈtɪʃən
ː	Long	eː
ˑ	Half-long	eˑ
̆	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˦˥	High rising
è	˨	Low	e᷅	˩˨	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
↓		Downstep	↗		Global rise
↑		Upstep	↘		Global fall

Figure 18.2 The chart of the International Phonetic Alphabet revised to 2005. (Reproduced by permission of the International Phonetic Association)

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	101 102		103 104			105 106	107 108	109 110	111 112		113
Nasal	114	115	116			117	118	119	120		
Trill	121		122						123		
Tap or Flap		184	124			125					
Fricative	126 127	128 129	130 131	132 133	134 135	136 137	138 139	140 141	142 143	144 145	146 147
Lateral fricative			148 149								
Approximant		150	151			152	153	154			
Lateral approximant			155			156	157	158			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
176	Bilabial	160	Bilabial	401	Examples:
177	Dental	162	Dental/alveolar	101 + 401	Bilabial
178	(Post)alveolar	164	Palatal	103 + 401	Dental/alveolar
179	Palatoalveolar	166	Velar	109 + 401	Velar
180	Alveolar lateral	168	Uvular	132 + 401	Alveolar fricative

OTHER SYMBOLS

169 Voiceless labial-velar fricative

170 Voiced labial-velar approximant

171 Voiced labial-palatal approximant

172 Voiceless epiglottal fricative

174 Voiced epiglottal fricative

173 Epiglottal plosive

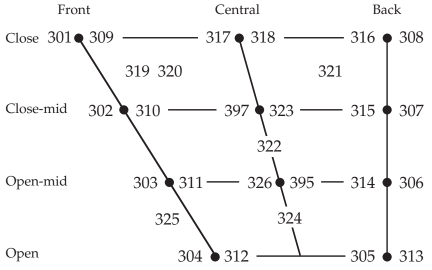
182 183 Alveolo-palatal fricatives

181 Voiced alveolar lateral flap

175 Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. 433 (509)

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

501 Primary stress

502 Secondary stress

ˌfoʊnəˈtɪʃən

503 Long eː

504 Half-long eˑ

505 Extra-short ĕ

507 Minor (foot) group

508 Major (intonation) group

506 Syllable break ɹi.ækt

509 Linking (absence of a break)

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. 119 + 402B

402A	Voiceless	n̥ d̥	405	Breathy voiced	b̤ a̤	408	Dental	t̪ d̪
403	Voiced	s̬ t̬	406	Creaky voiced	b̰ a̰	409	Apical	t̺ d̺
404	Aspirated	tʰ dʰ	407	Linguolabial	t̼ d̼	410	Laminal	t̻ d̻
411	More rounded	ɔ̹	420	Labialized	tʷ dʷ	424	Nasalized	ẽ
412	Less rounded	ɔ̜	421	Palatalized	tʲ dʲ	425	Nasal release	dⁿ
413	Advanced	u̟	422	Velarized	tˠ dˠ	426	Lateral release	dˡ
414	Retracted	e̠	423	Pharyngealized	tˤ dˤ	427	No audible release	d̚
415	Centralized	ë	428	Velarized or pharyngealized 209				
416	Mid-centralized	e̽	429	Raised	e̝ (ɹ̝ = voiced alveolar fricative)			
431	Syllabic	n̩	430	Lowered	e̞ (β̞ = voiced bilabial approximant)			
432	Non-syllabic	e̯	417	Advanced Tongue Root	e̘			
419	Rhoticity	327 ɚ	418	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
512	519	Extra high	524	529	Rising
513	520	High	525	530	Falling
514	521	Mid	526	531	High rising
515	522	Low	527	532	Low rising
516	523	Extra low	528	533	Rising-falling
517	Downstep		510	Global rise	
518	Upstep		511	Global fall	

Figure 18.3 The number chart of the International Phonetic Alphabet revised to 2005. (Reproduced by permission of the International Phonetic Association)

degree of articulatory opening before a segment is open enough to be recognized phonetically as a vowel. Lateral approximants also have oral opening, but at the sides (or one side) of the mouth, hence the label Lateral fricative and Lateral approximant. There are two other critical parameters represented in the matrix of symbols. Consonant sounds that are voiceless (without vibration of the vocal folds) are designated by the symbol at the left in a pair of symbols, while sounds that are voiced (when vocal fold vibration is present) are designated by the symbol at the right in a pair of symbols in the pulmonic consonant columns. In the case of vowels, symbols representing an unrounded vowel (with spreading of the lips) are found at the left in a pair of symbols that have otherwise identical articulatory shapes, with the symbols representing a rounded vowel (with rounding of the lips) being designated by the symbol at the right in a pair. These conventions reflect the hierarchical interplay of possible stricture modifications of the principal articulators, which can be thought of very generally (in stream-of-speech order) as the larynx (within it, the glottis and the pharynx), the velopharyngeal port, the tongue, the lips, and the jaw. Any system of phonetic notation needs to take the articulatory components of these five major regions into account, since any individual consonantal or vocalic sound implies a setting of these five areas, whether active or neutral. This will be particularly apparent in the description and transcription of voice quality (section 2.8).

In 1989, an equivalent system of "IPA Numbers" was instituted as a means of unambiguously indicating any symbol or sequence of symbols on the chart (IPA, 1989b). The 2005 version of the Number Chart is shown in Figure 18.3. This reference system is useful because it has the same visual orientation as the symbol chart. Since 1989, it has become possible to refer to individual symbol shapes using Unicode values, which are 16-bit hexadecimal versions based originally on ASCII and extended with the Unicode Consortium (2007) and the Text Encoding Initiative (2002) concepts to incorporate a unique value for each character within a worldwide set of writing system definitions including phonetic notation.

With respect to vowels, there has been some debate about how accurately the articulatory basis corresponds to the auditory basis for each symbol. Catford (1977, pp. 168–9, 176; 1981: 27–8) argues that early vowel quality descriptions evolved not just as auditory targets but that there is proprioceptive articulatory merit to their classification; that is, a vowel set can be learned articulatorily while developing proprioceptive cues based on the articulatory parameters outlined in the vowel chart. This view asserts the validity of the tongue-position model of vowel specification. Others argue that the targets are purely acoustic and that articulatory positioning may vary (Ladefoged, 1967). One moderating factor in evaluating whether tongue positions are as evenly articulatorily spaced as the Cardinal Vowel model posits is the model of the vocal tract itself. The issue is to identify where in the vocal tract the various resonances originate that contribute to vowel quality. Catford and Esling (2006) present a revised view of the vocal tract, based on Esling (2005), which could be thought of as incorporating "two vocal tracts in one" or an oral articulating compartment and a laryngeal articulating compartment. This two-part vocal tract is represented in Figure 18.4. Following

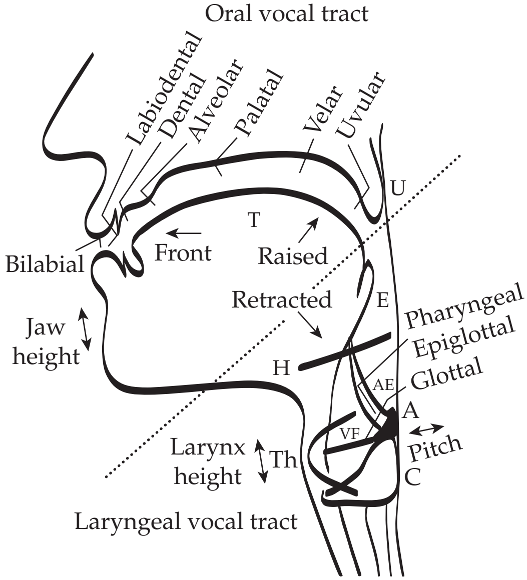


Figure 18.4 A revised depiction of the oral and laryngeal vocal tracts in which tongue motion is separated into three components and in which laryngeal activity includes glottal, aryepiglottic laryngeal constrictor, and height components. T = tongue; U = uvula; E = epiglottis; H = hyoid bone; AE = aryepiglottic folds; A = arytenoid cartilages; VF = vocal folds; Th = Thyroid cartilage; C = cricoid cartilage. (From Esling, 2005, reproduced by permission of the University of Toronto Press)

this model, it is predictable that the more open a vowel is, particularly "open back" vowels, the more likely it is that it will be susceptible to the effects of the laryngeal articulator. It is therefore to be expected that even the Cardinal Vowels in their evolution, particularly the ones in the retracted region of the vocal tract, would have been susceptible to the resonance characteristics of the "back-cavity" laryngeal articulator. For notational purposes, however, and leaving possible laryngeal coloring aside for the moment, a system based on proprioceptively determined tongue heights in the "front-cavity" oral vocal tract is entirely plausible and workable.

The implication of a vocal tract with a substantial laryngeal component built into it is that vowels acquire new representational groupings. Based on three directions of tongue movement, vowels are grouped into fronted, raised, and retracted sets. The retracted, laryngeal component of the system makes it clear that vowels in the lower-right corner of the chart fall within the domain of laryngeal, specifically pharyngeal "coloring." This region corresponds to the Pharyngeal and Glottal columns at the right-hand side of the consonant chart. To accommodate this revised view of the vocal tract, the 2005 IPA vowel chart is modified in Figure 18.5 to show the three articulatory regions in which the various auditory vocalic qualities

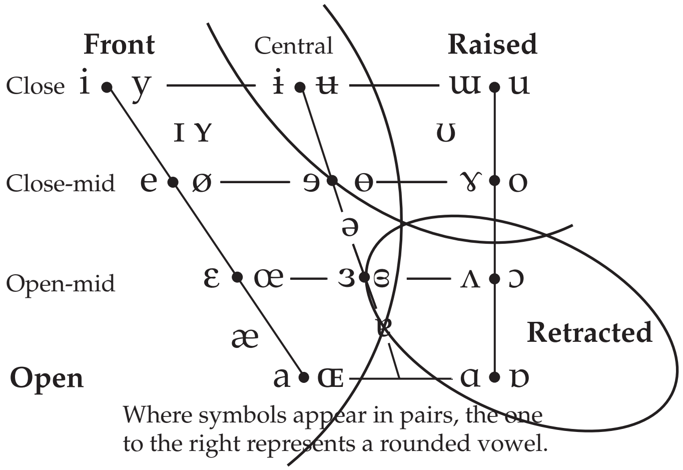


Figure 18.5 A revised conceptualization of the IPA vowel chart to correspond to the laryngeal articulator model of the vocal tract in Figure 18.4 (From Esling, 2005, reproduced by permission of the University of Toronto Press)

lie. As is the case with the Cardinal Vowels, the symbolic notation for these vowel qualities represents a long tradition of auditory and articulatory mapping, based on minimal sound-category differences in particular languages and studied phonetic comparison and training across a wide variety of languages. The regrouping merely points out the potential effects of resonances in the laryngeal articulator and that certain vowels in the mapping are more susceptible to it than are others. An example of how these effects can be noted with supplementary symbolic notation is given in section 2.2.

Perhaps the most important thing that can be said about the International Phonetic Alphabet is that the values which the symbols represent are distinct auditory qualities (combinations of bursts, silences, noise, or resonance) that have been attested as occurring in one or more languages as meaningfully contrastive speech sounds. Also, the sounds represented in series in the IPA chart are a fully expanded conceptualization of values derived from the phonological series of many languages in which these sounds occur. An attempt has been made to fill in every cell in the chart where distinctive sound contrasts have been observed. This has been done for the fricatives, for example, where contrastive speech sounds have been attested in all possible place-of-articulation categories and have been designated with a symbol.

2 Challenges to Notational Categories

There are six issues that can be taken as cases in point to illustrate where changes may need to be made to the system of phonetic notation as theories evolve to explain how categories of sounds are related. These issues include (1) the pharyngeal/epiglottal place of articulation, (2) the differentiation of vowel quality, (3) the concept of secondary articulations, (4) the notions of juncture and stress, (5) the notion of strength of articulation, and (6) tongue-front articulations. Mention is also made of the need for notational taxonomies for sounds that occur in disordered speech (section 2.7) and also for the description of voice quality (section 2.8).

2.1 Revisions to place of articulation

One outcome of the laryngeal articulator model is the remapping of pharyngeals and epiglottals onto the same place of articulation. This remapping could be viewed as a spatial reconfiguration. As reinforced by standard works such as Catford (1977), auditory description and categorical mapping is extremely robust. The challenge is finding notational equivalents for the distinctive sound categories. In the 1989 IPA chart, epiglottals were added as "Other Symbols": voiceless epiglottal fricative, voiced epiglottal fricative, and epiglottal plosive. The theory of the vocal tract prevalent at the time of the 1989 revision would have dictated that "epiglottal" be considered a categorical place of articulatory stricture between "pharyngeal" and "glottal." The reason for not including a separate "epiglottal" column among the "Pulmonic Consonants" was one of practical economy to avoid making the chart too wide. For the same practical reason, "alveolo-palatal" was listed among the "Other Symbols" rather than among the "Pulmonic Consonants" where it would have otherwise slotted in between "Retroflex" and "Palatal." The "epiglottals" had been painstakingly differentiated auditorily and catalogued by Catford (1968) with terms that sometimes combined "pharyngeal" and "epiglottal" and sometimes combined "pharyngeal" and "glottal" as ways of distinguishing the sound qualities. Taking the laryngeal articulator model as an alternative point of reference, "pharyngeal" and "epiglottal" are arguably not separate places but rather a function of the laryngeal aryepiglottic constrictor mechanism (Esling, 1996, 1999); and "glottal" and "pharyngeal" are interpreted as different degrees of valve control within the laryngeal aryepiglottic constrictor mechanism (Esling, 2005; Edmondson & Esling, 2006). The mapping of the "epiglottals" that were newly introduced in 1989 can therefore be reconceived as a conflation of place of articulation but as an elaboration of manners of articulation. In this way, for example, the epiglottal plosive can be included among "Pulmonic Consonants" as a pharyngeal plosive – the elusive "pharyngeal stop" attested by Kinkade (1967) but without a theory to support its identity at the time. The voiceless epiglottal fricative and voiced epiglottal fricative can be treated as a voiceless epiglottal trill and a voiced epiglottal trill, respectively, because trilling of the aryepiglottic folds and frication typically co-occur, in the same way that trilling and frication at the uvula often co-occur. Alternatively, [ʜ] and [ʢ] could be qualified as two new pharyngeal fricatives with higher larynx height settings than the "standard" or "basic" pharyngeal fricatives [ħ] and [ʕ] already listed in the "pharyngeal" column, depending on whether larynx height is independent of the aryepiglottic constrictor mechanism in their articulation. As a result, the epiglottal fricatives can be added to the existing set of pharyngeals, whose values have all been rethought as new manners of articulation or as new qualities of the "back-cavity" resonator. The notational symbols used to transcribe the sounds continue to be logical and robust, since the auditory sound qualities that served as the original basis for their transcription are as distinctive as before in the languages in which they occur. The new articulatory information helps to clarify what the symbols should be taken to stand for and what their distribution should be. The symbols themselves retain most of their earlier meaning, given the logical phonetic order of their development, and are extremely usable characters to fit into revised phonetic interpretations.

2.2 *Revisions to vowel characterization*

The phonetic notation for vowel sounds also has a basis in the articulatory shape of the vocal tract, which can be subject to reinterpretation. The vowel diagram in the IPA chart, remarkably consistent in content from 1926 through 2005, separates vowels into close and open in terms of jaw height (high and low in terms of tongue height), front and back with reference to tongue positioning, and unrounded or rounded in terms of lip configuration. Research into the operation of the larynx as an articulator and the pharynx as a resonating space suggests that the "back" vowels could be more explicitly divided into "raised" and "retracted" vowel qualities as shown in Figure 18.5 (Esling, 2005; Edmondson & Esling, 2006). The retracted conceptualization results from an articulatory reinterpretation of how the rear section of the vocal tract works. This in no way changes the history of phonetic listening that has come to characterize an [æ] or an [a] or an [ɑ] or an [ɒ] in the practice of vowel identification and transcriptional notation. What the remapping adds is primarily an explanation of what the various influences are that can change the quality of the vowels in a particular region in articulatory, auditory, and acoustic terms. It also reveals potential regroupings of symbols and how resonating cavity relationships in the vocal tract might differ within each particular grouping. The symbols, therefore, retain an auditory value that is understood and accepted on the basis of common training and practice. The tendencies for the so-called "coloring" of particular vowel sounds to change should not be regarded as linear functions (e.g., as high-to-low or front-to-back continua) but rather as functions of the influence of fronting, raising, or retracting. Of the three areas, retracting is the area that can exert the greatest change in background coloring on vowel quality. This is evident in the phonologies of so-called "ATR" languages, where [i ɪ], [u ʊ], and even [a ɑ] contrasts need to be transcribed not only with different symbols to indicate the quality difference but also with a diacritic to indicate the absence or presence, respectively, of laryngeal constriction, using the currently available "advanced tongue root" and "retracted tongue root" diacritics [̘] and [̙]. Phonetically, the contrastive pairings would be marked as: [i ɪ̙], [u ʊ̙], and [a ɑ̙] (Padayodi, 2008). Phonemically, the contrasts could be indicated most economically by only the absence or presence of the 'retracted tongue root′ symbol to convey that the second member of the pair has laryngeal constriction, thus: /i/ /i̞/, /u/ /u̞/, and /a/ /a̞/. Redundant marking, however, may have its advantages in both a phonetic or a phonemic marking scheme. This principle of redundancy in phonetic notation derives from the fact that the contrastive features that distinguish one phonemic consonant or vowel from another in a language may also be multiple and therefore redundant. The ATR/–ATR example demonstrates that vocalic differences may either be represented as differences in vowel quality, or as the effect of a secondary resonating chamber, or both.

2.3 *Notation of secondary articulations*

Secondary aspects of primary articulations present a challenge to notational conventions. The issue has to do with how long a secondary feature lasts and whether it should be marked on a consonant symbol, after the consonant, or on the vowel. Arabic emphatics are a case in point. Every variety of Arabic has a series of sounds with secondary "coloring" whereby, for example, [s t l] contrast with another series of sounds that have secondary tongue raising or retraction. This quality of "backing" can be captured in general by the tilde diacritic through the consonant, meaning 'Velarized or pharyngealized', e.g. [s̴ t̴ ɫ] (Catford, 1977, p. 193). In the 1989 revision of the IPA, this usage is superseded by superscript diacritics that follow the consonant symbol, i.e. [sˤ tˤ lˤ] if the quality is "Pharyngealized" (as in Thelwall & Sa′adeddin, 1999), or [sˠ tˠ lˠ] if the quality is judged to be "Velarized," or even [sʶ tʶ lʶ] if the quality is to be labeled "Uvularized." This practice, however, introduces possible ambiguity as to the sequencing of events. Since the vowel in such contexts in Arabic is also modified, another way to write the contrast would be [sa ta la] versus [sɒ tɒ lɒ]. The distinction is made redundant using the notation [sa ta la] versus [sˤɒ tˤɒ lˤɒ], but necessarily so if it is important to show that the secondary quality pervades the entire syllable. It would probably be a challenge to conceptual logic to note the difference only with vowel symbols, without appearing to qualify the consonants, e.g. [sᵃa tᵃa lᵃa] versus [sᶛɒ tᶛɒ lᶛɒ], because vowels are not usually used to qualify consonant symbols, and especially as this practice might confound the need to also mark contrastive length in Arabic.

Secondary labialization is another case in point. Simultaneous lip rounding on an [s] was once rendered with a subscript [w] as [s̫] (IPA, 1949, p. 17). In 1989, a following superscript was adopted, thus: [sʷ]; but it could be argued that this practice does not distinguish explicitly between the possibility of simultaneous rounding on the consonant and sequential rounding between the consonant and the following vowel. Thus, there could conceivably be a need to distinguish an especially rounded pronunciation of [sʷuːn] as in *soon* and of [sʷwuːn] as in *swoon* to show both simultaneity and consecutiveness of the rounding feature. This could be accomplished equally well by reverting to a generalized rounding diacritic, thus [s̹uːn] versus [s̹wuːn]. Normally, in carrying out phonetic transcription, a chosen practice is explained so that the particular usage adopted in the case at hand is apparent (Abercrombie, 1964).

2.4 Notation of stress and juncture

The device used for marking stress (most generally correlated with intensity differences) in IPA notation is a raised bar preceding a stressed syllable [ˈ]. Secondary stress is noted with a lowered bar [ˌ]. This usage is preferred to using a mark over the syllable itself, since a superscript diacritic over a vowel could be confused with an indication of tone (pitch movement) rather than stress. Differences in intensity are usually carried by the vowel of a syllable, but this is not always the case. In strings of consonants, for example, a case may need to be made that one of the elements in the string begins the focus of stress (Czaykowska-Higgins, 1993; Shaw et al., 1999). But even where consonants and vowels occur in alternating succession from syllable to syllable, there are options in the placement of the stress symbol. Typically, the stress mark is placed at the beginning of the stressed syllable, i.e., before the initial consonant. However, Payne (2005) has used an alternative formula to mark stress in Italian by focusing on the main stress-bearing unit – the vowel – rather than on the consonantal onset element of the syllable. In this interpretation, the stress mark appears immediately before the vowel symbol (typically represented by "V") of the stressed syllable of a bisyllabic sequence, regardless of whether there is a single consonant, a geminate consonant, or no consonant preceding it (consonants being represented by "C"). Thus, instead of marking a pair of syllables as ˈCVCV, or ˈCCVCV in the case of a geminate consonant (or splitting the geminate, as is often done, CˈCVCV), the stress mark is put immediately before the vowel in either case, CˈVCV or CCˈVCV.

In using phonetic notation, a decision has to be made as to where divisions occur between articulatory units. This is often accomplished by indicating a break between syllables with the diacritic for "syllable break" [.] to indicate where a sequence of syllables should be divided. This is the notion of "juncture" in the American phonemic tradition (Trager & Smith, 1951). It is of theoretical interest that this break or juncture is not necessarily a pause; it could relate to the lengthening of a vowel or consonant segment, to the stress or phonetic intensity placed on one syllable relative to another, or even to the relative pitch of units in the sequence. The precise acoustic or physical nature of the break may be indeterminate or attributable to more than one articulatory event, but the auditory/perceptual conceptualization of a break can be captured at a general phonetic level by a single notational device. This relates to the need to indicate where "stronger" articulations occur.

2.5 Notation of strength of articulation

Strength of articulation can be implied by some symbols, as can speed of articulation and also aerodynamic force. In other cases, force may need to be specified explicitly. The force with which articulators are used is implied somewhat, in a muscular sense, by the inherent scale of manners of articulation in the IPA chart. Plosives and nasals (higher up on the chart) require a greater degree of contractive muscular activity to close the articulators than do fricatives or approximants (lower down on the chart), which are produced with less stricture. Implosives, ejectives, and clicks, on the other hand, imply aspects of greater force, if only because a second articulator is used in coordination with the primary oral articulator to generate the sound – the larynx in the case of implosives and ejectives, and lingual closure in the case of clicks. Since these sounds are produced with nonpulmonic airstream mechanisms, the force with which the oral articulators operate may be greater than for their pulmonic counterparts. Speed may also be greater for the nonpulmonic consonants, as in the case of ejectives or clicks, where the oral release causes an abrupt acoustic burst. Taps and flaps are another category where speed differs from their fully stopped counterparts. Tapped [ɾ] implies a quicker oral articulation than stopped [d], and the flap [ɽ] is a faster dynamic action than stopped [ɖ]. Trills are sounds with a different requirement on the pulmonic mechanism to generate the force required for the action and also with a consequently inherently greater speed difference from other types of sounds. This force is aerodynamic rather than muscular, supplying greater pulmonic force for the production of the trill. Trills, as a consequence, may involve over 25 rapid bursts in the case of [r] and other oral trills (Hardcastle, 1976, p. 132) or a 50–60 Hz pattern in the case of aryepiglottic trilling [ʜ ʢ] (Esling & Edmondson, 2002). Apart from these inherent elements of particular sounds, it has been proposed that force of articulation can be noted with special diacritics where needed: subscript double quotes for greater force, subscript "corner" for less force (Duckworth et al., 1990 – discussed in section 2.7). When it comes to the length of time it takes to produce articulatory events (aside from those consonants or vowels marked explicitly with a length mark [ː]), some events are inherently longer than others, e.g. pharyngeals (Esling et al., 2005), but these details are not signaled explicitly in the symbolization used.

2.6 *Notation for tongue-front detail*

Oral articulations are particularly numerous in the region of the front of the tongue, especially when the tongue tip is raised. The possibilities of maneuverability afforded by the tongue front are more proliferated and elaborated than for other articulatory regions. Some cells of the IPA chart are shaded to represent categories judged articulatorily impossible, but there are no such impossibilities noted for the front lingual tip-up region of the chart. Still, the notational devices provided in this area of the chart are not adequate to represent all of the articulatory detail that the tongue front can perform. In fact, discovering the full range of articulatory elaboration in this region is one of the frontiers of theoretical phonetics that merits concentrated cross-linguistic research and corresponding notational refinement.

An example of tongue-front articulation that is not specified in the IPA notational scheme is the notion of grooving. Sweet (1877, pp. 19–20) discussed "narrow" and "wide" in the context of tongue shape in the production of vowels, but that reference is perhaps more relevant to what has been discussed in the second half of the twentieth century in terms of tense and lax vowels. Grooved tongue shape in fricatives remains a challenging articulatory phenomenon to define and to represent symbolically. Longitudinal grooving of the median lingual sulcus differs between [s] and [ʃ], but this detail is only an implicit component of the symbols. There is no separate diacritic to indicate more or less, or narrower or wider grooving. Usually, distinctions between different types of "s" sounds are indicated with diacritics that imply an alteration in the place of articulation of the sibilant. That is, [s] can be distinguished from [s̟] (advanced) and from [s̠] (retracted). It is also possible to apply diacritics to specify an apical [s̺] or a laminal [s̻], which are designations of "place" but on the active articulator. Conventional practice typically accounts only for the action of an active articulator against a passive articulator, such as "apico-alveolar" where the active tongue tip articulates against the passive alveolar ridge.

A similar situation characterizes the ambiguity of place of articulation between Alveolar, where the tongue apex can be used to produce "apical" sounds or the tongue blade can be used to produce "laminal" sounds, and Palatal, which implies sounds that are made with the tongue tip down. Between the two, a number of terms have been used to describe intervening sounds: Palatoalveolar, Postalveolar, and Alveolo-palatal. In terms of place of articulation, it could be argued that all three of these categories involve the same general location. Their origins, however, are connected with particular sounds that are widely distinguished across languages, primarily involving fricatives. Again, rather than solely being a function of place of articulation, the differences result from the position of the tongue tip (whether it is up or down) and the degree and type of grooving required. The layout of the chart implies a continuum of gradual tongue-tip retraction from [s] to [ʃ] to [ʂ]. The previous appellation for [ʃ ʒ], in rather clumsy place-of-articulation terms, was Palatoalveolar. The terminology evolved to Postalveolar, ostensibly to highlight the continuity of tongue-tip retraction and to avoid a conflict with the patently tongue-tip-down category of Alveolo-palatal. The difference between Retroflex and Alveolo-palatal (which would normally be the next place of articulation after Retroflex and before Palatal on the chart except for lack of space) is that the tongue tip is up and retracted ("receded" may be a better term) for the former, while it is down (behind the lower teeth) for the latter. So the difference is really attributable to the active articulator rather than to the passive place of articulation. In fact, in this sense, Retroflex is not really a place of articulation on the palate but rather a manner of tongue-tip shaping. The fricatives in the Alveolo-palatal category are [ɕ ʑ]. Acoustically, they have the highest noise frequency of the fricative manner of articulation, that is, they have more sibilance than their backer Palatal counterparts [ç ʝ], and this is the feature that distinguishes them most clearly auditorily. But "s" sounds also have high noise frequencies, and there is more articulatory similarity in the Alveolar to Postalveolar to Alveolo-palatal sounds [s̟ s s̠ ʃ ɕ] than the symbolization may portray. In terms of auditory distance, it is clear that they are not Retroflex, but articulatorily it remains unclear in the notation how much apical articulation, how much laminal articulation, or how much grooving is involved.

2.7 Notation for disordered speech

In clinical applications of phonetics, it has been found necessary to develop new sets of symbols and new categories to transcribe atypical speech (Duckworth et al., 1990). The motivation and chronology of this development are recounted in the *Handbook of the IPA* (IPA, 1999, pp. 186–93), where the 1997 chart of "ExtIPA Symbols for Disordered Speech" (known as the "Extended IPA") is also printed. This system adds six new place-of-articulation categories (Dentolabial, Labioalveolar, Linguolabial, Interdental, Bidental, and Velopharyngeal) and three new manner-of-articulation categories (Lateral central fricative, Nareal fricative, Percussive) to those found in standard IPA usage. In addition, a number of new conventions are proposed, including diacritic usage and the symbolization of pausing, loudness, speed, and timing features of connected speech. Most values in the ExtIPA chart are more extreme articulations than those commonly found in normal speech. For example, dentolabial, labioalveolar, bidental, and velopharyngeal articulations are not normally encountered in the sound systems of languages of the world; and the three new manners of articulation in the ExtIPA chart are not required in the IPA chart because their use can also be confined to the description of disordered speech. Nevertheless, "normal" is a relative designation, and some of these less commonly encountered categories could prove useful in developing wider-reaching phonetic theories. At the very least, the ExtIPA taxonomy broadens the symbolic capability of the basic IPA notational system with considerably more detail than was formerly possible. The proliferation of notational categories (for instance, designating the simultaneous presence of whistling, of nasal airflow, or of velopharyngeal friction, or specifying an opposite direction of airflow) offers an opportunity to expand on the ways in which particular articulatory components can be related to others; and this expanded notational freedom can lead to more creative thinking about the production of speech in general, not only of atypical speech.

2.8 Notation for voice quality

Based on the categorization of voice quality (long-term and usually habitual vocal tract settings) found in Laver (1980), a set of symbolization to mark changes in voice quality has evolved. The Voice Quality Symbol "VoQS" system (Ball et al., 2000) is intended as a hypernotational scheme for marking long pieces of segmental transcription. Within the square-bracket demarcation boundaries for phonetic transcription [ ] or the slashes for phonemic/phonological transcription / /, the VoQS notation of the quality to be applied to or superimposed on an entire string of symbols is placed between curly brackets { }. Thus, an extremely nasal string of speech would be isolated between {3Ṽ . . . and . . . 3Ṽ}, and a stretch of speech with slightly raised larynx voice and harsh phonation would be contained between {1L̝V! . . . and . . . 1L̝V!}, where the numbers represent scalar degrees. Various combinations of diacritics can be added to the global voice symbol V (standing for "voice" in this case rather than for "vowel," which is its common usage in phonological notation) to convey multiple meanings economically. For example, a stretch of consonants and vowels with uniformly superimposed moderate nasalization, lingual retroflexion, and whispery voice phonation could be represented between the brackets: {2Ṿ̃˞ . . . 2Ṿ̃˞}. Not all combinations of traits are so amenable, however, to symbolic grouping on a single carrier letter; and ExtIPA designations and their scale may have to be symbolized on successive V symbols or on multiple carrier symbols such as J for jaw position, as in open jaw voice {J̞ . . . J̞}.

3 Elaborating the IPA System of Notation

An elaborated consonant chart is presented in Figure 18.6 to serve as a reference and tracking device for consonantal articulatory possibilities. Some of the issues discussed in section 2 are incorporated into this chart, which is based on the 2005 IPA chart as well as on some aspects of the Extended IPA chart. The elaborated chart (1) incorporates an elaboration of place-of-articulation columns implied in the 2005 chart, supplemented by additions from the ExtIPA system, (2) illustrates the combination into a single category of place of articulation in the pharyngeal area as required by the laryngeal articulator model, (3) elaborates the range of manners of articulation to account for a full set of possible categories, (4) demonstrates graphically the reasons why the 2005 IPA chart is more succinct and economical than this more extensive version, and (5) highlights areas of current conflict in articulatory description that have yet to be resolved.

Pharyngeals and epiglottals are represented together and combined in the same column. This approach recognizes that there exists a plosive category within the pharyngeal articulator, [ʡ]. The sound termed the "Epiglottal plosive" is therefore seen as the culmination of active closure by the aryepiglottic pharyngeal stricture mechanism. The so-called "Epiglottal fricatives" are represented as pharyngeal trills, [ʜ ʢ], since the place of articulation is identical to [ħ ʕ], but trilling of the aryepiglottic folds is more likely to occur in tighter settings of the laryngeal constrictor or with more forceful airflow. The same "epiglottal" symbols could represent pharyngeal fricatives that have a higher larynx position than [ħ ʕ], but a higher larynx position is also more likely to induce trilling than in a pharyngeal fricative with a lowered larynx position. Because [ʜ ʢ] and [ħ ʕ] occur at the same Pharyngeal/Epiglottal place of articulation (Esling, 1999), the logical phonetic distinction to make between them is in manner of articulation, trill versus fricative. This new classification alters their value from the 2005 IPA chart. Following arguments that frication is difficult to produce or to distinguish when glottal voicing and constriction in the pharynx are so close to each other (Laufer, 1996), but also recognizing a tradition of practical phonological necessity, [ʕ] is retained in the fricative row but also added to the approximant row with the explicit diacritic specification [ʕ̞]. Although not a recognized phonemic distinction, the equivalent affricates are also included in an added row in the elaborated chart; that is, [ʡħ ʡʕ] have not been described as distinctive sounds, but their phonetic

CONSONANTS

	Bilabial	Labiodental	Dentolabial	Linguolabial	Dental	Alveolar	Postalveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal/ Epiglottal	Glottal
Plosive	pʰ p b	p̪ b̪	p͆ b͆	t̼ d̼	t̪ʰ t̪ d̪	tʰ t d	t̠ʰ t̠ d̠	ʈʰ ʈ ɖ	ȶʰ ȶ ȡ	cʰ c ɟ	kʰ k ɡ	qʰ q ɢ	ʡ	ʔ
Nasal	m̥ m	ɱ̊ ɱ	m̥͆ m͆	n̼̊ n̼	n̪̊ n̪	n̊ n	n̠̊ n̠	ɳ̊ ɳ	ȵ̊ ȵ	ɲ̊ ɲ	ŋ̊ ŋ	ɴ̥ ɴ		
Trill	ʙ̥ ʙ			r̼̊ r̼	r̪̊ r̪	r̥ r	r̠̊ r̠					ʀ̥ ʀ	ʜ ʢ	
Tap or Flap	ⱱ̟	ⱱ				ɾ̥ ɾ ɺ		ɽ̊ ɽ				ʀ̆	ʢ̆	
Fricative	ɸ β	f v	f͆ v͆	θ̼ ð̼	s̪ z̪ θ ð	s z	s̠ z̠ ʃ ʒ	ʂ ʐ	ɕ ʑ	ç ʝ	x ɣ ɧ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ̼ ɮ̼	ɬ̪ ɮ̪	ɬ ɮ	ɬ̠ ɮ̠							
Approximant	β̞ ʍ ɥ w	ʋ		ɹ̼	ð̞	ɹ̥ ɹ		ɻ̊ ɻ		ʝ̞ j ɥ	ɣ̞ ʍ ɰ w	ʁ̞	ʕ̞	
Lateral approximant				l̼	l̪	l	l̠	ɭ		ʎ	ʟ			
Affricate	pɸ bβ	p̪f b̪v	p͆f͆ b͆v͆	t̼θ̼ d̼ð̼	t̪s̪ d̪z̪ t̪θ d̪ð	ts dz	t̠s̠ d̠z̠ tʃ dʒ	ʈʂ ɖʐ	ȶɕ ȡʑ	cç ɟʝ	kx ɡɣ	qχ ɢʁ	ʡħ ʡʕ	
Lateral affricate						tɬ dɮ								
Implosive	ɓ̥ ɓ					ɗ̥ ɗ		ᶑ̥ ᶑ		ʄ̊ ʄ	ɠ̊ ɠ	ʛ̥ ʛ		
Ejective	p’	f’			θ’ t̪θ’	t’ s’ r’ ts’ tɬ’	ʃ’ tʃ’	ʈ’ ʂ’ ʈʂ’	ȶɕ’	c’ ç’ cç’	k’ x’ kx’	q’ χ’ qχ’		
Click	ʘ				ǀ	ǃ ǁ				ǂ				

Figure 18.6 An elaborated phonetic chart of consonants based on the 2005 IPA chart.

sequencing is clear and easily transcribable. They account for the gliding portion at the release of full aryepiglotto-epiglottic closure (in which the active aryepiglottic folds articulate against the passive epiglottis). The nonpulmonic rows of the Pharyngeal/Epiglottal column have been left empty, although they are not judged impossible. The Pharyngeal/Epiglottal and Glottal places of articulation are separated by dark lines into a laryngeal zone, consistent with the laryngeal articulator model presented in Figure 18.4. Significantly, many articulations that are possible at other vocal tract locations are shaded as impossible within the laryngeal articulator zone. Nevertheless, more categories are recognized and included here than in the official IPA chart.

It is worth pointing out that it is not accidental in the 1993/1996/2005 IPA charts that the left edge of the vowel chart lies directly beneath the Palatal place of articulation. If it were spatially possible, the right edge should lie beneath the Velar column, as in the 1926 chart. Thus, the front vowels [i y] are related notationally to the Palatal consonants, and the raised (up and back) vowels [ɯ u] are vocoid versions of the articulations found in the Velar column. This diagrammatic relationship is more explicit in the 1926 chart (Figure 18.1), where both columns are wide enough to show that the lingual positioning for [i] is related to [j] and that the lingual positioning for [u] is related to [w]. Following the logic of the vowel groupings in Figure 18.5, vowels in the open front corner [a ɶ] are also reflections of Palatal tongue positioning, but also of jaw opening – hence the labels "close" and "open"; while the retracted vowels [ɑ ɒ] should be located beneath the Pharyngeal/Epiglottal column. To represent this relationship graphically might necessitate turning the Pharyngeal/Epiglottal and Glottal columns on their sides and placing them horizontally beneath the columns from Bilabial to Uvular. Such a solution is not elegant, however, as most vowels remain oral, and only two columns of consonantal symbols and one corner of the vowel space are related to the laryngeal articulator. Despite the notational challenge, the conceptual auditory relationship between the retracted vowels and consonants produced in the laryngeal articulator region remains clear.

The predominance of front lingual articulations is apparent from the elaborated chart. Lines divide the chart into four sections, isolating *labial* from *front lingual* from *back lingual* from *laryngeal* articulations. Dentolabial has been borrowed from the ExtIPA chart to complement Labiodental, although it is more common in clinical contexts and is perhaps more useful for instructive phonetic purposes than for describing sounds that occur in natural language. Such a statement is always difficult, since not all of the sounds that occur in language today have been described yet; and some may never be described before their languages become extinct. It is nevertheless instructive to demonstrate that stops at the Dentolabial place of articulation (and even at the Labiodental place of articulation) are more difficult to hold than stops at more common places of articulation. Sounds from Linguolabial to Retroflex in the *front lingual* section are articulated with the dexterity (if that is an accurate analogy) of the pointed part of the tongue. As noted above, no cells in this region have been shaded as impossible, although some may remain unlikely. Those cells that are not filled in the 2005 chart are given explicit symbol combinations here. Most of the Linguolabial sounds will not be encountered in world languages, but it is instructive to show that the available diacritic can be applied to a variety of sound types. There is some question as to whether Linguolabial stops should be written with a "t" or a "p" if the lips are also closed. Such borderline issues make for good theoretical debate. A full Alveolo-palatal column has been added at the beginning of the *back lingual* section, where sounds are articulated with the tongue tip down. The stop symbols [ȶ ȡ] are less familiar than their fricative counterparts [ɕ ʑ] in IPA usage, but many descriptions particularly of Asian languages use them quite routinely. It remains a challenge to phonetic description to distinguish accurately between sounds that are produced with the tongue tip down but with stricture by the front portion of the tongue (as is the case with Alveolo-palatals) and sounds that are produced with the blade of the tongue (laminally) as opposed to with the tip (apically). Many common *front lingual* stops, especially in Asian languages, have been found to be "Apicodental-laminoalveolar" stops rather than simply Dental or Alveolar (Harris, 2006). Such detailed distinctions are not yet adequately represented on the elaborated chart.

Some other common articulatory possibilities are also not shown on the elaborated chart. Aspirated plosives, such as [pʰ] or [tʰ], do appear explicitly, but breathy-voiced stops such as [bʱ] or [dʱ] do not, and not all types of secondary articulations can be shown, e.g., the numerous secondary effects that can qualify clicks. As is the case with the 2005 IPA chart, many articulatory possibilities remain implied because of space restrictions as well as for descriptive economy. It is also implicit that defining a diacritic is taken as adequate to explain how it combines with another character, even though each diacritic is restricted to a limited set of symbols; for example, aspiration only follows voiceless and generally plosive symbols. Voiced plosives can be accompanied by breathy voice, which can either use the breathy-voiced symbol as a diacritic [dʱ] (implying a sequence) or the breathy-voiced fricative diacritic [d̤] (implying simultaneity). Such decisions of usage are also subject to phonological considerations and descriptive tradition. Although there are only two phonation types with explicit diacritic characters specified in the IPA chart, breathy voiced [̤] and creaky voiced [̰], diacritics can be borrowed from VoQS notation, and sequences can be qualified using VoQS bracketing. It is significant that breathy voiced and creaky voiced (laryngealized) notations represent phonatory opposites, the former being open or nonconstricted, and the latter being laryngeally constricted. Thus, the laryngealized diacritic [̰] could also be used to specify varieties of harsh voice.

The Nasal manner of articulation is the second row in the chart because the oral vocal tract is stopped (as for a plosive) during nasal airflow in their production. In the elaborated chart, voiceless nasals are included since their occurrence is not uncommon. The voiceless diacritic is the only diacritic that can be used either above [̊] or below [̥] another symbol without changing its meaning. Trills are also specified for voicelessness, as are two Taps, but audibility can be an issue, so the places of articulation where they can occur are restricted. A Labiodental flap symbol was added to the IPA chart in 2005. Its shape implies that the lips and the teeth are the articulators by default. An attested bilabial version of the flap (Anonby, 2006) is therefore marked with an advancing diacritic to indicate that the two lips are involved. The Fricative row has been split in two, primarily in order to indicate how the front of the tongue is used in two different ways. Both sibilant, grooved [s̪ z̪] sounds, which could also be written [s̟ z̟], and the traditionally "interdental" [θ ð] sounds occupy the Dental place of articulation. Similarly, both [s̠ z̠] and the more usual [ʃ ʒ] sounds are Postalveolar, but with differing grooving characteristics. And a sound with a double articulation, purportedly Postalveolar and Velar [ɧ], but which may also have a distinct rounding component when it occurs in Swedish, is given a place on the chart.

Approximants are also split into two levels – the controversially intermediate category of "lowered fricatives" and the more commonly used and sometimes doubly articulated approximants (with two points of oral stricture). To specify a phonetic difference between [ʝ j̜̞ j] is a very fine line, but it is argued to have meaningful distinctiveness in Spanish (Martínez-Celdrán, 2004). The double articulations [ʍ ɥ w] appear twice; they have a lip-rounding component and therefore also appear in the Bilabial column. What could be viewed as a technical problem is that the most open consonants [j ɰ] are also implied to have inherently spread lips, in the same way that the vowels [i ɯ] are unrounded, i.e. spread. However, a consonant like [j] may not have much lip spreading in its natural phonological incidence. This is what justifies an intermediate, nonspread (and nonrounded) category [j̜̞ ɣ̞] between the two, in particular to distinguish the sequence [ɣ ɣ̞ ɰ]. Lateral fricatives and approximants also present an interesting case. Unlike the central (or median) fricatives and approximants, where the passage through the vocal tract is medially located, the laterals have central closure, like a stop. It should also be noted that there is no voiceless [l̥] counterpart to [l] specified, since it would be perceived as the lateral fricative [ɬ] because of the replacement of periodic voicing by turbulent noise.

Affricates are elaborated in a similar way to the fricatives. Affricates are only referred to obliquely on the IPA chart. Even on the elaborated chart, it is clear that not all possibilities can be represented. Only homorganic possibilities are specified. Heterorganic affricates and double articulations are not included, just as there is no room to include double-stop categories such as [k͡p ɡ͡b]. Nonpulmonic consonants – implosives, ejectives, and clicks – are given separate rows in the elaborated chart. Both voiceless and voiced implosives are listed, using the voiceless diacritic following the usage in McLaughlin (2005). The ejectives are less numerous than the pulmonic stops and affricates; and the clicks, as mentioned above, are not annotated for secondary articulatory features. A large proportion of symbols represent sounds that involve stopping the airflow through the vocal tract. These sounds, and combinations of them, could produce a larger number of combined characters than it is practical to represent even in a chart intended for instructional, educational purposes.

Several qualifications, therefore, need to be made to the chart as elaborated in Figure 18.6. The [ʕ] symbol can also be placed in the Tap row to indicate a rapidly articulated closure and opening of the aryepiglottic folds. This phenomenon has not been fully explored, but a variety of ’ain /ʕ/ in Iraqi Arabic that is too short to be an epiglottal stop [ʡ] and has too much of a burst to be a fricative or an approximant is a likely candidate. A logical symbol would be the basic [ʕ] with a breve (short) mark at the top. The same would apply for a uvular tap. Several attested and/or logical ejective fricatives appear along with the stops and affricates in the Ejective row. Since some are more common, or likely, than others, there are gaps. The example of [r’] is of course necessarily understood to be voiceless [r̥’], since that is a property of the production of all glottalic egressive sounds. In the last row of the table, each click may have several secondary accompaniments, such as in the series [| |ʰ ᵑ| ᶢ| |ˀ |ᵡ |ᴴ] (Miller, 2007; Miller et al., 2009), but only the basic primary values are represented here. This qualification can be seen to be a general one. Most consonants in most rows can be embellished by the addition of secondary articulatory features such as palatalization [ʲ], uvularization [ʶ], or pharyngealization [ˤ]. Single or “singleton” consonants can also be made “geminate” (double or longer in length), which can be indicated by the IPA length mark [ː]. Such a range of elaborations, however, cannot be represented exhaustively in a single coherent chart.

Constructing an elaborated chart of notational representations such as this is perhaps useful as a theoretical phonetic exploratory device but at the same time may not be ideal as a tool for describing the sounds of the world’s languages. First and foremost, not all of the sounds represented on this elaborated chart are attested in languages of the world; and there are therefore far more symbols than are needed in practical phonetic transcription. At the same time, much of the fine phonetic detail usually represented by diacritic notation is absent. For the purposes of phonetic instruction, however, the elaborated chart broadens the conceptual field of articulatory possibilities while retaining an image of gaps of articulatory unlikelihood or of articulatory impossibility.

REFERENCES

Abercrombie, D. (1953) Phonetic transcriptions. *Le Maître Phonétique*, 100, 32–4.

Abercrombie, D. (1964) *English Phonetic Texts*. New York: Faber & Faber.

Abercrombie, D. (1991) *Fifty Years in Phonetics: Selected Papers*. Edinburgh: Edinburgh University Press.

Anonby, E. J. (2006) Mambay. *Journal of the International Phonetic Association*, 36, 221–33.

Ball, M. J., Esling, J. H., & Dickson, B. C. (2000) The transcription of voice quality. In R. D. Kent & M. J. Ball (eds.), *Voice Quality Measurement* (pp. 49–58). San Diego: Singular Publishing Group.

Bell, A. G. (1872) *Visible Speech as a Means of Communication Articulation to Deaf-Mutes. American Annals of the Deaf and Dumb,* 17, 1–21.

Bell, A. G. (1906) *Lectures upon the Mechanism of Speech*. New York: Funk & Wagnalls.

Bell, A. M. (1867) *Visible Speech: The Science of Universal Alphabetics, or Self-Interpreting Physiological Letters, for the Writing of All Languages in One*

Alphabet, inaugural edn. London: Simpkin, Marshall & Co.

Catford, J. C. (1968) The articulatory possibilities of man. In B. Malmberg (ed.), *Manual of Phonetics* (pp. 309–33). Amsterdam: North-Holland.

Catford, J. C. (1977) *Fundamental Problems in Phonetics.* Edinburgh: Edinburgh University Press.

Catford, J. C. (1981) Observations on the recent history of vowel classification. In R. Asher & E. Henderson (eds.), *Towards a History of Phonetics* (pp. 19–32). Edinburgh: Edinburgh University Press.

Catford, J. C. & Esling, J. H. (2006) Articulatory phonetics. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn., vol. 9 (pp. 425–42). Oxford: Elsevier.

Czaykowska-Higgins, E. (1993) Cyclicity and stress in Moses-Columbia Salish (Nxa'amxcín). *Natural Language and Linguistic Theory*, 11, 197–278.

Daniels, P. T. & Bright, W. (eds.) (1996) *The World's Writing Systems*. Oxford: Oxford University Press.

de Wolf, G. D., Gregg, R. J., Harris, B. P., & Scargill, M. H. (eds.) (1998) *Gage Canadian Dictionary*. Toronto: Gage Educational Publishing Company.

Duckworth, M., Allen, G., Hardcastle, W. J., & Ball, M. J. (1990) Extensions to the International Phonetic Alphabet for the transcription of atypical speech. *Clinical Linguistics and Phonetics*, 4, 273–80.

Edmondson, J. A. & Esling, J. H. (2006) The valves of the throat and their functioning in tone, vocal register and stress: Laryngoscopic case studies. *Phonology*, 23, 157–91.

Esling, J. H. (1996) Pharyngeal consonants and the aryepiglottic sphincter. *Journal of the International Phonetic Association*, 26, 65–88.

Esling, J. H. (1999) The IPA categories "pharyngeal" and "epiglottal": Laryngoscopic observations of the pharyngeal articulations and larynx height. *Language and Speech,* 42, 349–72.

Esling, J. H. (2005) There are no back vowels: The laryngeal articulator model. *Canadian Journal of Linguistics*, 50, 13–44.

Esling, J. H. & Edmondson, J. A. (2002) The laryngeal sphincter as an articulator: Tenseness, tongue root and phonation in Yi and Bai. In A. Braun & H. R. Masthoff (eds.), *Phonetics and Its Applications: Festschrift for Jens-Peter Köster on the Occasion of his 60th Birthday* (pp. 38–51). Stuttgart: Franz Steiner.

Esling, J. H., Fraser, K. E., & Harris, J. G. (2005) Glottal stop, glottalized resonants, and pharyngeals: A reinterpretation with evidence from a laryngoscopic study of Nuuchahnulth (Nootka). *Journal of Phonetics*, 33, 383–410.

Hardcastle, W. J. (1976) *Physiology of Speech Production*. London: Academic Press.

Harris, J. G. (2006) A palatographic study of places of articulation in Thai and some other Southeast Asian languages: Dentals, alveolars, and palatals. In J. E. Harris (ed.), *Readings in Articulatory Phonetics*, vol. 1: *Consonants and Phonation Types* (pp. 63–91). Bangkok: Ek Phim Thai Co.

IPA (1926) IPA chart. *Le Maître Phonétique*, Troisième Série, No. 16, octobre–décembre 1926.

IPA (1949) *The Principles of the International Phonetic Association: Being a Description of the International Phonetic Alphabet and the Manner of Using It, Illustrated by Texts in 51 Languages.* London: International Phonetic Association.

IPA (1989a) Report on the 1989 Kiel Convention. *Journal of the International Phonetic Association*, 19, 67–80.

IPA (1989b) The IPA 1989 Kiel convention Workgroup 9 report: Computer coding of IPA symbols and computer representation of individual languages. *Journal of the International Phonetic Association*, 19, 81–2.

IPA (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Kemp, J. A. (2006) Phonetic transcription: History. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn., vol. 9 (pp. 396–410). Oxford: Elsevier.

Kinkade, M. D. (1967) Uvular-pharyngeal resonants in Interior Salish. *International Journal of American Linguistics*, 33, 228–34.

Ladefoged, P. (1967) *Three Areas of Experimental Phonetics.* Oxford: Oxford University Press.

Laufer, A. (1996) The common [ʕ] is an approximant and not a fricative. *Journal of the International Phonetic Association*, 26, 113–17.

Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

MacMahon, M. K. C. (1996) Phonetic notation. In P. T. Daniels & W. Bright (eds.), *The World's Writing Systems* (pp. 821–46). Oxford: Oxford University Press.

Martínez-Celdrán, E. (2004) Problems in the classification of approximants. *Journal of the International Phonetic Association*, 34, 201–10.

McLaughlin, F. (2005) Voiceless implosives in Seereer-Siin. *Journal of the International Phonetic Association*, 35, 201–14.

Miller, A. L. (2007) Guttural vowels and guttural co-articulation in Ju|'hoansi. *Journal of Phonetics*, 35, 56–84.

Miller, A. L., Brugman, J., Sands, B., Namaseb, L., Exter, M., & Collins, C. (2009) Differences in airstream and posterior place of articulation among N|uu clicks. *Journal of the International Phonetic Association*, 39, 129–61.

Padayodi, C. M. (2008) Kabiye. *Journal of the International Phonetic Association*, 38, 215–21.

Payne, E. M. (2005) Phonetic variation in Italian consonant gemination. *Journal of the International Phonetic Association*, 35, 153–81.

Rey, A. & Rey-Debove, J. (eds.) (1988) *Le petit Robert 1: dictionnaire de la langue française*. Paris: Dictionnaires Le Robert.

Roach, P. & Hartman, J. (eds.) (1997) *English Pronouncing Dictionary*, 15th edn. Cambridge: Cambridge University Press.

Shaw, P. A., Blake, S. J., Campbell, J., & Shephard, C. (1999) Stress in hən'q'əmin'əm' (Musqueam) Salish. In M. Caldecott, S. Gessner, & E.-S. Kim (eds.), *University of British Columbia Working Papers in Linguistics, Proceedings of the Workshop on Structure and Constituency in Languages of the Americas*, 2, 131–63.

Sweet, H. (1877) *A Handbook of Phonetics.* Oxford: Clarendon Press.

Sweet, H. (1880–1) Sound notation. *Transactions of the Philological Society*, 177–235.

Sweet, H. (1971) *The Indispensable Foundation: A Selection from the Writings of Henry Sweet*, ed. E. J. A. Henderson. Oxford: London University Press.

Text Encoding Initiative (2002) *The TEI Consortium: Guidelines for Electronic Text Encoding and Interchange.* Charlottesville, VA: University of Virginia Press.

Thelwall, R. & Sa'adeddin, M. A. (1999) Arabic. In IPA, 51–54.

Trager, G. L. & Smith, H. L. (1951) *An Outline of English Structure* (Studies in Linguistics, Occasional papers 3). Norman, OK: Battenberg Press.

Unicode Consortium (2007) The Unicode Standard, Version 5.0.0, defined by: *The Unicode Standard, Version 5.0*. Boston: Addison-Wesley.

Wells, J. C. (1976) The Association's alphabet. *Journal of the International Phonetic Association*, 6, 2–3.

Wells, J. C. (2006) Phonetic transcription and analysis. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn., vol. 9 (pp. 386–96). Oxford: Elsevier.

19 Sociophonetics

PAUL FOULKES, JAMES M. SCOBBIE, AND DOMINIC WATT

1 Introduction

In general terms, sociophonetics involves the integration of the principles, techniques, and theoretical frameworks of phonetics with those of sociolinguistics. However, there has been considerable variation both in the usage of the term and the definition of the field, so that sociophonetic research may orient more towards the concerns of sociolinguists on the one hand or phoneticians on the other.

The first recorded use of the term "socio-phonetic" (*sic*) is by Deshaies-Lafontaine (1974), a dissertation on variation in Canadian French carried out squarely within the emergent field of Labovian or variationist sociolinguistics. The term was coined as a parallel to "sociolinguistic" in order to capture the project's emphasis on phonetic rather than syntactic or lexical variables (Deshaies, p.c.).[1] Among phoneticians, sociophonetics has been used as a thematic label at the quadrennial International Congress of Phonetic Sciences (ICPhS) since 1979. The nine papers presented under this heading at the 1979 conference followed the pattern set by Deshaies-Lafontaine in addressing central questions in sociolinguistics with reference to phonetic variables. Contributions included papers by eminent sociolinguists such as Labov (on vowel normalization) and Romaine (on variation and change in Scottish /r/). Probably the first example of explicitly sociophonetic work published in a prominent journal is the variationist study of Viennese German by Dressler and Wodak (1982, although they prefer the epithet "sociophonological," with synonymous intent).

Since these early studies the quantity of research which can be described as sociophonetic has increased rapidly, particularly since the mid 1990s, and the scope of that research has become ever broader. To support this observation it is instructive to survey the set of papers presented at the 2003 ICPhS. Around 90 papers were presented on topics with a sociolinguistic connection, such as variation, change, socially informed fieldwork, and speaking style. However, discussions also included subjects as diverse as the phonological relationship between liquids, descriptive accounts of Albanian and Cocos Malay, loanword pronunciation, conversation analysis, methods for developing large corpora, and psycholinguistic experiments on information processing.

Indeed, given the recent growth of the field and the disparate paths it has taken, providing an adequate definition of sociophonetics is far from straightforward. The discipline draws upon a rich empirical corpus which is generated through a wide set of methods and which is exploited to address a diverse range of theoretical questions. Circumscription of the field is similarly problematic. The boundaries of the discipline have become increasingly porous, such that sociophonetic research now amalgamates theories and methods not only from phonetics and sociolinguistics but also from related fields including psycholinguistics, clinical linguistics, first language (L1) and second language (L2) acquisition, theoretical phonology, and computational linguistics. At the same time, though, the eclecticism of sociophonetics may be misinterpreted as indicating a lack of clear focus. Its development through the marriage of traditionally separate disciplines has also led, perhaps, to its being viewed as a relatively peripheral concern to its contributory fields. However, whilst it remains something of a loose confederation of industries, the inclusion of this chapter in the second edition of this handbook bears witness to the claims of sociophonetics to be an independent discipline. With this in mind our goal in this chapter is both to survey the current state of the field and in doing so also attempt to delimit and define it. We will point to a number of important contemporary theoretical issues in phonetics more broadly where we think progress can best be made through work that draws on sociophonetic concepts and methods.

In our view, the unifying theme of sociophonetic work is the aim of identifying, and ultimately explaining, the sources, loci, parameters, and communicative functions of socially structured variation in speech. In this view the goals of sociophonetics include accounting for how socially structured variation in the sound system is learned, stored cognitively, subjectively evaluated, and processed in speaking and listening. Such work contributes to the development of theoretical models in phonetics and sociolinguistics, spanning speech production and perception, with a clear focus on the origin and spread of change. Sociophonetic methods and data also contribute to theoretical models in phonology, acquisition, and long-term storage of linguistic knowledge, because of the field's focus on fine phonetic detail, and structured variation.

Methodologically, socially structured variation offers great opportunities for experimental phoneticians to exploit, because micro-typological studies fall neatly between cross-linguistic and idiolexical comparisons. The fine granularity of the differences between related socially located linguistic systems provides an invaluable research tool, albeit one which has to date largely been exploited in cross-dialectal research defined geographically rather than socially. Phonetic research often draws on homogeneous pools of subjects to suppress variation, but intersubject differences in fine phonetic detail which function socially can be used in order to understand both the variation and the aspects which are common across closely related systems, by using structured pools of subjects.

We continue this overview with a brief explanation of socially structured variation, identifying the kinds of phenomena that are investigated in sociophonetic studies. We then survey representative studies, first those with a focus on speech production, followed by a discussion of sociophonetically informed studies of speech perception. We highlight a number of important methodological issues and the principal theoretical contributions made by sociophonetic work. Finally, we address the potential contributions of sociophonetics to applied fields.[2]

2 Defining Sociophonetic Variation

Speech provides various kinds of information simultaneously, a fact recognized by the earliest scholars of language (see for example Chambers' 2002 discussion of the Sanskrit grammarian Pāṇ̥ini, writing in c. 600 BC). In modern linguistics it has become customary to draw a binary division between linguistic and extra-linguistic, or indexical, information.

Linguistic information is that conveyed in respect of propositional meaning – what we might loosely characterize as assertions about the world that can be abstracted from their contexts of utterance. We can generally identify multiple utterances as linguistically equivalent even if they are pronounced in different ways. We do this by selectively attending to the linguistic information carried by the signal, as we do when, for instance, writing down utterances using a standardized orthography.

As well as permitting the communication of semantic information, however, speech also provides a rich source of indexical information reflecting, for example, a speaker's background, pragmatic intent, and emotional state (Abercrombie, 1967).[3] Some sources of indexical information emerge as the product of the universal constraints of biology and physics. Organic factors, such as vocal tract anatomy and physiology, exert a clear influence on vowel formant values and fundamental frequency such that relatively abrupt differences can be observed between adult males, females, and children. Similar effects underlie developmental aspects of first language acquisition and physical changes through the life course, most noticeably in old age. Other indexical features are purely a social product, that is, arbitrary associations between linguistic variants and types of speaker or speech. To illustrate, a much cited example is that of coda (postvocalic) /r/ in English (e.g., Scobbie, 2006b). Studies of so-called rhotic varieties (e.g., in North America, Ireland, and Scotland) have found statistical differences in the production of coda /r/ across social classes (e.g., Labov, 1972; Reid, 1978; Romaine, 1978; Stuart-Smith, 2007a). In these varieties, members of higher socio-economic groups typically use audible consonantal rhoticity more than those of lower social groups. The rhotic nature of coda /r/[4] can therefore be said to index social class. The arbitrariness of the association between rhoticity and class is illustrated when we compare the same phonetic quality in non-rhotic dialects (e.g., in England), where the opposite social evaluation of coda /r/ can be observed: a rhotic pronunciation of the final consonant in a word like *car* when it is a coda is often taken as a sign of low rather than high social status (Wells, 1982, p. 35).

The example of English coda /r/ illustrates both the importance of social factors in directing change and, at this level of description, its phonetic arbitrariness. In both rhotic and non-rhotic communities variation encompasses phonetic weakening and vocalization of coda /r/, a commonality which phonetic and phonological theory attempt to explain. But considering the phenomenon in isolation from its social context, through study of articulatory, acoustic, or perceptual data, is likely to be less revealing than research which incorporates extra dimensions from sociophonetics which explore the functionality and indexical meanings of the variants and their role in the establishment of structured variation and the transmission of change.

The indexical functions of a linguistic variable are usually manifested in statistical differences in a form's distribution across speakers, groups, or speech styles, rather than resulting from categorical usage or non-usage of a particular variant. This is certainly true in the case of studies of English /r/, for instance. However, examples of the latter can be found: Macaulay (1991) found that use of the monophthong [ʉ] rather than [ʌʉ] in words like *house* and *down* in Scottish English was frequent in the speech of his working-class informants, but was altogether absent among his middle-class speakers. Discrete phonological differences have also been reported for males and females in other languages, including Gros Ventre (Taylor, 1982) and Chukchi (Dunn, 2000).

A central question in the study of indexical features is how universal and social constraints interact. Johnson (2006), for example, compares vowel formant spaces for males and females in 17 languages and dialects. The magnitude of differences in F1/F2 distribution varies substantially across the languages, indicating that biology alone cannot account for the variation (see also Rosner & Pickering, 1994, pp. 49–73). Furthermore, there is evidence from studies of children which shows differences in average fundamental frequency, formant values, and voice quality between prepubescent boys and girls. These findings suggest that some gender-appropriate speech behavior must be learned in childhood rather than being determined solely by anatomical differences between the sexes (Sachs, 1975; Lee et al., 1995; Sederholm, 1998; Whiteside, 2001). It can therefore be difficult, even impossible, to disentangle socially-influenced variation from variation which is the product of biology and physics. In light of the difficulties presented to the analyst by this convolution of sources of variability, we recognize as sociophonetic any aspect of systematic phonetic variation in which the indexed factor is at least *in part* the product of social construction (following Foulkes & Docherty, 2006, p. 412).

3 Sociophonetic Studies of Speech Production

3.1 Sources of variation

The majority of work carried out under the rubric of sociophonetics has focused on identifying the indexical roles of features of speech production. It has been established that speech varies in systematic ways as a function of a very wide range of social factors. This has been one of the fundamental contributions of Labovian sociolinguistics more generally, and is reflected in the early sociophonetic works referred to in the introduction. One of Labov's principal motivations at the inception of his highly influential studies of English in New York City was to show, contra earlier works on American urban varieties such as Hubbell (1950), that linguistic variation is not random, inexplicable, or theoretically irrelevant. This manifesto, laid out explicitly in Labov (1966a), informs all subsequent work carried out under the variationist sociolinguistics banner.

Early sociolinguistic work focused on sources of variation identified as correlating with broad demographic categories such as social class, age, speaker sex, and ethnicity, and with speaking style (e.g., Labov, 1966b; Labov et al., 1972). Figure 19.1 represents a typical example, illustrating quantified data for (ng) from Norwich English (Trudgill, 1974). The sociolinguistic variable (ng) refers to the final nasal in gerunds (*walking*, etc.) as well as in polysyllabic monomorphemes such as *ceiling* and in the words *anything*, *everything*, *nothing*, and *something*. In all varieties of English so far studied, speakers alternate between using the standard velar nasal [ŋ] and the nonstandard alveolar [n]. However, the data in Figure 19.1 show that variant selection is not random. Instead, it displays *structured heterogeneity* (Weinreich et al., 1968), correlating with both the sex and social class of the speaker.

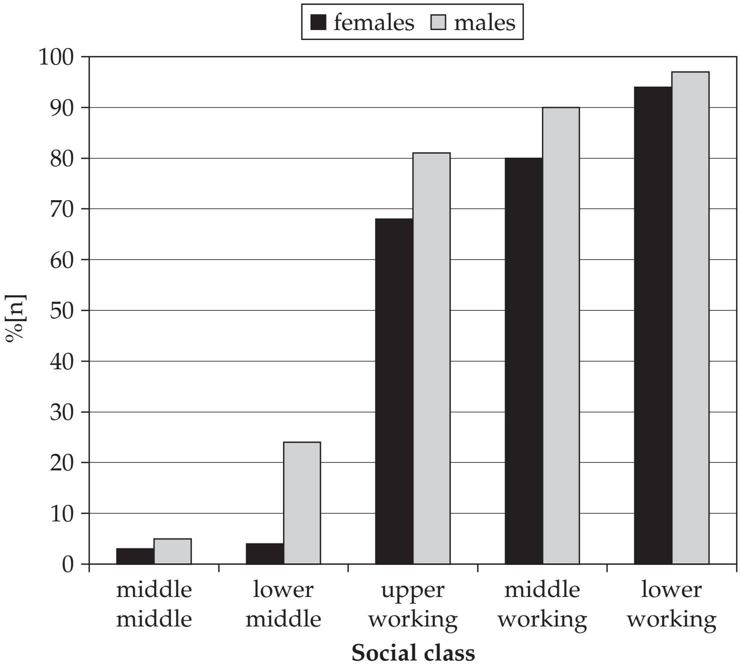


Figure 19.1 Usage of non-standard variant [n] for (ng) in Norwich English. (Adapted from Trudgill, 1974, p. 94)

Use of [n] is lowest for the middle-class groups, and rises across the social continuum, with an abrupt leap for the upper working class relative to the lower middle class. In all social groups the men use more [n] than the women. Use of [n] in Norwich is therefore an index of speaker gender and especially of social class.

Figure 19.1 is a typical example of early sociolinguistic research in that it shows that vernacular or nonstandard forms tend to be more frequent for males and for members of lower socio-economic groups. However, plenty of counter-examples have since been reported which reflect particularly marked social differences in the community under investigation. One such example is Mees and Collins' (1999) analysis of the distribution of glottalization features in Cardiff English, in which, contrary to the pattern found in other parts of the UK, glottal pronunciations of preconsonantal and prepausal /t/ are apparently evaluated locally as prestigious. In another example, Watt and Milroy (1999) discuss variation in Newcastle English in the vowels of the FACE, GOAT, and NURSE lexical sets (following the notation system devised by Wells, 1982). The patterns of variation do not align on a standard–nonstandard continuum. Instead, younger females appear to be leading a change away from the traditional local vernacular forms [fɪəs], [gʊət], and [nɔːs] towards other nonstandard variants that have wider currency across the north of England: [feːs], [goːt] and [nøːs].

Much sociolinguistic work since the 1960s, informed by theoretical advances from adjacent fields such as sociology and anthropology, has been devoted to refining our understanding of the relevant social sources of variation. One important result of such work has been a move beyond broad demographic categories in both methodology and theorizing. We offer below a brief review of some important advances in our understanding of the main sources of learned variability.

3.1.1 Social class, communities, and networks Ash (2002) discusses problems in defining and measuring social class, and assesses the value of class for studies in non-Western societies (see for example Haeri's (2003) critique of conventional notions of class, gender, and speech style among Cairene Arabic speakers). Yet social/stylistic stratification of some kind, e.g., diglossia (Ferguson, 1959), can occur in many different societies, including those where Western analytic categories may not straightforwardly apply. And even in Western societies, concepts such as social class require deconstruction in order to identify more precisely the root causes and functions of linguistic variation. For example, studies in the *communities of practice* framework (Eckert & McConnell-Ginet, 1999; Meyerhoff, 2002) show how linguistic forms are affected by an individual's chosen membership of groups. Groups may be defined at a local level (such as belonging to a particular sports team or professional organization), but they may also reflect more widespread and abstract patterns of social practice (exemplified by adolescents and their choices in behaviors, clothing, slang terminology, and arguably vernacular pronunciation on a nationwide basis).

Mendoza-Denton's work on the language of young Latina gang girls in California, for instance, links sociophonetic and discourse variation to gang members' "expectation-violating" modes of (nonlinguistic) behavior, such as fighting, that transgress what is conventionally considered appropriate for young women (Mendoza-Denton, 1996, 1999). Linguistic choices are viewed as one type of symbolic resource in the construction and maintenance of identities. In a similar vein, Bucholtz (1998, 1999) demonstrates that "nerds" (students who consciously adopt an "uncool" identity) differentiate themselves from their peers through various social practices including phonetic and linguistic choices. Californian nerds, for instance, did not participate in ongoing vowel changes such as fronting of GOAT and GOOSE, used released forms of word-final /t/ instead of the typical unreleased or glottalized forms, and avoided current slang. Other studies on the role of phonetic variants in establishing social identity include Eckert (2000) on socially polarized school groups, Kiesling (1998) on members of American college fraternities and sports teams, and Bunin Benor (2001) on orthodox Jewish women. There is a relatively wide collection of studies on the phonetic properties of gay, lesbian, and bisexual speakers and speech styles, including Pierrehumbert et al. (2004), Munson et al. (2006), and Munson (2007). It is increasingly apparent from studies such as these that individuals exercise a considerable degree of choice – whether conscious or subconscious – over the phonetic forms they use in their speech, within the constraints imposed by intelligibility considerations. These choices can make an essential contribution to the indexing of personal stance, identity, and communicative function. The freedom which speakers have to define, use, change, and move between different identity-based sociolects starkly shows the pitfalls which phoneticians and phonologists risk by failing to take social variation into account when positing functional explanations for patterns that may exist in a standard variety, but in a different form in other varieties.

The individual's degree of entrenchment in a group has also been assessed through studies which take account of the speaker's *social network* (Milroy, 1987b). Generally speaking, the more central the place of the individual in a group, the stronger their adherence to the group's norms of behavior and the greater the normative influence of linguistic forms associated with that group. For example, Labov (1972) investigated several sociolinguistic variables in the speech of the rival Harlem street gangs, the Aces and the Thunderbirds. Members of both gangs used many nonstandard phonological forms, including [n] for (ng) and [d] or [v] for (th) (in *brother*, etc.). The nonstandard usage of core members was relatively similar across the gangs, and considerably higher than that for peripheral associates of either gang. Strength of social networks may also be affected by factors such as differences in geographical *mobility* (Britain, 2002), the extent of *routine* behaviors such as those based around sport and leisure activities (Britain, 1997), and general social structures such as patterns of employment (Milroy, 1987b).

Network strength and structure interact in complex ways with the more conventional demographic categories typically employed in sociolinguistic research. For example, Dubois and Horvath's (1998, 1999) study of variability in dental fricatives among Louisiana Cajun English speakers revealed a resurgent use of the traditional stops [t̪] and [d̪] (for /θ/ and /ð/respectively) by younger informants (20–39 years of age). The informants were distinguished by their membership of networks which were described as either "closed" (enclave or otherwise insular communities) or "open" (in which individuals are more participative in wider society). The revitalization of (th)-stopping, they argue, is best accounted for by attending to speakers' membership of open and closed networks as well as by their gender. Young closed network men and women tended to "recycle" the traditional stop variants, but among open network speakers this habit was only observed among men. (th)-stopping in Cajun English thus has multiple indexical functions which differ depending on the relative openness of the networks in which its speakers are integrated. Dubois and Horvath link the resurgence of the stop variants to a revival of positive associations with the formerly stigmatized Cajun culture and identity.

3.1.2 Age and life stages Biological age is obviously an important contributor to phonetic and linguistic differences through childhood (Vihman, 1996) and again in later life (Beck, this volume). However, age differences may also reflect socially determined divisions of the age continuum, or *life stages*. Eckert (1997) argues that in most Western societies there are three main life stages: childhood, adolescence, and adulthood. Each is defined by major differences in typical lifestyle, which in turn exert radically different influences both on general behavior and specifically language use. The number of stages, their distinct patterns of social behavior, and their effects on language may differ from society to society or change over time. Speakers of particular ages in particular communities may also show marked linguistic differences as a consequence not only of long-term social convention, but also because of major historical or social events. In his study of dialect change in Gaza Arabic, for example, Al Shareef (2002) divided speakers into age groups according to their experiences of mass migration into Gaza following political events in 1948 and 1967. Those who were adults before 1948 were found to maintain their original local dialects, while those born after migration showed effects of contact with other dialects.

In childhood, lifestyle is dominated by the family setting. Children receive the bulk of their linguistic input from the immediate family and they conform broadly to the norms of the input model(s). For example, Foulkes and Docherty (2006) examined the use of pre-aspirated voiceless stops in Newcastle English, finding a close correlation between usage by 2- to 4-year-old children and that of their mothers. Kerswill and Williams (2000) offer a clear illustration of the importance of the home model in their study of children's speech in Milton Keynes, a town which underwent huge expansion in the 1960s to create a commuter residence for London and other cities in the south east of England. As a result the new town experienced very high rates of in-migration from all over the UK and further afield. In Kerswill and Williams' study, 4-year-old children showed great phonological diversity as a group, reflecting the mix of dialects in their homes. Children of rhotic parents, for example, displayed rhoticity themselves.

Input variation has also been the focus of a series of studies on acquisition of Scottish English, including Hewlett et al. (1999) and Scobbie (2005). They examined the acquisition of the complex phonological pattern referred to as the Scottish

Vowel Length Rule (SVLR). Most varieties of English display vowel length differences dependent on the voicing of the following consonant such that the vowels of *brood* and *bruise* are longer than that of *brute*. The effect of the SVLR is that vowels preceding voiced stops are also short, such that in SVLR accents *brood* and *brute* are short, while *bruise* is long. The impact on acquisition of SVLR of different parental models was assessed, considering whether children have two, one, or no Scottish parents. Their findings indicate that the pattern is harder to learn if one or neither parent is Scottish. Scobbie (2005) shows in detail how a "mixed" speaker can incorporate aspects of both parental and community targets to result in new and typologically marked patterns. Sociophonetic studies can therefore be highly instructive for our understanding of phonological acquisition. A great deal can be learnt from how children form their own system (Vihman, 1996) on the basis of different input systems, whether the inputs vary as part of the community's sociolinguistic norms (the standard sociolinguistic focus), or between the family and the community at large, or even within the family.

Through childhood and into adolescence the social role of the peer group begins to take over. Linguistically, the adolescent period is frequently characterized by a shift away from the family model in favor of high usage of nonstandard forms, high usage of forms that are innovative in any ongoing sound change, and homogeneity of usage across the peer group. The Milton Keynes study again illustrates such a shift. Compared with the 4-year-olds, children aged 8 and 12 years showed increasing divergence from their parental models, but increasing focusing of the dialect within the peer group. Minority forms such as rhoticity had disappeared by age 6 (Kerswill & Williams, 2000).

In adulthood, lifestyle may settle again, and the exigencies of career choices may lead to particular language varieties taking on a certain *market value* for the speaker, depending on communicative function. For example, Coupland (1980) discusses the value of both local Cardiff speech and standard varieties for a worker in a travel agency. Pappas (2006) gives an account of stigmatized local pronunciations of /l/ and /n/ in Patra Greek, showing that use of stigmatized variants, and attitudes towards them, correlate with speakers' orientations towards metropolitan versus provincial lifestyles and their associated employment aspirations (see also Brouwer & van Hout, 1992).

3.1.3 Sex and gender Speaker sex has often been investigated in phonetic as well as sociolinguistic studies, perhaps because the biological effects of speaker sex on speech are in many respects obvious and impossible to avoid (for phonetic studies with a particular focus on sex-based differences, see, e.g., Byrd, 1994; Whiteside et al., 2004; Simpson & Ericsdotter, 2007; Jacewicz et al., 2007). For their part, sociolinguists have come to focus more on the socially constructed and "performed" roles of gender rather than the binary category of biological sex (Butler, 1990; Hall & Bucholtz, 1995; Eckert, 2000; Cheshire, 2002). The importance of gender in understanding language use is grounded on the observation that males and females tend to compete with, and evaluate themselves against, members of their own gender (Eckert, 2000, p. 122). To understand variation in speech it is therefore important to explore differences *within* gender groups rather than focus on simple male/female comparisons. The specific question for sociophonetics is how these factors of sex and gender interact. The extent to which biological sex differences might directly underpin apparently arbitrary gender differences with respect to sociolinguistic behavior has received less attention, however. Gordon and Heath (1998) attempt to draw explicit links between biological sex and male/female differences in the extent of participation in ongoing sound changes by pointing to "intrinsic" sex-based preferences in the sound symbolic domain. They argue on the basis of synchronic and historical data from a range of languages that women tend to lead vowel changes towards the close front region of the vowel space, while men are predisposed towards changes involving vowel retraction and rounding.

The importance of gender as opposed to sex is illustrated in studies such as that by McConnell-Ginet (1983), Johnson (2006), and Stuart-Smith (2007b). Stuart-Smith, for example, provides a detailed study of [s] production by 32 English speakers from Glasgow. Anatomical differences between males and females predict acoustic differences in [s], such that the smaller vocal tract typical for a woman would produce an [s] with aperiodic energy at a higher overall frequency than the larger male vocal tract (Stevens, 1998, p. 398). However, in Stuart-Smith's study, the acoustic data from working-class girls patterned with those from males. Figure 19.2 shows the long-term average spectrum (LTAS) for the [s] in the word *ice* as produced by a young working-class female (upper pane) and for a young middle-class female (lower pane). The distribution of energy in the spectrum for the working-class female resembles that typically found for Glaswegian males, with a much lower center of gravity than that found for middle-class women.

Following the reasoning expressed by Eckert (2000) and others adopting a communities of practice framework, Stuart-Smith interprets the girls' phonetic patterns not as an attempt to sound like males, but to distance themselves from middle-class girls and the social identity they present.

3.1.4 Regional variation In addition to the social dimensions of variation we should also comment on studies of regional variation, since speech also indexes a person's geographical identity. Regional studies have a particularly long history, and in fact, from the perspective of our definition of sociophonetics, it is possible to regard the pioneers of nineteenth-century dialectology as the first sociophoneticians (e.g., Wenker, 1895). Their work not only yielded descriptive documentation of geographical variation, it also showed awareness of the social variation within communities through the predominant focus on older rural males as the harbingers of maximally archaic forms, as well as a recognition that traditional dialects were undergoing change through processes such as standardization (Chambers & Trudgill, 1998). Contemporary analyses of regional variation operate with more complex notions of space which acknowledge "distance" between locations as having social and psychological dimensions rather than being defined solely in terms of geographical proximity (Britain, 2002). Such factors may include

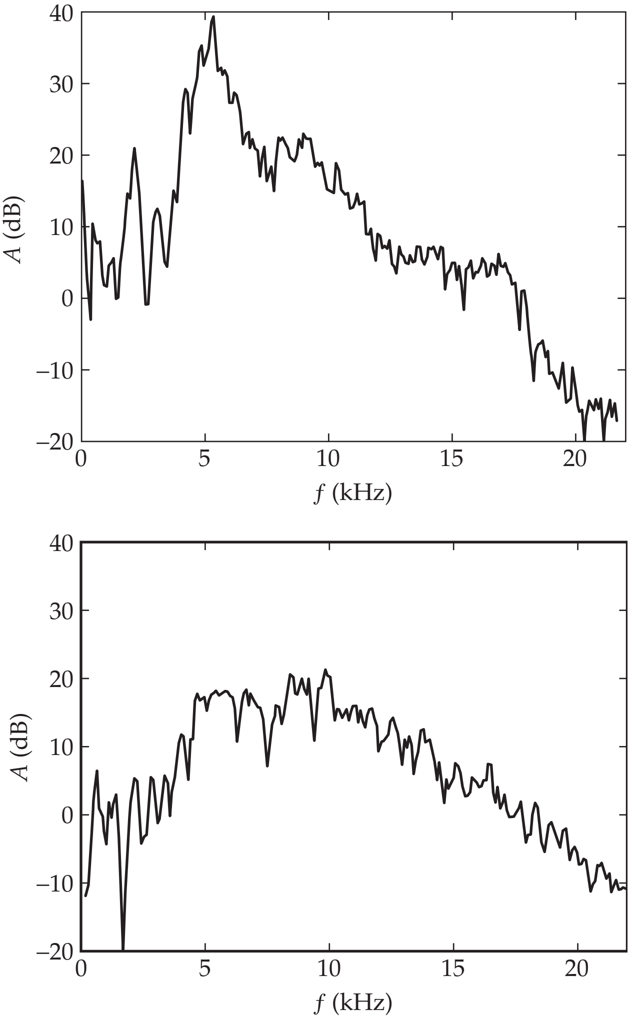


Figure 19.2 Long-term average spectra (LTAS) for [s] in *ice* spoken by a young working-class Glaswegian female (upper panel) and a young middle-class Glaswegian female (lower panel) (see Stuart-Smith, 2007b).

political boundaries and differing orientations towards larger economic centers (e.g., Boberg, 2000; Woolhiser, 2005; Llamas, 2007).

The wider mobility of some groups, implicit and explicit processes of national and supralocal standardization, and people's exposure to and awareness of other regional varieties have been prime areas of interest for sociophoneticians. Advances in telecommunications, recording, and analysis technologies have facilitated the exploration of interaction and interference between a wide range of subtly different phonetic systems. For an overview of recent studies carried out within this framework and the development of increasingly sophisticated theoretical models to which they contribute see Auer et al. (2005).

3.1.5 Ethnicity, race and bilingualism Ethnicity is a social product as opposed to a biological given (Fought, 2002) (like gender as opposed to sex) and indeed can be entirely nonbiological if based on religion or culture. Both ethnic marking in L1 and the role of ethnicity in creating an L2 variety have been examined. The relationship between linguistic variation and ethnicity has been a prominent focus in sociolinguistics since Labov's earliest works, which included investigations of the phonological patterns of the Portuguese and Wampanoag Native American minorities in Martha's Vineyard (Labov, 1963), and Puerto Ricans and African Americans in New York City (Labov et al., 1968). A great canon of work has since been produced on African American English (AAE), describing its current features and also tracing its development (see, e.g., Wolfram, 1969; Mufwene et al., 1998; Green, 2002; Wolfram & Thomas, 2002). Phonological features have been less studied than other aspects of the grammar, and the bulk of work has concentrated on differences between AAE and other varieties with little attention being paid to variation within AAE (Wolfram & Schilling-Estes, 1998, p. 174). It is often assumed, in fact, that AAE varies relatively little geographically (but see, e.g., Hinton & Pollock, 2007, for counter-evidence), and collectively AAE speakers appear to resist participation in major sound changes such as the Northern Cities Shift (e.g., Wolfram & Schilling-Estes, 1998; Milroy & Gordon, 2003).

The English of several other ethnic communities has been studied in North America (e.g., Anderson, 1999; Fought, 1999, 2003; Ryback-Soucy & Nagy, 2000; Schilling-Estes, 2000; Thomas, 2000; Boberg, 2004). However, the role of ethnicity in shaping accent differences has been studied relatively rarely elsewhere, or with respect to languages other than English. Exceptions include Māori English (Britain, 1992; Holmes, 1997), ethnic varieties of English in Australia (Clyne et al., 2001), ethnic marking in Israeli Hebrew (Yaeger-Dror, 1994a), and ethnic groups divided on religious grounds in the UK (Milroy, 1987b; McCafferty, 1999, 2001). Heselwood and McChrystal (2000) present a preliminary study of the accent features of Panjabi-English bilinguals in Bradford, revealing marked gender differences in the use of L2-influenced variants. For example, the males used a greater number of retroflexed variants of /t/ and /d/, a feature characteristic of Panjabi itself. It appears that the males may be adapting phonological features of one language for use as markers of ethnicity in the other. This "recycling" of sociolinguistic features is also reported by Dyer (2002) in her study of the English steel town, Corby. The town saw a large influx of Scottish steel workers in the 1960s. Subsequent generations have abandoned most Scottish phonological features, but some have been retained as markers of local Corby identity and a means by which young Corby speakers differentiate themselves from inhabitants of neighboring areas.

Strategic sociolinguistic choices can also be made by second language learners and bilinguals: van der Haagen (1998), for instance, investigated Dutch learners of English and found correlations between their attitudes to American and British varieties and their pronunciation preferences. Choice of language or variety may contribute to a speaker's social identity. Khattab (2006, 2007, 2009) shows that English–Arabic bilingual children learn and exploit a range of phonetic variants comparable to that of their monolingual peers. Moreover, phonetic interference patterns from English onto Arabic can be interpreted not as imperfect learning (bi-directional interference) but as strategic devices to achieve conversational goals. For example, the children frequently adapted English words to Arabic phonology when attempting to satisfy their parents by speaking Arabic in fieldwork sessions. Scobbie (2005) links such results back to monolingual cross-dialectal variation and the influence of parents with "incomer" accents, while Evans et al. (2007), Lambert et al. (2007), and Hirson and Sohail (2007) focus on the effect of being in a bilingual community. Variability of rhotics in Panjabi-English bilinguals also reflects the growing trend of examining first- and second-generation immigrant variation, further linking topics in acquisition and bilingualism to social and ethnic identity.

3.1.6 Intra-speaker variation The sources of sociophonetic variation discussed so far are all important for the observation of inter-speaker differences within a community. Intra-speaker variation, meanwhile, has been investigated mainly through analysis of speech across several modes of communicative setting which collectively can be labeled *style*. In Labov's early work style was viewed as a continuum defined by degree of self-monitoring. Vernacular usage was hypothesized to occur where speakers paid minimal attention to their speech, while it was predicted that increasing formality (accessed, for example, in tasks such as reading aloud) would lead to increased self-consciousness and self-monitoring (Labov, 1972, p. 208).

Labov's original formulation of style has been acknowledged to be simplistic, however. It is now recognized that speakers adjust phonetic parameters in response not only to their own self-monitoring but also in response to a range of external factors including topic, physical setting, and audience. A more standard variety of pronunciation may be used, for example, where a speaker deems it appropriate for the listener (e.g., because of the formality of the setting, or when the speaker judges that the listener might understand a vernacular form less readily than its standard equivalent). Bell (1984) refers to this effect as *audience design*. Lindblom (1986, 1990) conceptualizes within-speaker variation in a similar way in the H&H (hyper- and hypo-speech) framework. Speakers position themselves along a hyper–hypo continuum according to the communicative setting, trading off the demands of limiting articulatory effort against those of ensuring intelligibility for listeners. The hyper end of the continuum is characterized by relatively canonical speech, with shifts to the hypo end permitting incrementally more underarticulation. The hyper–hypo range may align with a standard–vernacular continuum, but need not necessarily do so. Wassink et al. (2007) compare the features of hyper-speech with those of Lombard and child-directed speech. Work in conversation analysis has also shown that fine-grained phonetic cues may be manipulated by participants to manage interaction, for example to delimit speaking turns, highlight repetitions, time interruptions, and indicate (dis)agreement (Local, 2003, 2007; Ogden, 2004). Interactants have been shown to monitor and systematically adapt details of consonant and vowel pronunciation, loudness, rhythm, timing, intonation, and pitch in their negotiation of conversation.

Intra-speaker variation has furthermore been investigated in a number of longitudinal studies. Such work is relatively commonplace in the acquisition literature, in which children's articulatory development is monitored (Davis, this volume). Sociolinguistic studies have also tracked the emergence of social and stylistic variation among children (e.g., Roberts, 1997; Docherty et al., 2006; Smith et al., 2007). Systematic sociolinguistic variation is present in the input children receive from the beginning, and it appears to be learned in tandem with aspects of the phonological system from the onset of speech production. There is, furthermore, evidence that input from adults is tailored to the developing social identity of children. For example, in a study of Newcastle English, Foulkes et al. (2005) found speech from mothers to girls contained more standard variants of /t/ than speech to boys, thus mirroring the gender-correlated patterns found in the adult community.

Fewer longitudinal studies have been carried out of adult communities or individuals, due both to logistical difficulties in performing such work and also as a result of the widespread assumption that linguistic patterns are essentially fixed once a speaker reaches adulthood. However, evidence from the few available studies suggests this assumption may be erroneous (e.g., Trudgill, 1988; Yaeger-Dror, 1994b; Sankoff et al., 2001; Nahkola & Saanilahti, 2004; Bowie, 2005; Sankoff & Blondeau, 2008). A particularly famous example is the study of Queen Elizabeth II's speech (Harrington et al., 2005). Over a 50-year period, as evidenced by her annual Christmas broadcasts, the Queen's speech has in some respects shifted in tandem with ongoing changes in standard British English, for instance in respect of lowering and retraction of the TRAP vowel (Hawkins & Midgley, 2005; Fabricius, 2007).

3.1.7 Summary The research reviewed in this section illustrates the complex range of external factors that exert systematic influence on phonetic and phonological form. It also shows that sources of variation are to a large extent "performed" rather than given: variation in speech is certainly *constrained* by biology, but is not wholly shaped by it. Socially determined factors complement those determined by biology, and interact with them, enabling speakers to use phonetic variation as a resource to achieve a range of social goals. Speaker-hearers are socially situated, and the social situation is rich in structured variation, so even traditional experimental laboratory-based phonetic research cannot afford *not* to exploit the opportunities which are available to experimentally control for variation by using socially structured pools of subjects (Scobbie, 2007a).

In general, inter-speaker differences have received too little focused attention in the phonetics and phonology literature, in which they are frequently treated as undesirable noise in the data. (Exceptions to this pattern include Abbs, 1986; Vaissière, 1988; Johnson et al., 1993; Syrdal, 1996; and Allen et al., 2003.) Similarly, sociolinguistic studies have often tended to gloss over differences between individuals' speech productions by pooling or averaging data for speaker groups (but see, e.g., Mees & Collins, 1999; Mufwene, 2001; Beckett, 2003; Piroth & Janker, 2004). The relevance of individual variation to our understanding of social and communicative aspects of language is, however, being recognized more widely by practitioners in both fields (Docherty, 2007). Neither phoneticians nor sociolinguists have addressed issues of ethnicity to the level of detail given to other factors (but see, e.g., Wolfram & Thomas, 2002; Fridland, 2003). Sociolinguists have also been criticized for the implicit determinism of some of their claims (see, e.g., Coulmas, 2005). However, it is now widely accepted that while factors such as region, class, and gender all have an important influence on speech, they do not determine how people speak (Johnstone & Bean, 1997, p. 236). Instead, the array of structured variation available to an individual, coupled with other factors such as ideology (Coupland, 1980; Woolard & Schieffelin, 1994; Milroy, 2001; Wassink & Dyer, 2004), can be seen as a rich resource from which the individual can choose elements in order to project his or her identity and achieve particular communicative goals.

3.2 *Loci of variation*

Segmental variation has been a dominant focus in sociophonetics, but it has become apparent that socially structured variation may be found at all levels of phonetic and phonological structure from subsegmental aspects of timing to suprasegmental properties of larger structural domains.

3.2.1 Segmental variation From the segmental point of view, socially influenced variation can be found at various levels: the phonemic system, phonotactic distribution and lexical incidence of phonemes and allophones, and segmental realization (Wells, 1982; Foulkes & Docherty, 2006). Such differences may be evident across dialects of a language, therefore indexing regional background, and they may also contribute to stylistic differences when speakers shift, for instance, from more to less standard varieties. They may also be subject to variation and change within a community, and thus become associated with subgroups.

Numerous examples of segmental variation have already been given. However, the importance of acoustic analysis of vowels must be acknowledged, since it forms the core of the highly influential sociolinguistic work of Labov et al. (1972). In a large-scale survey of vocalic variation in American English, individual vowel productions were represented by plotting on x–y scattergrams the frequencies of the first and second formants of vowels measured from their midpoints, or points of greatest formant displacement. This was an application of a technique already long established in the mainstream phonetics literature (e.g., Joos, 1948; Peterson & Barney, 1952), and has since been employed in (socio)phonetic work on many languages and dialects (see, e.g., Thomas, 2001, for varieties of English; Gordon et al., 2000, for Chickasaw; Kim, 2005 for Finland Swedish; or Cieri, 2005, for Italian dialects). Of particular interest has been the potential of F1/F2 data to be diagnostic of sound change. For instance, overlap of significant numbers of tokens of ostensibly contrastive phonemic categories on the F1/F2 plane may be taken to indicate suspension of phonetic contrast through merger, while changes in a vowel category's field of dispersion is often interpreted as qualitative drift. A much-studied example is American English /a/ (as in *cot*) versus /ɔ/ (as in *caught*). Figure 19.3 from Majors (2005), to take a random example, appears to indicate for Missouri English that while a contrast is maintained by speakers in St. Louis (lower panel) the vowels are qualitatively nondistinct, as well as backer relative to /u/, in Springfield speech.

Incursion of one vowel's field of dispersion into the area occupied by another may trigger a *chain shift*, whereby movements of neighboring vowels are co-ordinated to preserve the system of contrasts (Docherty & Watt, 2001). A great deal of research has been devoted to the progress of chain shifts in varieties of North American English, in particular the Northern Cities Shift, which is operative in the urban varieties of a large swathe of the north-central United States (Labov et al., 1972). Labov (1994, 2001, forthcoming) has generalized from these studies a number of *principles of linguistic change*, for which he claims both predictive power and robust cross-linguistic validity.

Instrumental phonetic analysis of vowel systems has thus been harnessed both to vindicate theories of sound change first elaborated in structuralist linguistics (e.g., Martinet, 1955; Hockett, 1965) and also as a means of tracking sound changes in progress. The success of this approach has led, however, to a tendency to sideline – or to ignore altogether – many of the more problematic aspects of the methodology. F1/F2 plots are often presented as though they directly represent speakers' vowel productions, despite the fact that they do not incorporate perceptually relevant features such as vowel duration, formant dynamics, formant bandwidth, and contributions to vowel quality made by the third and higher formants. Coincident frequencies of the lowest two formants for two vowel tokens do not necessarily entail their perceptual identity (e.g., Faber & Di Paolo, 1995; Labov et al., 1991; Majors, 2005; Labov & Baranowski, 2006; see also section 4.3 below). Further doubt is cast on the perceptual importance of formant frequencies during the "steady-state" portions of vowels by the results of experiments with silent-center syllables (McLeod & Jongman, 1993; Nittrouer, 2005), which show that formant transitions to and from flanking consonants may carry a good deal of the perceptual load when listeners are asked to identify vowels. Likewise, aspects of vowel production including phonetic quality and gradient phonemicity (Scobbie et al., 1999; Scobbie & Stuart-Smith, 2008), cross-speaker differences in formant transitions (Thomas, 2000; McDougall, 2004) and in formant frequencies (Nolan & Grigoras, 2005) make understanding the complex relationship between vowel production, vowel acoustics, and vowel perception yet more elusive.

One challenge which continues to stimulate research in vowel perception concerns the way(s) in which listeners compensate for formant frequency differences between talkers by normalizing the acoustic consequences of, firstly, vocal tract length (i.e., differences between men, women, and children), and secondly, regional, social, and idiosyncratic accent differences. For sociophonetic purposes, it is clearly desirable to attempt to preserve variation deriving from sources of the second type, while minimizing the effects of variation resulting from vocal tract length differences, since the latter are likely to be of secondary interest. Several reviews

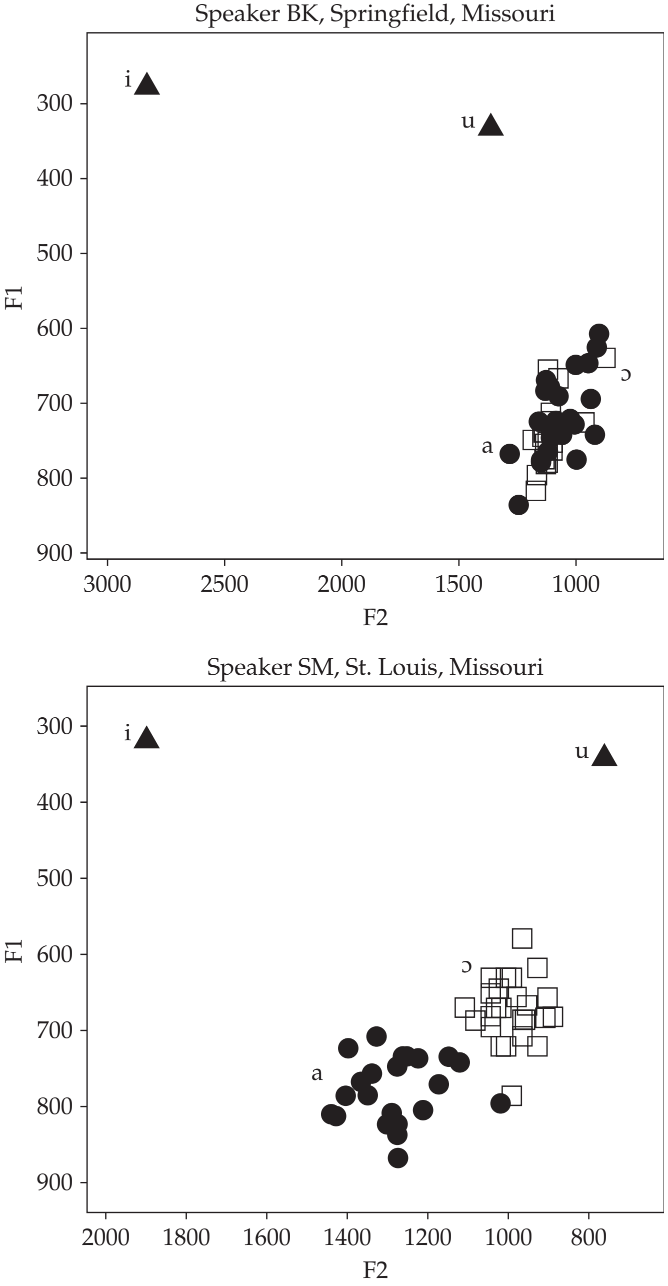


Figure 19.3 Vowel midpoint "static" F1/F2 plots for two Missouri English speakers (lower plot, St. Louis speaker 1; upper plot Springfield speaker 9), showing values for individual tokens of /a/ and /ɔ/ and mean values for /i/ and /u/. (Adapted from Majors, 2005)

of the performance of various normalization algorithms have been published (e.g., Rosner & Pickering, 1994), but the most recent of these (Adank et al., 2004) approaches the issue from a specifically sociophonetic direction by evaluating the relative merits of a range of "vowel intrinsic" and "vowel extrinsic" normalization methods. Watt and Fabricius (2002) describe a routine which can be used to improve the mapping of male and female speakers' F1/F2 spaces as estimated by increases in the spaces' coextensiveness. This is done by rescaling F1 and F2 values relative to the speaker's "centroid," which has as its coordinates the grand mean of the mean F1 and F2 values for /i/, /a/ (or whatever is the most open vowel in the variety under investigation), and a point in formant space at which the speaker's minimum F2 is equivalent to his or her minimum observed F1 value. The last was considered necessary for the routine to be of utility when investigating varieties of British English lacking fully close, back, and rounded vowels. The best means of normalizing other acoustic measures, especially where these have been abstracted from frequency continua (e.g., spectral moments), are less clear.

Consonantal variables have typically been analyzed auditorily rather than acoustically in sociolinguistic work, but recent studies have begun to apply sophisticated analytic techniques to large and heterogeneous data samples. Stuart-Smith's (2007b) study of Glasgow [s], for example (see section 3.1.3 above), quantified acoustic data in terms of the mean center of gravity, spectral peak, and acoustic slope of the fricative. Kissine et al. (2003) followed similar procedures in a study of devoicing of Dutch fricatives, as did Munson et al. (2006) in their comparison of gay and heterosexual speakers. Formant analysis has also been extended to sonorant consonants such as /r/ (Foulkes & Docherty, 2000) and /l/ (Carter, 2003; Livijn, 2002). Heselwood (2007) utilized a combination of acoustic, laryngographic, and nasoendoscopic techniques in his investigation of productions of the *'ayn* in a range of varieties of Arabic.

Other investigative techniques used in sociophonetic work include electropalatography (Wright, 1989; Hardcastle & Barry, 1989), MRI (Zhang et al., 2003) and ultrasound (Mielke et al., forthcoming; see further below).

3.2.2 Subsegmental variation One of the main contributions of instrumental techniques has been to reveal that systematic variation in speech production can reach down to very fine-grained detail. Phoneticians have regularly and frequently documented subtle differences in articulatory targets across languages. Indeed, Pierrehumbert et al. (2000, p. 285) conclude that "there are no two languages in which the implementation of analogous phonemes is exactly the same . . . even the most common and stereotypical phonetic processes are found to differ in their extent, in their timing, and in their segmental and prosodic conditioning." Clear exemplifications of this are provided by cross-linguistic research on features such as voice onset time (VOT) (Lisker & Abramson, 1964; Cho & Ladefoged, 1999), the effects of prosodic context (Turk & Shattuck-Hufnagel, 2000; Suomi, 2005), and connected speech processes (Kohler, 1990; Cucchiarini & van den Heuvel, 1999; Nicolaidis, 2001; Farnetani & Recasens this volume).

Similar differences have also been observed both across and within dialects of the same language. Fourakis and Port (1986) showed dramatic differences in the occurrence of epenthetic stops in dialects of English. While five American subjects produced epenthetic stops categorically in words such as *dense* and *false*, four South Africans never did. These patterns indicate that epenthesis is not the automatic product of universal vocal tract dynamics on articulatory gestures, but must instead reflect learned patterns of articulation. Docherty and Foulkes (1999, 2005) draw a similar conclusion in a study of voiceless stop realizations in two dialects of English. Tokens in the data varied in terms of the extent of continued voicing from preceding vowels, and the presence or absence of release bursts, formant transitions, and pre-aspiration. The patterns can be interpreted as variation within the co-ordination of articulatory gestures. However, some of these patterns were also associated with particular demographic groups within the dialects, such that certain patterns were indexical of, for example, speaker gender. Using electropalatographic (EPG) techniques, Nolan and Kerswill (1990) examined consonantal place assimilation at word boundaries. Their data showed that participants from lower socio-economic groups produced significantly more assimilated tokens than those from the higher ones.

Sociophonetic data provide a very powerful tool for investigating theoretical models of phonetics because they allow experimental examination of slightly different linguistic systems, while holding many other factors constant, something that is far harder, indeed almost impossible, to achieve in cross-linguistic research (Scobbie, 2007a). A particularly interesting subcase of variation is where the phonetic targets of a group of speakers are scattered in a region of phonetic space that would normally be regarded as extending right through adjacent category spaces. Study of fine variation may be an end in itself, but when the "same" phonological opposition is spread through phonetic space in a socially structured way, we are then able to probe directly the phonetics–phonology interface. For example, Scobbie (2005, 2006a) shows via analysis of VOT and prevoicing that the contrast between "voiced" and "voiceless" stops varies widely among Shetland Islanders with different parental backgrounds. The phonetic targets of VOT, for example, span the range from lead to long lag, without any loss of contrastiveness between "/p/" and "/b/", or any sense of a categorical shift from one system of voicing (aspiration-based) to another (voicing-based). The finding of an inverse relationship between the extent of aspiration for /p/ and the rate of prevoicing for /b/ is also found in Aberdeen (Watt & Yurkova, 2007), stratified by age. Kissine et al. (2003) have analyzed a number of cues to the Dutch /v/~/f/ contrast, and while trading relations between acoustic cues to contrast have been studied extensively (Repp & Liberman, 1987; Hodgson & Miller, 1996), sociophonetic variation seems to provide a natural setting for such research.

3.2.3 Suprasegmental variation Suprasegmentals have been studied less frequently in sociolinguistic than phonetic work, largely on account, perhaps, of the difficulty in establishing the functional equivalence of alternate linguistic forms in data samples consisting of uncontrolled materials (see Milroy & Gordon, 2003, pp. 185ff.). Quantification and comparison of patterns is thus rendered particularly complex. Intonational meanings, for example, comprise several strands of information reflecting (at least) grammatical structure, the pragmatic function of the utterance, and the speaker's stance (Cruttenden, 1997).

Regional and social variability has, nevertheless, been studied in respect of many suprasegmental features. Britain (1992), for example, investigates the development of high rising tone in declaratives among speakers of New Zealand English. The innovative pattern was particularly associated with younger speakers, females, and Māoris. It was also found to play particular discourse roles, serving as a means of monitoring listener attention, and helping the speaker to maintain the conversational turn. Other accounts of regional and social variation in intonation include Fletcher et al. (2005), Grabe et al. (2000), Nolan and Farrar (1999), and van Leyden (2004) for English; Dalton and Ní Chasaide (2003, 2005) for Irish; Bruce and Gårding (1978) for Swedish; Selting (2004) and Bergmann (2006) for German; Heffernan (2006) for Japanese; and Ogden and Routarinne (2005) for Finnish, with a specific focus on the discourse functions of rising intonation. Sociophonetic studies of tone languages include Stanford (2007) on Sui, and Hildebrandt (2007) on Manange and other Bodish languages.

Rhythmic features have frequently been compared across languages, and the development of quantitative analytic methods such as the pairwise variability index has permitted quantification which reflects the traditonal categories of stress- and syllable-timing (e.g., Grabe & Low, 2002; see also Ramus et al., 1999). These methods have been applied successfully to dialects of the same language. For example, Carter (2005) presents an analysis of Spanish speakers' acquisition of rhythm in L2 English, while Cedergren and Perreault (1995) discuss age and class effects on syllable timing in Montréal French, and Keane (2006) addresses variation in high and low diglossic varieties of Tamil.

Sociophonetic examinations of vocal setting and voice quality are relatively rare, which may reflect the relative analytic complexity of such features where the descriptive protocol developed by Laver (1980) is applied to large samples of speakers. Stuart-Smith (1999) documents voice quality for 32 speakers of Glasgow English, finding systematic variation related to their age, gender, and social class. More limited studies, involving either impressionistic statements or a focus on the distribution of a particular phonation type, include Henton and Bladon (1988), Esling (1991), and Gobl (1988). Trent (1995) found voice quality differences aided accurate identification of ethnicity when listeners were presented with samples of speech produced by African American and Caucasian speakers, while Podesva (2007) considers the contribution of falsetto phonation to a speaker's construction of a gay identity.

3.2.4 Summary It appears that systematic variation can occur in speech production at all levels of phonetic structure that have been studied in detail in a sociophonetic framework. However, it remains an open question whether certain phonetic or phonological parameters are more or less predisposed to bear the burden of indexical meaning. Labov (2006) appears sceptical that sociophonetic variation can occur in principle in any domain. It has often been noted, for example, that regional variation in English is largely carried by vowel realization (Wells, 1982, p. 178). By contrast, it has been claimed that features such as lexical stress placement appear to vary rather little across English dialects (Wells, 1995). It is of empirical interest to assess whether patterns of sociophonetic variation are constrained by the phonological system of the language, or by other systematic aspects of variation such as those induced by, for example, prosodic structure. Comparing the effects on variation of both internal (grammatical) and external (social) constraints is typical in sociolinguistic studies (Tagliamonte, 2006). However, attempts to assess the *influence* of internal constraints on external ones are relatively rare (Docherty, 2007). Notable examples of such research are Mendoza-Denton et al. (2003), who take account of word frequency, and the extensive survey by Raymond et al. (2006) of factors affecting /t, d/ deletion in English.

4 Sociophonetic Studies and Speech Perception

Although sociophonetic research has been primarily concerned with speech production, attention has increasingly turned to speech perception. Thomas (2002a) provides a detailed review of perceptual studies which are of relevance for sociophonetics. The majority of this work falls into four main categories, each of which is discussed further below.

4.1 *Identifying indexical features*

Many studies have shown that listeners can extract cues to a speaker's social or regional background from the speech signal. Geographical origin has perhaps been tested most frequently (e.g., Bush, 1967; Munro et al., 1999; van Bezooijen & Gooskens, 1999; Clopper & Pisoni, 2004), but there are also examples with a focus on ethnicity (e.g., Trent, 1995; Baugh, 1996), social class (Sebastian & Ryan, 1985), gender/sex (Lass et al., 1979), and sexuality (Munson et al., 2006). Clopper et al. (2006) found a complex interaction of the speaker's gender and regional origin, and "dialect markedness," to influence listeners' judgments of perceptual similarity among four regional dialects of American English.

Foulkes et al. (in press) tested whether listeners could use differences in voiceless stop realization as a cue to speaker gender, using child talkers aged 2–4 years from Tyneside, UK. Their predictions were based on observations of gender-correlated differences in studies of adult Tynesiders' speech. For example, in word-medial intervocalic position (*butter*, *happy*, *baker*), plain oral variants are statistically more frequent for females, while males prefer glottalized forms (Docherty & Foulkes, 1999). In the experiment by Foulkes et al., a group of listeners from Tyneside heard a set of single word stimuli containing voiceless stops and were asked to identify the sex of the child. Control groups of Americans and British listeners from other regions were also recruited. Results showed differences across the groups in the predicted directions (Figure 19.4). The Tyneside listener group (but

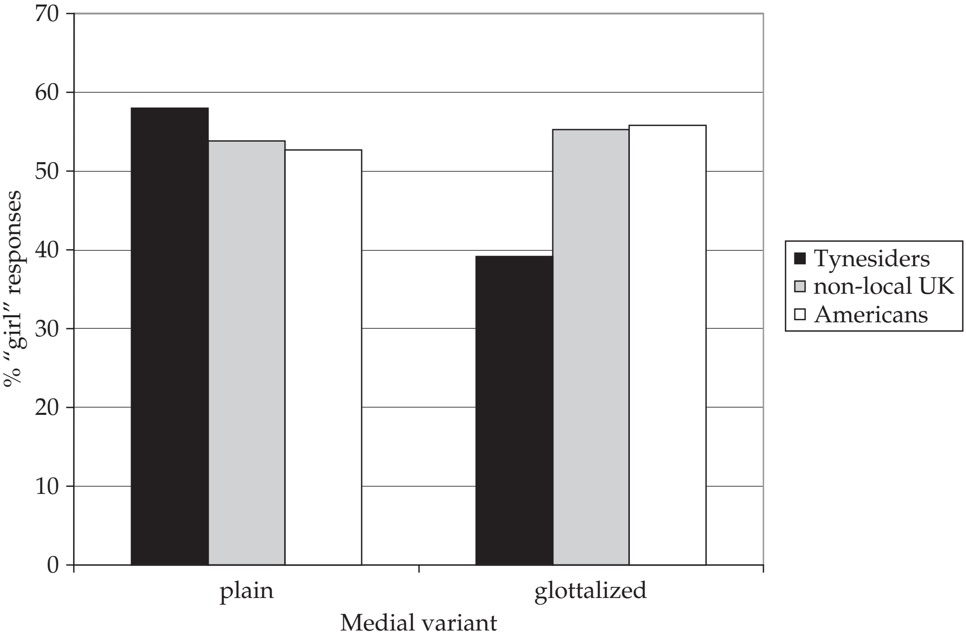


Figure 19.4 Percentage of "girl" responses to word-medial tokens by listener group and variant. (Adapted from Foulkes et al., in press)

not the control groups) gave significantly fewer "girl" responses to stimuli with glottalized stops than they did to stimuli with a plain variant, and they also gave fewer "girl" responses to glottalized tokens than either control group did. The results support the conclusion that local listeners display tacit knowledge of statistical associations between phonetic variants and socially defined categories of speaker.

The main aim of studies such as these has usually been to identify which features listeners use in the identification process, and whether these coincide with indexical features observed in studies of speech production in the communities concerned. However, it has also been found that listeners vary in their ability to perform these tasks, and furthermore that there is variability in the weight of perceptual cues across speakers, listeners, and situations. Some of the variability is no doubt the result of differences in experimental designs: studies have differed, for example, in the length and type of stimuli used, from extended extracts to single vowels (Walton & Orlikoff, 1994), and natural, filtered (Lass et al., 1980) or resynthesized speech (Graff et al., 1986). Understanding this variability in listener performance relative to speech input is of particular relevance in the forensic domain when lay listeners may be called upon to give evidence about a voice they have heard in a criminal context (Bull & Clifford, 1984; Blatchford & Foulkes, 2006).

Research categorized under the label *perceptual dialectology* (Preston 1999; Long & Preston, 2002) can also be considered sociophonetic, although its experimental methods typically do not involve the presentation of recorded samples to listeners. Rather, they assess listeners' awareness and memory of regional and social varieties through tasks such as identifying dialect boundaries on maps.

4.2 Evaluating indexical features

Listeners' subjective evaluations of indexical features have been investigated in a number of studies. As Thomas (2002a) notes, formal assessment of listeners' interpretations of linguistic variation dates back at least to Pear (1931). Choice of language or variety, alternative pronunciations, and variation in acoustic and phonetic parameters of voice may all affect the way listeners judge the personality of a speaker. Various techniques have been employed including the matched guise paradigm (Giles & Powesland, 1975; Cargile et al., 1994) and experimentally modified stimuli. Some studies have linked listener attitudes to particular aspects of speech. For example, van Bezooijen (1988) suggests that her Dutch listeners drew upon prosodic differences in their evaluation of "strong" personality, but social status and intelligence were linked more to segmental features. This remains, however, an under-researched area and one in which judgments and social evaluations are perhaps more likely to be locally determined rather than based on universal associations of phonetic features with character traits. Coupland and Bishop (2007) show that evaluations of regional accents of English vary in respect of the speaker's gender and the listener's own background. Moreover, many judgments differed by age of listener, indicating that attitudes may change over time. Standard accents, for example, were rated less highly by young listeners, who by contrast gave much more favorable ratings than older listeners did to London, Australian, and West Indian accents.

Understanding attitudinal responses to linguistic variation is particularly important in studies of inter-ethnic communication (Lambert et al., 1960; Gumperz, 1982; Purnell et al., 1999). Attitudes have also been discussed in relation to issues such as the impact of the media on children (Lippi-Green, 1997), success in the job market (Milroy & Milroy, 1998), the likelihood that a jury will convict or acquit (Dixon et al., 2002), and the transmission of linguistic change (van Bezooijen, 2005).

4.3 Perception of ongoing change

A number of investigations have assessed whether listeners are able to perceive sound changes in progress such as phonemic mergers (e.g., Hay, Warren et al., 2006). Presumably, the diffusion of changes throughout a population depends upon listeners' abilities to detect others' innovatory usages before they can adopt them in their own speech, even if those differences are apparently indistinguishable by outside observers, including phoneticians using a conventional battery of analytical methods (see section 3.2.1). The presence and persistence of marginal contrasts (e.g., near-mergers) in phonological systems also have implications for principles such as maintenance of perceptual distance that underpin models of the cognitive representation of speech and language.

The results of these studies are, however, somewhat mixed. It has been argued that listeners lose the ability to perceive contrasts that are disappearing, even if they retain the contrast in production (e.g., Janson & Schulman, 1983, on a vowel merger in Swedish). However, Labov et al. (1991) argue that such results may reflect the artificiality of the test scenarios (listeners being asked to label isolated vowels, or synthetic stimuli). Labov et al. used more natural stimuli to test Philadelphians' ability to discriminate pairs such as *merry* and *Murray*, which are near-homophonous or indeed fully merged for many speakers in Philadelphia. Their results suggested that listeners could recover distinctions, albeit with difficulty. Similar findings are reported by Di Paolo and Faber (1990).

4.4 Impact of social and regional variation on perception and processing

The perceptual studies reviewed above have primarily been carried out to test listeners' reactions to features of interest from a sociolinguistic perspective. Parallel developments have seen researchers in experimental speech perception consider the impact of social, regional, and inter-speaker variation on tasks such as lexical access and phoneme identification.

Gender differences in speech have received relatively scant attention in the perception literature. Strand (1999) found that her subjects' perception of the category boundary between /s/ and /ʃ/ was influenced by the gender of a person in a picture they were shown while performing the listening task. Similar findings are reported for vowel categorization by Johnson et al. (1999). In experiments by Niedzielski (1999), listeners were exposed to voice samples they were told were either Canadian or Michigan English. From a set of synthesized vowels they were then asked to choose exemplars most appropriate to the variety they had heard earlier. Their decisions differed according to the variety they believed they had been listening to. Sociolinguistically distributed cues present in the acoustic signal and inferred social factors interact, for example in the perception of Vietnamese tone (Brunelle & Jannedy, 2007). It appears, then, that sociolinguistic expectations may influence basic speech perception to quite a marked degree, with increasing evidence of the implicitness of the expectations (see also Elman et al., 1977; Janson, 1983; Hay, Warren, & Drager, 2006; Warren et al., 2007).

Previous exposure to particular accents or to the voices of individual talkers may also have an impact upon processes such as word recognition, making this an active area of research where phonetic, sociolinguistic, and psycholinguistic concerns are converging (e.g., Delvaux & Soquet, 2007). Nathan et al. (1998) found better word recognition rates by children who were familiar with the talker's accent. Evans and Iverson (2007) report similar results, with speakers of southern English accents proving better at identifying southern-accented words embedded in noise than were speakers of northern accents. With respect to individual voices, Nygaard et al. (1994) found that subjects performed better at word recognition tasks than control subjects who had not previously heard the talkers' voices. This implies that information about the features of individual voices is retained by listeners and can be drawn upon if necessary. Nygaard and her colleagues argue on this basis that lexical representations of words must contain speaker-specific details alongside the more abstract information that permits lexical access when listening to novel talkers (see also Hawkins & Smith, 2001; Lachs et al., 2002; Hawkins, 2003; Nygaard, 2005; Docherty, 2007; and Mitterer & Ernestus, 2008 for a contrary view).

4.5 Summary

These perceptual studies reviewed above show that listeners can and do access indexical information in the course of listening to speech. Thus there must be some cognitive storage and processing of that information. The question remains as to what form this representation takes, and how it is stored and processed relative to more traditional conceptions of "linguistic" knowledge (see further section 6 below).

5 Methodological Issues

It is clear from the disparate lines of work we have reviewed that one of the defining features of sociophonetics is that it draws upon a particularly wide array of methodologies that have generally been developed in other, longer-established disciplines. We do not propose to provide a critique of all methods which have been used in sociophonetic studies. However, it is worth drawing attention to major differences between the methods typically employed in sociophonetic studies of speech production compared with those of laboratory phonetics and phonology. We also take the opportunity to highlight some of the newer methodologies currently being applied to sociophonetic work.

5.1 Data collection

Data collection methods are perhaps the most obvious point of disparity, reflecting differences in the prevailing research questions of sociolinguistics on the one hand and phonetics/phonology on the other.

In Labovian sociolinguistics the vernacular has always been regarded as the most prized speech style. The focus of such research on establishing variation within and across communities has furthermore meant that data are usually collected from a heterogeneous speaker sample, and often involve a range of speech styles. These emphases lead sociolinguistics and sociophonetics into a position of conflict with most other empirical fields, including experimental phonetics, phonology, and acquisition, which have usually been most concerned with citation forms and (implicitly) standard varieties. Experimental studies in these fields have, moreover, generally used heavily controlled and/or artificial materials and experimental tasks. Materials are often elicited through randomized word-lists, sometimes consisting of nonsense words which conform to the phonotactics of the language in question. Data are usually gathered in ideal acoustic conditions, often from relatively small and homogeneous groups of speakers of standard dialects, often colleagues or university students, speaking in isolation.

The contrasting approaches naturally have both strengths and weaknesses. The control exercised over laboratory materials facilitates analysis and comparison of spoken material across languages, contexts, or speakers, with, in principle, all factors held constant except for the features under scrutiny. Materials gathered in the field, on the other hand, can be difficult because of impaired technical quality, the unpredictable returns of spontaneous data (overlapping speech, the lack of sufficient tokens of the features of interest), and because analysis needs to cater for many potential factors which may influence phonetic forms. Duration studies, for example those of vowel length or VOT, are exceptionally hard to perform on uncontrolled data because of the effects of factors such as phonological context and overall speech rate. More generally, acoustic images of data from spontaneous interaction do not always reduce easily to the neat templates provided in acoustics textbooks.

Analysis may therefore need to begin with an assessment of the phonetic categories apparent in the data, even if these do not conform to prior expectations based on, for instance, the IPA definition of a sound. For example, in an analysis of voiceless stop realizations in Newcastle English, Docherty and Foulkes (1999, 2005) constructed an acoustic profile of each token, describing the presence or absence of acoustic features as well as quantifying key parameters. This detailed record allowed tokens to be categorized for the purposes of discussing patterns of social distribution. The same technique was applied to children's speech in a subsequent study (Docherty et al., 2006). Examining phonetic tokens in such detail permitted an essential degree of refinement in the consideration of whether children were mastering the acquisition of the stops. The analysis was able to take into account the full range of variant forms found in the ambient language, and thus also circumvented a potentially misleading reliance on standard or citation forms as the putative targets for acquisition. Khattab (2006, 2007, 2009) followed a similar approach in her study of Arabic-English bilingual acquisition, showing the importance of establishing targets for acquisition that take account both of local norms and of variation in the community's speech patterns.

While laboratory materials are very useful in many ways, they are problematic in others. They only scratch the surface of the informants' phonetic repertoire and thus limit the theoretical inferences that can be drawn with respect to speech planning or phonological knowledge. It may be easy to elicit particular strings of phonemes through nonsense materials, but extremely difficult to ensure that the pronunciation of these high-level units approaches in any way the forms that are observed in natural use. Laboratory materials may also lack severely in naturalness. Elicitation using read materials, for example, has often been criticized by sociolinguists.

If the spoken vernacular differs markedly from the standard written form, reading aloud may represent a discrete linguistic task rather than a point towards the formal or hyper end of a style continuum (Milroy & Gordon, 2003, p. 201). This is perhaps clearest in the case of diglossic communities such as those of the Arabic-speaking world, but the same issue can be raised in any community. We should also bear in mind that reading tasks may be threatening, inappropriate, or unworkable for many speakers such as young children or members of nonliterate communities, and materials tailored to provide minimal pairs for segmental phonological or phonetic analysis may result in speech which is too narrow for prosodic analysis. University-educated adults make convenient and sophisticated research subjects, and phoneticians are well aware that many competent language users from other groups present methodological difficulties. Elicitation methods may therefore need creative adaptation, such as the use of picture-based tasks (Khattab, 2006, 2007, 2009), and interactive or distracting activities such as map tasks (e.g., Grabe, 2004) or spot-the-difference tasks (Bradlow et al., 2007). In any case, it is essential that the experimental method and materials are appropriately paired, and interpreted with respect to the situation.

Sociolinguistic methods present different problems. Collecting suitable data entails the consideration of many sampling and fieldwork issues, including how to define the speech community and the relevant social or demographic divisions, how to elicit appropriate linguistic styles, and how to obtain adequate material in field settings. Ladefoged (2003), Milroy and Gordon (2003), and Tagliamonte (2006) provide excellent overviews of such issues. The necessity of sampling from different speaker groups means that corpora collected in sociophonetic fieldwork can become particularly large, and thus time-consuming both to collect and analyze.

In light of the various decisions that must be made, and the need to tailor fieldwork and analysis to the speech community as well as the potentially diverse research questions at stake, there can be no fixed protocol for sociophonetic data collection. However, a common base for research in the Labovian variationist tradition would be to collect a combination of spoken materials, some relatively well controlled and some to reflect the natural repertoire of the language users being studied. Speaker samples are often constructed around broad social categories such as age, class, and sex/gender, and clearly contrasting groups are selected rather than attempting to sample the whole community along demographic continua. For example, in studies of "class," speakers may be selected from markedly different neighborhoods, judged via local knowledge of typical housing and work types, rather than through the application of complex systems to quantify the relevant demographic factors that contribute to an individual's social class (e.g., Watt & Milroy, 1999; Stuart-Smith, 1999).

Depending on their aims, other types of sociophonetic work may present stark contrasts in terms of the amount and range of data collected, and the size and range of speaker samples. Work on the phonetics of conversation, for instance, may involve intricate analysis of small sets of conversational fragments, with no specific interest in the social background of the speaker (e.g., Local, 2003; Plug, 2005). More ethnographically informed research may be based on coarser analysis of extensive speech samples from relatively few people (e.g., Mendoza-Denton, 1996, 1999; Hay & Drager, 2007). Surveys of the geographical distribution of phonetic forms may involve large speaker samples, but trade this requirement against a relatively small quantity of material (Labov et al., 2006).

5.2 Data analysis

Acoustic analysis is now more widespread and much easier to perform than ever before thanks to the availability of free analysis packages such as Praat and Wavesurfer, and specialized additions for analysis, storage, and presentation of data (e.g., Akustyk). However, it is important to bear in mind that acoustic data are not inherently superior to data derived from careful auditory analysis. While the latter may be coarser, it has the advantages of being faster, and processed through the best normalization mechanism yet developed: the human ear and perceptual system. Although auditory analysis may be argued to run the risks of human error and subjectivity, acoustic analysis can be subject to similar problems. Acoustic data, and thus the theoretical claims made on the basis of the data, are all affected by the analyst's choice of recording equipment, software package, analysis settings, measurement criteria, and location of measurement for a given token. Differences generated by such decisions may be far from trivial. Illustrations of striking variation in acoustic data include Harrison's (2004) comparison of formant measurements using different software systems and settings, and the effects on formant data of both telephone transmission (Künzel, 2001; Byrne & Foulkes, 2004), and microphone types (Plichta, 2004). For a summary of best practices in handling acoustic data see the regularly updated reviews on Plichta's website (http://bartus.org/akustyk).

Instrumental articulatory phonetics has had a limited impact in sociophonetics. Research projects conducted in Cambridge by Kerswill, Nolan, and Wright in the late 1980s used EPG (Stone, this volume), and convincingly showed the value of articulatory data (Kerswill, 1985; Kerswill & Wright, 1990; Wright & Kerswill, 1989). Standard arguments that instrumental analysis can be more powerful and reliable than transcription were extended through study of an important sociolinguistic variable, namely /l/ vocalization, which is difficult to analyze acoustically. Acoustically subtle aspects of articulation can be explored with EPG or other techniques, but this approach was not subsequently adopted in sociolinguistics, not least perhaps due to issues of cost and convenience. In more recent work, Scobbie and Wrench (2003) and Scobbie et al. (2007) have undertaken fairly standard lab-based phonetic studies which, by focusing on broad dialectal and subtle inter-speaker variation in /l/ vocalization, again make a case for sociophonetic articulatory research.

It is possible to gather some types of data about articulation as simply as making an audio recording, for example with a camera, but quantitative and intra-oral data constitute a more difficult proposition. A priori, articulatory data are equivalent to acoustic data as a means to a sociolinguistic end, and are merely problematic for two main logistical reasons. First, equipment to collect such data (and the expertise to use them meaningfully) is not as available as audio-recording equipment. Second, the relative invasiveness of the data collection process may be expected to interfere with speaker behavior, especially if speakers have to be recorded in laboratories. On the other hand, articulatory data can make a novel and important contribution to the analysis both of the complex and unpredictable relationship that exists between the sounds of speech and the vocal tract configurations that generate them, and also to the social variables which shape interaction. It is debatable whether any model of the speaker in his/her social context can be complete without articulatory data. One view (Thomas, 2002b, p. 168), explicitly relating to the variable (r), holds that only *acoustic* instrumental analysis is relevant. Such a view adopts a listener-oriented theory of variation, in which it is only what people hear that matters. The alternative view holds that there is also a pressing need to investigate how speakers physically create sounds in a social context, in order to examine the role of the speaker as a sociolinguistic *agent*. The aim of speakers may indeed be to reproduce in acoustic space the sociolinguistic variants which they themselves hear around them in order to convey social meaning appropriately, but their speech production strategies to achieve this goal may well differ from those used by other people.

A technique which appears particularly promising for articulatory sociophonetic research is Ultrasound Tongue Imaging (UTI) (Gick & Wilson, 2006; Stone, this volume), which has some advantages over EPG in immediacy of use, especially for obtaining qualitative articulatory information on the location, shape, and movement of a large part of the mid-sagittal section of the tongue. Its main disadvantages are that it is hard to obtain good acoustic–articulatory temporal alignment and accurate spatial images of the tongue (Wrench & Scobbie, 2006) and to ensure stability of the probe (Scobbie et al., 2008), and it is not agreed how to quantify the tongue images for statistical analysis. This last point may be a particular problem for sociophonetic work in that data are required from relatively large samples of speakers. EPG is excellent for the study of anterior constrictions, while UTI seems ideal for looking at secondary articulations like velarization or pharyngealization, because the articulations are slow-moving open constrictions. Perhaps ultimately both techniques will be used simultaneously for sociophonetic research (Wrench & Scobbie, 2003).

To evaluate the approach, Scobbie et al. (2008) have explored the methodological ramifications of the use of UTI within an otherwise standard sociophonetic design, with subjects aged 12–13. Initial results indicate that word-list style speech differs little if UTI measurement is introduced in a field setting. Perhaps these speakers are more likely to be influenced by the presence of a conversational partner who is a friend than by the experiment and the equipment per se. Moreover, while it might seem obvious that an observer effect could be greater when the speaker is aware that their speech is being measured articulatorily than where the measurement is merely acoustic (through an audio recording), we suspect that participants do not strongly associate the measurement of their oro-facial physiology to speech and accent. The implications of speaking into a microphone are far more obvious: a researcher is going to listen and make judgments based on the sound of the voice. For the lay person, articulatory data from inside the mouth are esoteric, physical, and removed from normal linguistic experience.

The resulting ECB08 corpus, undertaken in schools and in the laboratory, confirms findings from purely laboratory-situated pilot studies (Scobbie & Stuart-Smith, 2005): some speakers attain the acoustic goal of sounding derhoticized (hence young and vernacular) using articulatory routines which nevertheless contain persistent strong rhotic-like gestures. As well as the more predictable gestural reduction, UTI reveals that in some cases, *covert* rhotic-like lingual articulations such as retroflexion may be masked by devoicing and temporal delay into post-utterance silence, so that they generate little or no rhotic auditory/acoustic effect (Figure 19.5). Thus, for reasons that are not yet clear, and in contexts and styles that are not yet understood, speakers can aim for a derhoticized acoustic target (which carries a particular sociolinguistic meaning) using an articulation with reveals a strong but relatively inaudible reflex of the diachronically previous and apparently still "underlying" affiliation. Similar behavior may also occur in Dutch (Scobbie et al., in preparation).

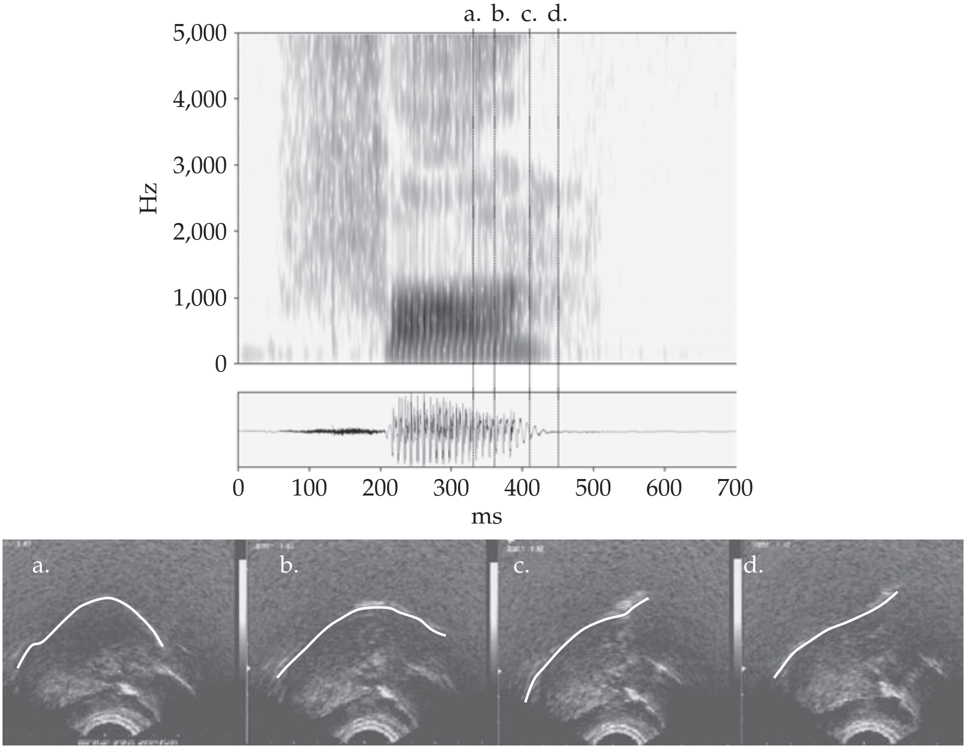


Figure 19.5 Four ultrasound images showing tongue-tip raising late in a pre-pausal derhoticized token of citation "for" from a Scottish English speaker. Anterior to right. The third formant remains high throughout the token [fʌ].

Finally, statistical analysis is generally essential in sociophonetic work, especially given the likelihood that complex sources of variation in the design and the focus on spontaneous speech may yield some messy data. Appropriate techniques must therefore be chosen carefully. Sociolinguists have generally been content to identify variable patterns with the Varbrul program (Rand & Sankoff, 1990). However, this program is not without its limitations (being restricted to categorical data, for example) and it enjoys little popularity in other fields (Pierrehumbert, 2006). Generic tools such as regression and analysis of variance find more favor in experimental phonetics, where the researcher has more control over the number of tokens per cell. Cluster analysis is also used in case studies where several variables are examined simultaneously (Horvath, 1985; Stuart-Smith et al., 2007). For a general survey of statistical techniques see Rietveld and van Hout (1993).

5.3 Outlook

While sociophonetics is characterized by its eclecticism with respect to data and methods, it is apparent that its techniques have so far been applied to relatively few languages. Work on English is especially dominant (as testified by the bias in examples cited in this chapter). However, sociophonetic studies of other languages are increasing in number, with recent work including studies of Albanian (Moosmüller & Granser, 2006), Arabic (Khattab et al., 2006), Irish (Dalton & Ní Chasaide, 2003, 2005), Latvian (Bond et al., 2006), and Shoshoni (Elzinga & Di Paolo, forthcoming).

Sociophoneticians are furthermore coming together to share good practice in analytic techniques, especially those that can be applied to spontaneous speech. The annual NWAV conferences in North America have staged workshops on sociophonetic techniques applied to speech production since 2004, and a first textbook is to appear (Yaeger-Dror & Di Paolo, 2010). Sociophonetics still needs phonetic scientists to develop better techniques for processing large quantities of data, often spontaneous speech, and for normalizing across different speakers (Van de Velde, 2007).

6 Theoretical Implications of Sociophonetic Studies

It will be apparent from the foregoing that sociophonetic data have been harvested to address a wide range of theoretical issues, reflecting the range of disciplines that have contributed to the development of sociophonetics as a field. We offer here a brief summary of the main theoretical areas of concern to sociophoneticians.

Given the historical origin of sociophonetics within sociolinguistics it is no surprise to find considerable overlap in their theoretical interests. Labov's work has always been principally concerned with providing explanations for language change: how changes originate and how they spread through grammars and communities (Labov, 1994, 2001, forthcoming; Milroy, 1992). Sociophoneticians have naturally focused on aspects of sound change. The contribution of sociolinguistic work in general to historical linguistics has been to complement the theoretical predictions of earlier schools, especially those of the neogrammarians and structuralists. The claims of such schools were largely based on concepts relating to the grammatical system, such as functional load and symmetry (McMahon, 1994). Sociolinguists have agreed that such factors may indeed contribute to determining which changes are more likely to occur, and what paths they might take. Indeed, as noted in section 3.2.1, Labov's chain shift model draws explicitly on structuralist notions of the phonological system.

However, sociolinguists have demonstrated that it is essential to make reference to human communities and human interaction in order to fully understand how and why changes take place where and when they do. Changes operate because communities are heterogeneous, and because speaker-listeners *evaluate* competing linguistic forms. They recognize that variants have indexical meanings and thus that their use may be more or less attractive, appropriate or valuable in particular social circumstances. Positively evaluated variants (such as coda /r/ in American English) generally spread at the expense of their less positively evaluated rivals. The contribution of theoretical tools from sociology, social psychology, and other neighboring disciplines cannot be underestimated in this regard. Frameworks such as social networks and communities of practice have both been imported into linguistics and have led to significant advances in our understanding of the structure of human interaction and its effects on language.

Experimental phonetics has itself made considerable advances in respect of the *actuation* problem, or the question of where and why a change begins. Experimental studies explain how phonetic innovations may arise as a result of the dynamic actions of the articulatory system, the effects of aerodynamic principles operating within the vocal tract, and the properties of the perceptual system. It has been shown, for example, that contrastive systems of high and low tones arise through reanalysis of fundamental frequency differences originally associated with consonant voicing (Hombert et al., 1979), while affrication of stops is most likely to develop adjacent to close vowels because of the likelihood that vocal tract narrowing will create turbulent airflow (Ohala 1983, 1989). Such explanations are limited, however, to phonetically transparent and cross-linguistically recurrent changes. They do not explain the more arbitrary developments found in abundance in sociolinguistic studies, such as the change in English /r/ which has taken opposite paths in different parts of the English-speaking world. Labov has attempted to make sense of the apparent arbitrariness of many changes by appealing to degrees of conscious awareness of variable forms on the part of speaker-listeners. Variables may be ranked as stereotypes, markers, or indicators, in decreasing order of awareness. Different types of change may affect the different types of variable. It remains a moot point whether phonetic forms can be shown to have universal degrees of *salience*, equally noticeable no matter what the community or language concerned (Docherty, 2007). Frequency effects may interact with those of social evaluation to determine the outcome of change (Bybee, 2001), as in the case of dialect leveling changes and new dialect formation (Kerswill & Williams, 2000; Trudgill et al., 2000; Trudgill, 2004). The features of new dialects, as in the case of New Zealand English, tend to be drawn from the common shared features of the contributing dialects, with minority forms becoming lost.

Sociophonetic data have made less of an impact on the main theoretical developments in phonetics and phonology. Following the pattern of Chomskyan linguistics in general, phonology and phonetics have largely pursued an active strategy of eliminating many aspects of variation, including socially structured variation, from their purview. Theories of speech production and perception (Elman & McClelland, 1986; Levelt, 1989; Löfqvist this volume) have certainly made reference to variation in spoken form, but in general this has been variation connected to prosodic context, segmental environment, speech rate, etc. Phonology has likewise tended to be concerned with aspects of variation that can be considered allophonic or the subject of phonological rules or processes (depending on the terminology used in the particular model).

As a consequence, there has been rather limited communication so far between sociophonetics on the one side and phonological and phonetic theory on the other. Sociophoneticians have not yet tested the full range of predictions made by theoretical models, while theoreticians have been slow to take account of sociophonetic data and the challenging testing ground that they provide. Some collaborative progress has been made, however, as the following examples illustrate. First, adjustments have been made to the machinery of several phonological models in response to the findings of sociophonetic work, including Optimality Theory (e.g., Nagy & Reynolds, 1997; Anttila, 1997; Coetzee, 2006). Second, work on the phonetics of conversation has established that many aspects of speech planning are mutually negotiated by partners in interaction, and that listeners orient to fine-grained aspects of phonetic detail in the construction of conversation. Findings such as these raise some serious challenges to many of the fundamental assumptions of modern linguistic theories, including segmental structures, the core role of the lexicon, and the emphasis on speech planning being the product of one party (Local, 2003, 2007). Finally, data from sociophonetic studies of speech production by adults and children are contributing to the refinement of exemplar models of phonological knowledge, first applied to language in speech perception research (Foulkes & Docherty, 2006; Hay, Nolan, & Drager, 2006; Johnson, 2006; and for a critique, Labov, 2006). Exemplar models are prime candidates to accommodate sociophonetic data, since they depart from tradition in taking aspects of variation as central facts to be accounted for and explained (Pisoni, 1997; Pierrehumbert, 2002). They also implicitly acknowledge that socially conditioned variation may overlap with aspects of variation which result from reflexes of the phonological system, being manifested in the same phonetic materials (Docherty et al., 2006).

7 Wider Applications of Sociophonetics

As well as contributing to linguistics, sociophonetic studies provide valuable resources for a range of applied fields.

Variation poses a perennial problem for speech technology. Natural variation must be catered for in speech and speaker recognition systems to ensure robust performance (Hoequist & Nolan, 1991; Laver, 1995; Bates et al., 2007). Descriptive sociophonetic accounts can contribute to such systems, identifying the loci and parameters of variation for speakers, dialects, and contexts. They may assist in refining speech synthesis programs, rendering them more natural-sounding and acceptable to listeners. Synthesis systems are also being developed which permit options for regional dialect and other indexical features of the speaker's voice (Fitt & Isard, 1999; Carlson & Granström, this volume). Sociophoneticians and speech technologists have furthermore often shared corpora of recordings in the pursuit of their respective goals (e.g., Glenn & Strassel, 2006).

In speech and language therapy, sociophonetic research provides a baseline of normal patterns of within-speaker and within-community variation. This can assist the speech and language therapist in distinguishing genuine pathology from nonstandardness in children's speech and language development (Howard & Heselwood, 2002; Oetting, 2005), and can also be used to inform appropriate diagnosis and treatment in adults (Milroy, 1987a, pp. 208 ff.; Docherty & Khattab, 2008). The close resemblance of certain innovatory speech forms to infantile and/or pathological pronunciations (e.g., labiodental /r/ or (th)-fronting and -stopping in English) can make this difficult. It increases the likelihood that non-disordered early adopters of such sound changes are more likely to be misidentified as phonologically delayed or disordered than their peers, with obvious repercussions in terms of workload and commitment of resources, and potential distress to the child and his/her caregivers. The sociophonetic literature can therefore be a valuable aid to speech and language therapists whose task it is to decide whether or not to recommend treatment, a consideration explictly recognized by a number of contributions to McLeod (2006). Among other things it may help them to identify the point at which intervention becomes unnecessary or self-defeating in cases where the change has been adopted by a critical mass of speakers of the variety (see further Watt & Smith, 2005).

A refined assessment of normal patterns of spoken variation may further assist in pedagogical issues. Educationalists are better equipped to assess educational needs in particular communities as a result, for example, of understanding the differences between standard written forms and the local pronunciation norms. A striking example where this issue came to public attention was in the United States in 1979, when the presiding judge upheld a suit brought against the School District Board of Ann Arbor, Michigan, by black parents claiming that the school system had violated their children's rights by failing to teach them standard English or to take their spoken dialect into account during their education (Freeman, 1982; Wolfram & Schilling-Estes, 1998, pp. 169 ff.).

Many sectors of commerce plan their business strategies with reference to linguistic factors, including assessments of the impact of spoken variation on their markets (Bell, 1991, pp. 135 ff.). For instance, the locations for telephone call centers may be chosen in part because the local speech variety is deemed to be attractive or acceptable to clients. Advertising campaigns may select specific regional accents or individual voices to maximize the appeal of the product to customers. Linguistic researchers are also coming to study the effects of language choices on workers as well as businesses, for example with reference to the use in call centers of written texts, explicit training in conversational style, and the forced choice of sociolinguistic forms (Mirchandani, 2004; Orr, 2007).

Finally, sociophonetics plays a central role in the growing field of forensic phonetics. Understanding cross-speaker and within-speaker variation is essential in the process of speaker comparison, in which the recorded voice of a criminal is compared with that of a suspect. Of particular importance are the establishment of the distribution of features across populations and the parameters of variation in different settings (for example, assessing the impact of speaking on a telephone, and the effects on the voice of emotion and intoxicants). The effects of aging may also be important in cases where there is a long delay between the recording of the crime and that of the suspect. In another forensic task, speaker profiling, the analyst is asked to describe the likely source of a voice in order to narrow the field of potential suspects, for example in cases where a telephone call or tape recording may be delivered by a kidnapper. The strength of conclusions that can be drawn depends largely on the documentary record of how linguistic features are distributed (for an example, see Ellis, 1994; French et al., 2006).

8 Conclusion

As this review has illustrated, sociophonetics is a diffuse research field, but one which is beginning to lay claim to be a core phonetic science. Its unifying characteristic is that it is born of a cross-fertilization of methods and theories, drawn especially from phonetics and sociolinguistics but increasingly grafting itself to the principles of other disciplines. The strength of this pedigree is that it enables sociophonetics to address some of the weaknesses in its component parts (Thomas, 2002b). The source materials of sociophonetics include not only standard dialects and citation speech, but a range of speech styles and they display a particular emphasis on spontaneous interaction. Data samples are typically large and elicited from heterogeneous samples. The methodologies employed by sociophoneticians range from controlled experimentation to ethnographically informed observation of speaking and listening in different situations. The analytic methods used in sociophonetics span a wide range of techniques, both instrumental and auditory. Sociophonetics is also informed by a variety of different theoretical models, and its results are in turn being used to address a wide range of theoretical issues.

The array of materials, methods, and models testifies to the recognition by sociophoneticians that speech is a multifaceted signal, replete with systematic variation resulting from many sources, and fulfilling a wide range of functions.

Placing the social complexity of speech center stage offers a strenuous challenge in explaining how aspects of variation are learned, stored, and processed. In particular, it remains to be seen whether sociophonetic variation and "pure linguistic" knowledge are best handled as discrete cognitive modules (Docherty & Foulkes, 2000; Docherty et al., 2006; Scobbie 2007b).

Labov has famously noted that he long resisted the term "sociolinguistics" because of its implication that there might be a successful linguistic theory which is not social (Labov, 1972, p. xiii). By the same token, many researchers in phonetics are now coming to the view that abstraction of speech from its social context limits the power of phonetic research. Thus, socially structured variation is both a topic of undeniable theoretical importance for the phonetic sciences, and a phenomenon that can be exploited by phoneticians of all types in the pursuit of the very widest range of research interests. Fine differences in the phonetic systems of individuals that are not merely physiological are part of the grammar, and should not be marginalized as "variation-as-noise," because they can be used to explore the fundamental bases of sound systems in far more subtlety than cross-linguistic differences. Subtle and otherwise, many speech phenomena are learned and used as part of the construction of social identity, making sociophonetics one of the key phonetic sciences.

NOTES

We record our thanks to Bill Haddican, Ghada Khattab, and Malcah Yaeger-Dror for their valuable comments on drafts of this chapter, more of which we wish we had space to incorporate. Jane Stuart-Smith kindly provided us with the material for Figure 19.2. We thank Tivoli Majors for providing the data for Figure 19.3. Scobbie gratefully acknowledges ESRC grant RES-000-22-2032.

[1] Although sociophonetics does not yet appear in published versions of the *Oxford English Dictionary*, the *OED* archives provide one instance of "socio-phonetic" which predates Deshaies-Lafontaine. The term is used by Halle (1963, p. 10) in a translation of a Russian text by Gvozdev from 1949. However, its sense there is metalinguistic and does not concur with modern usage. Our thanks to Gillian Evans of the *OED* and Mike MacMahon for drawing this to our attention.

[2] We acknowledge that the authors of the works we cite may not themselves use the term sociophonetic to describe their research.

[3] A number of sociolinguists utilize Silverstein's (2003) distinction between "first-order" and "second-order" indexicality in accounting for differences in functional aspects of linguistic variation. First-order indexicality refers to the (objective) association of particular patterns of linguistic behavior with globally or locally meaningful social groups, while second-order indexicality pertains to speakers' subjective metalinguistic knowledge of the social and communicative roles played by variable linguistic forms.

[4] The use of the symbol /r/ here is intended to cover a range of possible phonetic forms. See, e.g., Stuart-Smith (2003) for discussion of variants in Scottish English.

REFERENCES

Abbs, J. H. (1986) Invariance and variability in speech production: A distinction between linguistic intent and its neuromotor implementation. In J. S. Perkell & D. H. Klatt (eds.), *Invariance and Variability in Speech Processes* (pp. 202–19). Hillsdale, NJ: Lawrence Erlbaum.

Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Adank, P., Smits, R., & Hout, R. van (2004) A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116, 3099–107.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003) Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 113, 544–52.

Al Shareef, J. (2002) Language change and variation in Palestine: A case study of Jabalia refugee camp. Doctoral dissertation, University of Leeds.

Anderson, B. L. (1999) Source-language transfer and vowel accommodation in the patterning of Cherokee English /ai/ and /oi/. *American Speech*, 74, 339–68.

Anttila, A. (1997) Deriving variation from grammar. In F. Hinskens, R. van Hout, & W. L. Wetzels (eds.), *Variation, Change and Phonological Theory* (pp. 35–68). Amsterdam: John Benjamins.

Ash, S. (2002) Social class. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *Handbook of Language Variation and Change* (pp. 402–220). Oxford: Blackwell.

Auer, P., Hinskens, F., & Kerswill, P. (eds.) (2005) *Dialect Change: Convergence and Divergence in European Languages*. Cambridge: Cambridge University Press.

Bates, R. A., Ostendorf, M., & Wright, R. A. (2007) Symbolic phonetic features for modeling of pronunciation variation. *Speech Communication*, 49, 83–97.

Baugh, J. (1996) Perceptions within a variable paradigm: Black and white racial detection and identification based on speech. In E. W. Schneider (ed.), *Focus on the USA* (pp. 169–82). Amsterdam: John Benjamins.

Beckett, D. (2003) Sociolinguistic individuality in a remnant dialect community. *Journal of English Linguistics*, 31, 3–33.

Bell, A. (1984) Language style as audience design. *Language in Society*, 13, 145–204.

Bell, A. (1991) *The Language of News Media*. Oxford: Blackwell.

Bergmann, P. (2006) Regional variation in intonation: Nuclear rising-falling contours in Cologne German. In F. Hinskens (ed.), *Language Variation: European Perspectives* (pp. 23–36). Amsterdam: John Benjamins.

Bezooijen, R. van (1988) The relative importance of pronunciation, prosody and voice quality for the attribution of social status and personality characteristics. In R. van Hout & U. Knops (eds.), *Language Attitudes in the Dutch Language Area* (pp. 85–103). Dordrecht: Foris.

Bezooijen, R. van (2005) Approximant /r/ in Dutch: Routes and feelings. *Speech Communication*, 47, 15–31.

Bezooijen, R. van & Gooskens, C. (1999) Identification of language varieties: The contribution of different linguistic levels. *Journal of Language and Social Psychology*, 18, 31–48.

Blatchford, H. & Foulkes, P. (2006) Identification of voices in shouting. *The International Journal of Speech, Language and the Law*, 13, 241–54.

Boberg, C. (2000) Geolinguistic diffusion and the U.S.–Canada border. *Language Variation and Change*, 12, 1–24.

Boberg, C. (2004) Ethnic patterns in the phonetics of Montreal English. *Journal of Sociolinguistics*, 8, 538–68.

Bond, Z. S., Stockmal, V., & Markus, D. (2006) Sixty years of bilingualism affects the pronunciation of Latvian vowels. *Language Variation and Change*, 18, 165–177.

Bowie, D. (2005) Language change over the lifespan: A test of the apparent time construct. *Penn Working Papers in Linguistics*, 11, 45–58.

Bradlow, A. R., Baker, R. E., Choi, A., Kim, M., & Van Engen, K. J. (2007) The Wildcat Corpus of Native- and Foreign-Accented English. *Journal of the Acoustical Society of America*, 121, 3072.

Britain, D. (1992) Linguistic change in intonation: The use of high rising terminals in New Zealand English. *Language Variation and Change*, 4, 77–103.

Britain, D. (1997) Dialect contact and phonological reallocation: "Canadian raising" in the English Fens. *Language in Society*, 26, 15–46.

Britain, D. (2002) Diffusion, levelling, simplification and reallocation in past tense BE in the English Fens. *Journal of Sociolinguistics*, 6, 16–43.

Brouwer, D. & Hout, R. van (1992) Gender-related variation in Amsterdam vernacular. *International Journal of the Sociology of Language*, 94, 99–122.

Bruce, G. & Gårding, E. (1978) A prosodic typology for Swedish dialects. In E. Gårding, G. Bruce, & R. Bannert (eds.), *Nordic Prosody* (pp. 219–28). Lund: Gleerup.

Brunelle, M. & Jannedy, S. (2007) Social effects on the perception of Vietnamese tone. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1461–4.

Bucholtz, M. B. (1998) Geek the Girl: Language, femininity, and female nerds. In N. Warner, J. Ahlers, L. Bilmes, M. Oliver, S. Wertheim, & M. Chen (eds.), *Gender and Belief Systems. Proceedings of the Fourth Berkeley Women and Language Conference* (pp. 119–31). Berkeley: Berkeley Women and Language Group.

Bucholtz, M. B. (1999) "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, 28, 203–23.

Bull, R. & Clifford, B. R. (1984) Earwitness voice recognition accuracy. In G. L. Wells & E. F. Loftus (eds.), *Eyewitness Testimony: Psychological Perspectives* (pp. 92–123). Cambridge: Cambridge University Press.

Bunin Benor, S. (2001) The learned /t/: Phonological variation in orthodox Jewish English. *Penn Working Papers in Linguistics*, 7, 1–16.

Bush, C. N. (1967) Some acoustic parameters of speech and their relationships to the perception of dialect differences. *TESOL Quarterly*, 1, 20–30.

Butler, J. P. (1990) *Gender Trouble: Feminism and the Subversion of Identity*. London: Routledge.

Bybee, J. (2001) *Phonology and Language Use*. Cambridge: Cambridge University Press.

Byrd, D. (1994) Relations of sex and dialect to reduction. *Speech Communication*, 15, 39–54.

Byrne, C. & Foulkes, P. (2004) The mobile phone effect on vowel formants. *The International Journal of Speech, Language and the Law*, 11, 83–102.

Cargile, A. C., Giles, H., Ryan, E. B., & Bradac, J. J. (1994) Language attitudes as a social process: A conceptual model and new direction. *Language and Communication*, 14, 211–36.

Carter, P. (2003) Extrinsic phonetic interpretation: Spectral variation in English liquids. In J. K. Local, R. A. Ogden, & R. A. M. Temple (eds.), *Papers in Laboratory Phonology VI: Phonetic Interpretation* (pp. 237–52). Cambridge: Cambridge University Press.

Carter, P. M. (2005) Prosodic variation in SLA: Rhythm in an urban North Carolina Hispanic community. *Penn Working Papers in Linguistics*, 11, 59–71.

Cedergren, H. & Perreault, H. (1995) On the analysis of syllable-timing in everyday speech. *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 4, 232–5.

Chambers, J. K. (2002) Studying language variation: An informal epistemology. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *Handbook of Language Variation and Change* (pp. 3–14). Oxford: Blackwell.

Chambers, J. K. & Trudgill, P. (1998) *Dialectology*, 2nd edn. Cambridge: Cambridge University Press.

Cheshire, J. (2002) Sex and gender in variationist research. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *Handbook of Language Variation and Change* (pp. 423–43). Oxford: Blackwell.

Cho, T. & Ladefoged, P. (1999) Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, 27, 207–29.

Cieri, C. (2005) Modeling phonological variation in multidialectal Italy. Doctoral dissertation, University of Pennsylvania.

Clopper, C. G., Levi, S. V., & Pisoni, D. B. (2006) Perceptual similarity of regional varieties of American English. *Journal of the Acoustical Society of America*, 119, 566–74.

Clopper, C. G. & Pisoni, D. B. (2004) Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32, 111–40.

Clyne, M., Eisikovits, E., & Tollfree, L. F. (2001) Ethnic varieties of Australian English. In D. Blair & P. Collins (eds.), *English in Australia* (pp. 223–38). Amsterdam: John Benjamins.

Coetzee, A. W. (2006) Variation as accessing "non-optimal" candidates. *Phonology*, 23, 337–85.

Coulmas, F. (2005) *Sociolinguistics: The Study of Speakers' Choices*. Cambridge: Cambridge University Press.

Coupland, N. (1980) Style-shifting in a Cardiff work-setting. *Language in Society*, 9, 1–12.

Coupland, N. & Bishop, H. (2007) Ideologised values for British accents. *Journal of Sociolinguistics*, 11, 74–93.

Cruttenden, A. (1997) *Intonation*, 2nd edn. Cambridge: Cambridge University Press.

Cucchiarini, C. & Heuvel, H. van den (1999) Postvocalic /r/-deletion in Dutch: More experimental evidence. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 3, 1673–6.

Dalton, M. & Ní Chasaide, A. (2003) Modelling intonation in three Irish dialects. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 1073–6.

Dalton, M. & Ní Chasaide, A. (2005) Tonal alignment in Irish dialects. *Language and Speech*, 48, 441–64.

Delvaux, V. & Soquet, A. (2007) Inducing imitative phonetic variation in the laboratory. *Proceedings of the 16th International Congress of Phonetic Sciences,* Saarbrücken, 369–72.

Deshaies-Lafontaine, D. (1974) A socio-phonetic study of a Québec French community: Trois-Rivières. Doctoral dissertation, University College London.

Di Paolo, M. & Faber, A. (1990) Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, 2, 155–204.

Dixon, J. A., Mahoney, B., & Cocks, R. (2002) Accents of guilt? Effects of regional accent, race, and criminal type on attributions of guilt. *Journal of Language and Social Psychology*, 21, 162–8.

Docherty, G. J. (2007) Speech in its natural environment: Accounting for social factors in phonetic variability. In J. Cole & J.-I. Hualde (eds.), *Laboratory Phonology IX* (pp. 1–35). Berlin: Mouton de Gruyter.

Docherty, G. J. & Foulkes, P. (1999) Newcastle upon Tyne and Derby: Instrumental phonetics and variationist studies. In P. Foulkes & G. J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 47–71). London: Arnold.

Docherty, G. J. & Foulkes, P. (2000) Speaker, speech, and knowledge of sounds. In N. Burton-Roberts, P. Carr, & G. J. Docherty (eds.), *Phonological Knowledge: Conceptual and Empirical Issues* (pp. 105–29). Oxford: Oxford University Press.

Docherty, G. J. & Foulkes, P. (2005) Glottal variants of (t) in the Tyneside variety of English: An acoustic profiling study. In W. J. Hardcastle & J. Mackenzie Beck (eds.), *A Figure of Speech: A Festschrift for John* Laver (pp. 173–99). London: Lawrence Erlbaum.

Docherty, G. J., Foulkes, P., Tillotson, J., & Watt, D. (2006) On the scope of phonological learning: Issues arising from socially structured variation. In L. Goldstein, D. H. Whalen, & C. T. Best (eds.), *Laboratory Phonology VIII* (pp. 393–421). Berlin: Mouton.

Docherty, G. J. & Khattab, G. (2008) Sociophonetic issues in speech impairment. In M. Ball, M. Perkins, N. Müller, & S. Howard. *The Handbook of Clinical Linguistics* (pp. 603–25). Oxford: Blackwell.

Docherty, G. J. & Watt, D. (2001) Chain shifts. In R. Mesthrie (ed.), *The Concise Encyclopedia of Sociolinguistics* (pp. 303–7). Amsterdam: Pergamon.

Dressler, W. U. & Wodak, R. (1982) Sociophonological methods in the study of sociolinguistic variation in Viennese German. *Language in Society*, 11, 339–70.

Dubois, S. & Horvath, B. (1998) Let's tink about dat: Interdental fricatives in Cajun English. *Language Variation and Change*, 10, 245–61.

Dubois, S. & Horvath, B. (1999) When the music changes, you change too: Gender and language change in Cajun English. *Language Variation and Change*, 11, 287–313.

Dunn, M. (2000) Chukchi women's language: A historical-comparative perspective. *Anthropological Linguistics*, 42, 305–28.

Dyer, J. M. (2002) "We all speak the same round here": Dialect levelling in a Scottish-English community. *Journal of Sociolinguistics*, 6, 99–116.

Eckert, P. (1997) Age as a sociolinguistic variable. In F. Coulmas (ed.), *Handbook of Sociolinguistics* (pp. 151–67). Oxford: Blackwell.

Eckert, P. (2000) *Linguistic Variation as Social Practice*. Oxford: Blackwell.

Eckert, P. & McConnell-Ginet, S. (1999) New generalizations and explanations in language and gender research. *Language in Society*, 28, 185–202.

Ellis, S. (1994) The Yorkshire Ripper enquiry: Part 1. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 1, 197–206.

Elman, J. L., Diehl, R. L., & Buchwald, S. E. (1977) Perceptual switching in bilinguals. *Journal of the Acoustical Society of America*, 62, 971–4.

Elman, J. L. & McClelland, J. L. (1986) Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (eds.), *Invariance and Variability in Speech Processes* (pp. 360–85). Hillsdale, NJ: Lawrence Erlbaum.

Elzinga, D. & Di Paolo, M. (forthcoming) Shoshoni. In M. Di Paolo & A. K. Spears (eds.), *Increasing Language Diversity in Linguistics Courses: Practical Approaches and Materials*. Columbus, OH: Ohio State University Press.

Esling, J. H. (1991) Sociophonetic variation in Vancouver. In J. Cheshire (ed.), *English Around the World* (pp. 123–33). Cambridge: Cambridge University Press.

Evans, B. & Iverson, P. (2007) Plasticity in vowel perception and production: A study of accent change in young adults. *Journal of the Acoustical Society of America*, 121, 3814–26.

Evans, B., Mistry, A., & Moreiras, C. (2007) An acoustic study of first- and second-generation Gujarati immigrants in Wembley: Evidence for accent convergence? *Proceedings of the 16th*

International Congress of Phonetic Sciences, Saarbrücken, 1741–4.

Faber, A. & Di Paolo, M. (1995) The discriminability of nearly merged sounds. *Language Variation and Change*, 7, 35–78.

Fabricius, A. (2007) Variation and change in the TRAP and STRUT vowels of RP: A real time comparison of five acoustic data sets. *Journal of the International Phonetic Association*, 37, 293–320.

Ferguson, Charles A. (1959) Diglossia. *Word*, 15, 325–40.

Fitt, S. & Isard, S. (1999) Synthesis of regional English using a keyword lexicon. *Proceedings of Eurospeech 99*, 2, 823–6.

Fletcher, J., Grabe, E., & Warren, P. (2005) Intonational variation in four dialects of English: The high rising tune. In S.-A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 390–409). Oxford: Oxford University Press.

Fought, C. (1999) A majority sound change in a minority community: /u/-fronting in Chicano English. *Journal of Sociolinguistics*, 3, 5–23.

Fought, C. (2002) Ethnicity. In J. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *Handbook of Language Variation and Change* (pp. 444–72). Oxford: Blackwell.

Fought, C. (2003) *Chicano English in Context*. Basingstoke: Palgrave.

Foulkes, P., & Docherty, G. J. (2000) Another chapter in the story of /r/: "Labiodental" variants in British English. *Journal of Sociolinguistics*, 4, 30–59.

Foulkes, P. & Docherty, G. J. (2006) The social life of phonetics and phonology. *Journal of Phonetics*, 34, 409–38.

Foulkes, P., Docherty, G. J., Khattab, G., & Yaeger-Dror, M. (in press) Sound judgements: Perception of indexical features in children's speech. In D. Preston & N. Niedzielski (eds.), *A Reader in Sociophonetics*. Berlin: Mouton.

Foulkes, P., Docherty, G. J., & Watt, D. (2005) Phonological variation in child-directed speech. *Language*, 81, 177–206.

Fourakis, M. & Port, R. (1986) Stop epenthesis in English. *Journal of Phonetics*, 14, 197–221.

Freeman, E. B. (1982) The Ann Arbor decision: The importance of teachers' attitudes towards language. *The Elementary School Journal*, 83, 40–7.

French, J. P., Harrison, P., & Windsor Lewis, J. (2006) Case report: R v. John Samuel Humble: The Yorkshire Ripper Hoaxer Trial. *International Journal of Speech, Language and the Law*, 13, 255–73.

Fridland, V. (2003) "Tie, tied and tight": The expansion of /ai/ monophthongization in African-American and European-American speech in Memphis, Tennessee. *Journal of Sociolinguistics*, 7, 279–98.

Gick, B. & Wilson, I. (2006) Excrescent schwa and vowel laxing: Cross-linguistic responses to conflicting articulatory targets. In L. Goldstein, D. H. Whalen, & C. T. Best (eds.), *Laboratory Phonology VIII* (pp. 635–60). Berlin: Mouton.

Giles, H. & Powesland, P. F. (1975) *Speech Style and Social Evaluation*. New York: Academic Press.

Glenn, M. G. & Strassel, S. (2006) Linguistic resources for meeting speech recognition. In S. Renals & S. Bengio (eds.), *Machine Learning for Multimodal Interaction* (pp. 390–401). Berlin: Springer.

Gobl, C. (1988) Voice source dynamics in connected speech. *KTH Speech Technology Laboratory Quarterly Progress and Status Report*, 1/1988, 123–59.

Gordon, M. & Heath, J. (1998) Sex, sound symbolism and sociolinguistics. *Current Anthropology*, 39, 421–49.

Gordon, M., Munro, P., & Ladefoged, P. (2000) Some phonetic structures of Chickasaw. *Anthropological Linguistics*, 42, 366–400.

Grabe, E. (2004) Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles & J. Peters (eds.), *Regional Variation in Intonation* (pp. 9–31). Tübingen: Niemeyer.

Grabe, E. & Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (eds.), *Papers in Laboratory Phonology VII* (pp. 515–46). Berlin: Mouton.

Grabe, E., Post, B., Nolan, F. J., & Farrar, K. (2000) Pitch accent realization in four varieties of British English. *Journal of Phonetics*, 28, 161–85.

Graff, D., Labov, W., & Harris, W. A. (1986) Testing listeners' reactions to phonological markers of ethnic identity: A new method for sociolinguistic research. In D. Sankoff (ed.), *Diversity and Diachrony* (pp. 45–58). Amsterdam: John Benjamins.

Green, L. J. (2002) *African American English: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Gumperz, J. (1982) *Discourse Strategies*. Cambridge: Cambridge University Press.

Haagen, M. van der (1998) *Caught between Norms: The English Pronunciation of Dutch Learners*. The Hague: Holland Academic Graphics.

Haeri, N. (2003) *Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt*. London: Palgrave Macmillan.

Hall, K. & Bucholtz, M. (eds.) (1995) *Gender Articulated: Language and the Socially Constructed Self*. New York: Routledge.

Halle, M. (1963) Phonemics. In T. A. Sebeok (ed.), *Current Trends in Linguistics*, vol. 1: *Soviet and East European Linguistics* (pp. 5–21). The Hague: Mouton.

Hardcastle, W. J. & Barry, W. (1989) Articulatory and perceptual factors in /l/ vocalisation in English. *Journal of the International Phonetic Association*, 15, 3–17.

Harrington, J., Palethorpe, S., & Watson, C. I. (2005) Deepening or lessening the divide between diphthongs? An analysis of the Queen's annual Christmas Broadcasts. In W. J. Hardcastle & J. M. Beck (eds.), *A Figure of Speech: A Festchrift for John Laver* (pp. 227–62). Hillsdale, NJ: Lawrence Erlbaum.

Harrison, P. (2004) Variability of formant measurements. MA dissertation, University of York. (www. jpfrench. com/docs/harrison-formant-dissertation.pdf)

Hawkins, S. (2003) Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, 373–405.

Hawkins, S. & Midgley, J. (2005) Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, 35, 183–99.

Hawkins, S. & Smith, R. (2001) Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Rivista di Linguistica*, 13, 99–188.

Hay, J. & Drager, K. (2007) Sociophonetics. *Annual Review of Anthropology*, 36, 89–103.

Hay, J., Nolan, A., & Drager, K. (2006) From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review*, 23, 351–79.

Hay, J., Warren, P., & Drager, K. (2006) Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34, 458–84.

Heffernan, K. (2006) Prosodic levelling during language shift: Okinawan approximations of Japanese pitch-accent. *Journal of Sociolinguistics*, 10, 641–66.

Henton, C. & Bladon, A. (1988) Creak as a sociophonetic marker. In L. Hyman and C. N. Li (eds.), *Language, Speech and Mind: Studies in Honor of Victoria A. Fromkin* (pp. 3–29). London: Routledge.

Heselwood, B. (2007) The "tight approximant" variant of the Arabic 'ayn. *Journal of the International Phonetic Association*, 37, 1–32.

Heselwood, B. & McChrystal, L. (2000) Gender, accent features and voicing in

Panjabi-English bilingual children. *Leeds Working Papers in Linguistics and Phonetics*, 8, 45–70.

Hewlett, N., Matthews, B., & Scobbie, J. M. (1999) Vowel duration in Scottish English speaking children. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 2157–60.

Hildebrandt, K. A. (2007) Tone in Bodish languages: Typological and sociolinguistic contributions. In M. Miestamo & B. Wälchli (eds.), *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations* (pp. 67–90). Berlin: Mouton de Gruyter.

Hinton, L. N. & Pollock, K. E. (2007) Regional variations in the phonological characteristics of African American Vernacular English. *World Englishes*, 19, 59–71.

Hirson A. & N. Sohail (2007) Variability of rhotics in Punjabi-English bilinguals. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1501–4.

Hockett, C. F. (1965) Sound change. *Language*, 41, 185–204.

Hodgson, P. & Miller, J. L. (1996) Internal structure of phonetic categories: Evidence for within-category trading relations. *Journal of the Acoustical Society of America*, 100, 565–76.

Hoequist, C. & Nolan, F. J. (1991) On an application of phonological knowledge in automatic speech recognition. *Computer Speech and Language*, 5, 133–53.

Holmes, J. (1997) Maori and Pakeha English: Some New Zealand social dialect data. *Language in Society*, 26, 65–101.

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979) Phonetic explanations for the development of tones. *Language*, 55, 37–58.

Horvath, B. (1985) *Variation in Australian English*. Cambridge: Cambridge University Press.

Howard, S. & Heselwood, B. (2002) The contribution of phonetics to the study of vowel development and disorders. In M. J. Ball & F. Gibbon (eds.), *Vowel Disorders* (pp. 37–82). Woburn, MA: Butterworth/Heinemann.

Hubbell, A. F. (1950) *The Pronunciation of English in New York City*. New York: King's Crown Press, Columbia University.

Jacewicz, E., Fox, R. A., & Salmons, J. (2007) Vowel space areas across dialects and gender. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1465–8.

Janson, T. (1983) Sound change in perception and production. *Language*, 59, 18–34.

Janson, T. & Schulman, R. (1983) Non-distinctive features and their use. *Journal of Linguistics*, 19, 321–36.

Johnson, K. (2006) Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34, 485–99.

Johnson, K., Ladefoged, P., & Lindau, M. (1993) Individual differences in vowel production. *Journal of the Acoustical Society of America*, 94, 701–14.

Johnson, K., Strand, E., & D'Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27, 359–84.

Johnstone, B. & Bean, J. M. (1997) Self-expression and linguistic variation. *Language in Society*, 26, 221–46.

Joos, M. (1948) Acoustic phonetics. *Language*, 24, 5–131, 133–6.

Keane, E. (2006) Rhythmic characteristics of colloquial and formal Tamil. *Language and Speech*, 49, 299–332.

Kerswill, P. (1985) A sociophonetic study of connected speech processes in Cambridge English: An outline and some results. *Cambridge Papers in Phonetics and Experimental Linguistics*, 4, 1–39.

Kerswill, P. & Wright, S. (1990) On the limits of auditory transcription: A sociophonetic perspective. *Language Variation and Change*, 2, 255–75.

Kerswill, P. & Williams, A. (2000) Creating a new town koine: Children and language change in Milton Keynes. *Language in Society*, 29, 65–115.

Khattab, G. (2006) Phonological acquisition in Arabic-English bilingual children. In Z. Hua & B. Dodd (eds.), *Phonological Development and Disorders: A Cross-Linguistic Perspective* (pp. 383–412). Clevedon, UK: Multilingual Matters.

Khattab, G. (2007) Variation in vowel production by English-Arabic bilinguals. In J. Cole & J.-I. Hualde (eds.), *Laboratory Phonology IX* (pp. 383–409). Berlin: Mouton.

Khattab, G. (2009) Phonetic accommodation in children's code-switching. In B. E. Bullock & J. Toribio Almeida (eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp. 142–59). Cambridge: Cambridge University Press.

Khattab, G., Al-Tamimi, F., & Heselwood, B. (2006) Acoustic and auditory differences in the /t-T/ opposition in male and female speakers of Jordanian Arabic. In S. Boudela (ed.), *Perspectives on Arabic Linguistics XIV* (pp. 131–60). Amsterdam: John Benjamins.

Kiesling, S. F. (1998) Variation and men's identity in a fraternity. *Journal of Sociolinguistics*, 2, 69–100.

Kim, Y. (2005) On the phonetics of unstressed /e/ in Stockholm Swedish and Finland Swedish. *Proceedings of FONETIK 2005*, Department of Linguistics, Göteborg University.

Kissine, M., Velde, H. Van de, & Hout, R. van (2003) An acoustic study of standard Dutch /v/, /f/, /s/ and /z/. In L. Cornips & P. Fikkert (eds.), *Linguistics in the Netherlands 2003* (pp. 93–04). Amsterdam: John Benjamins.

Kohler, K. J. (1990) Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 69–92). Dordrecht: Kluwer.

Künzel, H. J. (2001) Beware of the "telephone effect": The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 8, 80–99.

Labov, W. (1963) The social motivation of a sound change. *Word*, 19, 273–309.

Labov, W. (1966a) The linguistic variable as a structural unit. *Washington Linguistics Review*, 3, 4–22.

Labov, W. (1966b) *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, W. (1972) *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1994) *Principles of Linguistic Change*, vol. 1: *Internal Factors*. Oxford: Blackwell.

Labov, W. (2001) *Principles of Linguistic Change*, vol. 2: *Social Factors*. Oxford: Blackwell.

Labov, W. (forthcoming) *Principles of Linguistic Change*, vol. 3: *Cognitive Factors*. Oxford: Blackwell.

Labov, W. (2006) A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34, 500–15.

Labov, W., Ash, S., & Boberg, C. (2006) *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Berlin: Mouton.

Labov, W. & Baranowski, M. (2006) 50 msec. *Language Variation and Change*, 18, 223–40.

Labov, W., Cohen, P., Robins, C., & Lewis, J. (1968) *A Study of the Non-standard English of Negro and Puerto Rican Speakers in New York City*. New York: Columbia University Press.

Labov, W., Karen, M., & Miller, C. (1991) Near-mergers and the suspension of phonemic contrast. *Language Variation and Change*, 3, 33–74.

Labov, W., Yaeger, M., & Steiner, R. (1972) *A Quantitative Study of Sound Change in Progress*. Philadelphia: US Regional Survey.

Lachs, L., McMichael, K., & Pisoni, D. (2002) Speech perception and implicit memory: Evidence for detailed episodic encoding. In J. S. Bowers & C. J. Marsolek (eds.), *Rethinking Implicit Memory* (pp. 215–35). Oxford: Oxford University Press.

Ladefoged, P. (2003) *Phonetic Data Analysis*. Oxford: Blackwell.

Lambert, W. C., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960) Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology*, 60, 44–51.

Lambert, K., Alam, F., & Stuart-Smith, J. (2007) Investigating British Asian accents: Studies from Glasgow. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1509–12.

Lass, N. J., Almerino, C. A., Jordan, L. F., & Walsh, J. M. (1980) The effect of filtered speech on speaker race and sex identifications. *Journal of Phonetics*, 8, 101–12.

Lass, N. J., Tecca, J. E., Mancuso, R. A., & Black, W. I. (1979) The effect of phonetic complexity on speaker race and sex identifications. *Journal of Phonetics*, 7, 105–18.

Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (1995) Voice types in automated telecommunications applications. In J. Windsor Lewis (ed.), *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor* (pp. 85–95). London: Routledge.

Lee, A., Hewlett, N., & Nairn, M. (1995) Voice and gender in children. In S. Mills (ed.), *Language and Gender: Interdisciplinary Perspectives* (pp. 194–204). London: Longman.

Levelt, W. J. M. (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Leyden, K. van (2004) *Prosodic Characteristics of Orkney and Shetland Dialects: An Experimental Approach*. Utrecht: LOT.

Lindblom, B. (1986) On the origin and purpose of discreteness and invariance in sound patterns. In J. S. Perkell & D. H. Klatt (eds.), *Invariance and Variability in Speech Processes* (pp. 493–510). Hillsdale, NJ: Lawrence Erlbaum.

Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 403–39). Dordrecht: Kluwer.

Lippi-Green, R. (1997) *English With an Accent*. London: Routledge.

Lisker, L. & Abramson, A. S. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422.

Livijn, P. (2002) Distribution of dental and retroflex l-sounds across some Swedish dialects. *Fonetik 2002, the XVth Swedish Phonetics Conference, Stockholm, May 29–31, 2002*, Quarterly Progress and Status Report, Centre for Speech Technology, KTH Stockholm, 25–8.

Llamas, C. (2007) "A place between places": Language and identities in a border town. *Language in Society*, 36, 579–604.

Local, J. (2003) Variable domains and variable relevance: interpreting phonetic exponents. *Journal of Phonetics*, 31, 321–39.

Local, J. (2007) Phonetic detail and the organisation of talk-in-interaction. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1–10.

Long, D. & Preston, D. R. (eds.) (2002) *Handbook of Perceptual Dialectology*, vol. 2. Amsterdam: John Benjamins.

Macaulay, R. K. S. (1991) *Locating Dialect in Discourse: The Language of Honest Men and Bonnie Lasses in Ayr*. Oxford: Oxford University Press.

Majors, T. (2005) Low back vowel merger in Missouri speech: Acoustic description and explanation. *American Speech*, 80, 165–79.

Martinet, A. (1955) *Économie des changements phonétiques*. Bern: Francke.

McCafferty, K. (1999) (London)Derry: Between Ulster and local speech – class, ethnicity and language change. In P. Foulkes & G. J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 246–64). London: Arnold.

McCafferty, K. (2001) *Ethnicity and Language Change.* Amsterdam: John Benjamins.

McConnell-Ginet, S. (1983) Intonation in a man's world. In B. Thorne, C. Kramarae, & N. Henley (eds.), *Language, Gender and Society* (pp. 69–88). Rowley, MA: Newbury House.

McDougall, K. (2004) Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 11, 103–30.

McLeod, H. & Jongman, A. (1993) Categorical perception of silent-center syllables. *Journal of the Acoustical Society of America*, 119, 2427–37.

McLeod, S. (ed.) (2006) *The International Guide to Speech Acquisition*. Clifton Park, NY: Thomson Delmar Learning.

McMahon, A. M. S. (1994) *Understanding Language Change*. Cambridge: Cambridge University Press.

Mees, I. M. & Collins, B. (1999) Cardiff: A real-time study of glottalization. In P. Foulkes & G. J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 185–202). London: Arnold.

Mendoza-Denton, N. (1996) Language ideology and gang affiliation among California Latina girls. In M. Bucholtz, A. C. Liang, L. A. Sutton, & C. Hines (eds.), *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference* (pp. 478–86). Berkeley, CA: University of California Press.

Mendoza-Denton, N. (1999) Fighting words: Latina girls, gangs, and language attitudes. In D. L. Galindo & M. D. Gonzales (eds.), *Speaking Chicana: Voice, Power and Identity* (pp. 39–56). Tucson, AZ: University of Arizona Press.

Mendoza-Denton, N., Hay, J., & Jannedy, S. (2003) Probabilistic sociolinguistics: Beyond variable rules. In R. Bod, J. Hay, & S. Jannedy (eds.), *Probabilistic Linguistics* (pp. 97–138). Cambridge, MA: MIT Press.

Meyerhoff, M. (2002) Communities of practice. In J. K. Chambers, N. Schilling-Estes, & P. Trudgill (eds.), *Handbook of Language Variation and Change* (pp. 526–48). Oxford: Blackwell.

Mielke, J., Baker, A., & Archangeli, D. (forthcoming) Variability and homogeneity in American English /ɹ/ allophony and /s/ retraction. In C. Fougeron & M. D'Imperio (eds.), *Laboratory Phonology X: Variation, Detail, and Representation*. Berlin: Mouton.

Milroy, J. (1992) *Linguistic Variation and Change*. Oxford: Blackwell.

Milroy, J. (2001) Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, 5, 530–55.

Milroy, J. & Milroy, L. (1998) *Authority in Language*, 3rd edn. London: Routledge.

Milroy, L. (1987a) *Observing and Analysing Natural Language*. Oxford: Blackwell.

Milroy, L. (1987b) *Language and Social Networks*, 2nd edn. Oxford: Blackwell.

Milroy, L. & Gordon, M. (2003) *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.

Mirchandani, K. (2004) Practices of global capital: Gaps, cracks and ironies in transnational call centres in India. *Global Networks*, 4, 355–73.

Mitterer, H. & Ernestus, M. (2008) The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, 109, 168–73.

Moosmüller, S. & Granser, T. (2006) The spread of Standard Albanian: An illustration based on an analysis of vowels. *Language Variation and Change*, 18, 121–40.

Mufwene, S. (2001) *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Mufwene, S. S., Rickford, J. R., Bailey, G., & Baugh, J. (eds.) (1998) *African-American English: Structure, History and Use*. London: Routledge.

Munro, M. J., Derwing, T. M., & Flege, J. E. (1999) Canadians in Alabama: A perceptual study of dialect acquisition in adults. *Journal of Phonetics*, 27, 385–403.

Munson, B. (2007) The acoustic correlates of perceived sexual orientation, perceived masculinity, and perceived femininity. *Language and Speech*, 50, 125–42.

Munson, B., Jefferson, S. V., & McDonald, E. C. (2006) The influence of perceived sexual orientation on fricative identification. *Journal of the Acoustical Society of America*, 119, 2427–37.

Nagy, N. & Reynolds, B. (1997) Optimality Theory and variable word-final deletion in Faetar. *Language Variation and Change*, 9, 37–55.

Nahkola, K. & Saanilahti, M. (2004) Mapping language changes in real time: A panel study on Finnish. *Language Variation and Change*, 16, 75–92.

Nathan, E., Wells, W. H. G., & Donlan, C. (1998) Children's comprehension of unfamilar regional accents: A preliminary investigation. *Journal of Child Language*, 25, 343–65.

Nicolaidis, K. (2001) An electropalatographic study of Greek spontaneous speech. *Journal of the International Phonetic Association*, 31, 67–85.

Niedzielski, N. (1999) The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18, 62–85.

Nittrouer, S. (2005) Perception of steady-state vowels and vowelless syllables by adults and children. *Journal of the Acoustical Society of America*, 117, S2402.

Nolan, F. and Farrar, K. (1999) Timing of f0 peaks and peak lag. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 961–7.

Nolan, F. & Grigoras, C. (2005) A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12, 143–73.

Nolan, F. J. & Kerswill, P. E. (1990) The description of connected speech processes. In S. Ramsaran (ed.), *Studies in the Pronunciation of English. A Commemorative Volume in Honour of A. C. Gimson* (pp. 295–316). London: Routledge.

Nygaard, L. C. (2005) Perceptual integration of linguistic and nonlinguistic properties of speech. In D. B. Pisoni & R. E. Remez (eds.), *The Handbook of Speech Perception* (pp. 390–413). Oxford: Blackwell.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994) Speech perception as a talker-contingent process. *Psychological Science*, 5, 42–6.

Oetting, J. B. (2005) Assessing language in children who speak a nonmainstream dialect of English. In M. J. Ball (ed.), *Clinical Sociolinguistics* (pp. 180–92). Oxford: Blackwell.

Ogden, R. (2004) Non-modal voice quality and turn-taking in Finnish. In E. Couper-Kuhlen & C. Ford (eds.), *Sound Patterns in Interaction* (pp. 29–62). Amsterdam: John Benjamins.

Ogden, R. & Routarinne, S. (2005) The communicative functions of final rises in Finnish intonation. *Phonetica*, 62, 160–75.

Ohala, J. J. (1983) The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (ed.), *The Production of Speech* (pp. 189–216). New York: Springer.

Ohala, J. J. (1989) Sound change is drawn from a pool of synchronic variation. In L. Breivik & H. Jahr (eds.), *Language Change* (pp. 173–98). Berlin: Mouton.

Orr, S. (2007) A sociophonetic study of speech and interaction in a Glaswegian call centre. Doctoral dissertation, University of Glasgow.

Pappas, P. A. (2006) Stereotypes and /n/ variation in Patra, Greece. In F. Hinskens (ed.), *Language Variation: European Perspectives* (pp. 153–68), Studies in Language Variation 1. Amsterdam: John Benjamins.

Pear, T. H. (1931) *Voice and Personality as Applied to Radio Broadcasting*. New York: Wiley.

Peterson, G. E. & Barney, H. L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175–84.

Pierrehumbert, J. B. (2002) Word-specific phonetics. In C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology VII* (pp. 101–39). Berlin: Mouton.

Pierrehumbert, J. B. (2006) The next toolkit. *Journal of Phonetics*, 34, 516–30.

Pierrehumbert, J. B., Beckman, M., & Ladd, D. R. (2000) Conceptual foundations of phonology as a laboratory science. In N. Burton-Roberts, P. Carr, & G. J. Docherty (eds.), *Phonological Knowledge: Conceptual and Empirical Issues* (pp. 273–303). Oxford: Oxford University Press.

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004) The influence of sexual orientation on vowel production. *Journal of the Acoustical Society of America*, 116, 1905–8.

Piroth, H. G. & Janker, P. M. (2004) Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics*, 32, 81–109.

Pisoni, D. B. (1997) Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (eds.), *Talker Variability in Speech Processing* (pp. 9–32). San Diego: Academic Press.

Plichta, B. (2004) Best practices in the acquisition, processing, and analysis of acoustic speech signals. www.historicalvoices.org/flint/extras/Audio-technology.pdf.

Plug, L. (2005) From words to actions: The phonetics of *eigenlijk* in two communicative contexts. *Phonetica*, 62, 131–45.

Podesva, R. J. (2007) Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, 11, 478–504.

Preston, D. R. (ed.) (1999) *Handbook of Perceptual Dialectology*, vol. 1. Amsterdam: John Benjamins.

Purnell, T., Idsardi, W., & Baugh, J. (1999) Perceptual and phonetic experiments in American English dialect identification. *Journal of Language and Social Psychology*, 18, 10–30.

Ramus, F., Nespor, M., & Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265–92.

Rand, D. & Sankoff, D. (1990) Goldvarb 2.1: A variable rule application for the Macintosh. Montréal: Centre de Recherches Mathématiques, Université de Montréal. (www.crm.umontreal.ca/~sankoff/GoldVarb_Eng.html)

Raymond, W. D., Dautricourt, R., & Hume, E. (2006) Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, 18, 55–97.

Reid, E. (1978) Social and stylistic variation in the speech of children: Some evidence from Edinburgh. In P. Trudgill (ed.), *Sociolinguistic Patterns in British English* (pp. 151–71). London: Arnold.

Repp, B. H. & Liberman, A. M. (1987) Phonetic categories are flexible. In S. Harnad (ed.), *Categorical Perception* (pp. 89–112). Cambridge: Cambridge University Press.

Rietveld, T. & Hout, R. van (1993) *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin: Mouton.

Roberts, J. (1997) Hitting a moving target: Acquisition of sound change in progress by Philadelphia children. *Language Variation and Change*, 9, 249–66.

Romaine, S. (1978) Post-vocalic /r/ in Scottish English: Sound change in progress? In P. Trudgill (ed.), *Sociolinguistic Patterns in British English* (pp. 144–58). London: Arnold.

Rosner, B. S. & Pickering, J. B. (1994) *Vowel Perception and Production*. Oxford: Oxford University Press.

Ryback-Soucy, W. & Nagy, N. (2000) Exploring the dialect of Franco-Americans of Manchester, New Hampshire. *Journal of English Linguistics*, 28, 249–64.

Sachs, J. (1975) Cues to the identification of sex in children's speech. In B. Thorne & N. Henley (eds.), *Language and Sex: Difference and Domination* (pp. 152–71). Rowley, MA: Newbury House.

Sankoff, G., Blondeau, H., & Charity, A. (2001) Individual roles in a real-time change: Montreal (r → R) 1947–1995. *Etudes and Travaux*, 4, 141–57.

Sankoff, G. & Blondeau, H. (2008) Language change across the lifespan: /r/ in Montreal French. *Language*, 83, 560–88.

Schilling-Estes, N. (2000) Investigating intra-ethnic differentiation: /ay/ in Lumbee Native American English. *Language Variation and Change*, 12, 141–74.

Scobbie, J. M. (2005) Interspeaker variation among Shetland Islanders as the long term outcome of dialectally varied input: Speech production evidence for fine-grained linguistic plasticity. *QMUC Speech Science Research Centre Working Paper WP2.*

Scobbie, J. M. (2006a) Flexibility in the face of incompatible English VOT systems. In L. Goldstein, D. H. Whalen, & C. T. Best (eds.), *Laboratory Phonology VIII: Varieties of Phonological Competence* (pp. 367–92). Berlin: Mouton.

Scobbie, J. M. (2006b) (R) as a variable. In K. Brown (editor-in-chief), *The Encyclopaedia of Language and Linguistics*, 2nd edn., vol. 10 (pp. 337–44). Oxford: Elsevier.

Scobbie, J. M. (2007a) Biological and social grounding of phonology: Variation as a research tool. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 225–8.

Scobbie, J. M. (2007b) Interface and overlap in phonetics and phonology. In G. Ramchand and C. Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces* (pp. 17–52). Oxford: Oxford University Press.

Scobbie, J. M., Pouplier, M., & Wrench, A. A. (2007) Conditioning factors in external sandhi: An EPG study of English /l/ vocalisation. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 441–4.

Scobbie, J. M., Sebregts, K., & Stuart-Smith, J. (in preparation) Acoustic, articulatory and phonological perspectives on rhoticity and /r/ in Dutch.

Scobbie, J. M. & Stuart-Smith, J. (2005) Ongoing variation and change in Glasgow liquids: A pilot ultrasound study. Paper presented at the Fifth UK Language Variation and Change Conference (UKLVC 5), University of Aberdeen, 12–14 September.

Scobbie, J. M. & Stuart-Smith, J. (2008) Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. In P. Avery, E. B. Dresher, & K. Rice (eds.), *Contrast: Perception and Acquisition: Selected papers from the Second International Conference on Contrast in Phonology* (pp. 87–113). Berlin: Mouton.

Scobbie, J. M., Stuart-Smith, J., & Lawson, E. (2008) *Final report on ESRC Grant RES000222032: Looking Variation and Change in the Mouth: Developing the Sociolinguistic Potential of Ultrasound Tongue Imaging*.

Scobbie, J. M., Turk, A. E., & Hewlett, N. (1999) Morphemes, phonetics and lexical items: The case of the Scottish Vowel Length Rule. *Proceedings of the*

14th International Congress of Phonetic Sciences, San Francisco, 1617–20.

Scobbie, J. M. & Wrench, A. A. (2003) An articulatory investigation of word final /l/ and /l/-sandhi in three dialects of English. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 1871–4.

Scobbie, J. M., Wrench, A. A., & Linden, M. van der (2008) Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. *Proceedings of 8th International Seminar on Speech Production*, Strasbourg.

Sebastian, R. J. & Ryan, E. B. (1985) Speech cues and social evaluation: Markers of ethnicity, social class, and age. In H. Giles & R. N. St Clair (eds.), *Recent Advances in Language, Communication, and Social Psychology* (pp. 112–43). London: Lawrence Erlbaum.

Sederholm, E. (1998) Perception of gender in ten-year-old children's voices. *Logopedics Phoniatrics Vocology*, 23, 65–8.

Selting, M. (2004) Dresden *Fallbogen* contours as an example of regionalized German intonation. *Canadian Journal of Linguistics*, 49, 289–326.

Silverstein, M. (2003) Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, 23, 193–229.

Simpson, A. P. & Ericsdotter, C. (2007) Sex-specific differences in f_0 and vowel space. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 933–6.

Smith, J., Durham, M., & Fortune, L. (2007) "Mam, my trousers is fa'in doon!": Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation and Change*, 19, 63–99.

Stanford, J. (2007) Dialect contact and identity: A case study of exogamous Sui clans. Doctoral dissertation, Michigan State University.

Stevens, K. N. (1998) *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Strand, E. (1999) Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, 18, 86–99.

Stuart-Smith, J. (1999) Glasgow: Accent and voice quality. In P. Foulkes & G. J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 203–22). London: Arnold.

Stuart-Smith, J. (2003) The phonology of Modern Urban Scots. In J. Corbett, J. D. McClure, & J. Stuart-Smith (eds.), *The Edinburgh Companion to Scots* (pp. 110–37). Edinburgh: Edinburgh University Press.

Stuart-Smith, J. (2007a) A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1449–52.

Stuart-Smith, J. (2007b) Empirical evidence for gendered speech production: /s/ in Glaswegian. In J. Cole & J.-I. Hualde (eds.), *Laboratory Phonology IX* (pp. 65–86). Berlin: Mouton.

Stuart-Smith, J., Timmins, C., & Tweedie, F. (2007) "Talkin' Jockney"? Variation and change in Glaswegian accent. *Journal of Sociolinguistics*, 11, 221–60.

Suomi, K. (2005) Temporal conspiracies for a tonal end: Segmental durations and accentual f0 movement in a quantity language. *Journal of Phonetics*, 33, 291–309.

Syrdal, A. (1996) Acoustic variability in spontaneous conversational speech of American English talkers. In H. T. Bunnell & W. Idsardi (eds.), *Proceedings of ICSLP96* (4th International Conference on Spoken Language Processing, University of Delaware), 1, 438–41.

Tagliamonte, S. (2006) *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Taylor, A. (1982) "Male" and "female" speech in Gros Ventre. *Anthropological Linguistics*, 24, 301–7.

Thomas, E. R. (2000) Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. *Journal of Phonetics*, 28, 1–26.

Thomas, E. R. (2001) *An Acoustic Analysis of Vowel Variation in New World English*. Durham, NC: Duke University Press.

Thomas, E. R. (2002a) Sociophonetic applications of speech perception experiments. *American Speech*, 77, 115–47.

Thomas, E. R. (2002b) Instrumental phonetics. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.), *Handbook of Language Variation and Change* (pp. 168–200). Oxford: Blackwell.

Trent, S. A. (1995) Voice quality: Listener identification of African-American versus Caucasian speakers. *Journal of the Acoustical Society of America*, 98, 2936.

Trudgill, P. (1974) *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Trudgill, P. (1988) Norwich revisited: Recent linguistic changes in an English urban dialect. *English World-Wide*, 9, 3–49.

Trudgill, P. (2004) *New-Dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.

Trudgill, P., Gordon, E., Lewis, G., & Maclagan, M. (2000) Determinism in new-dialect formation and the genesis of New Zealand English. *Journal of Linguistics*, 36, 299–318.

Turk, A. E. & Shattuck-Hufnagel, S. (2000) Word-boundary-related duration patterns in English. *Journal of Phonetics*, 28, 397–440.

Vaissière, J. (1988) Prediction of velum movement from phonological specifications. *Phonetica*, 45, 122–39.

Velde, H. Van de (2007) Phonetic variation in a sociolinguistic context. Oral presentation, Journées des Sciences de la Parole, Charleroi, March.

Vihman, M. M. (1996) *Phonological Development: The Origins of Language in the Child*. Oxford: Blackwell.

Walton, J. H. & Orlikoff, R. F. (1994) Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech and Hearing Research*, 37, 738–45.

Warren, P., Hay, J., & Thomas, B. (2007) The loci of sound change effects in recognition and perception. In J. Cole & J.-I. Hualde (eds.), *Laboratory Phonology IX* (pp. 87–112). Berlin: Mouton de Gruyter.

Wassink, A. B. & Dyer, J. (2004) Language ideology and the transmission of phonological change: Changing indexicality in two situations of language contact. *Journal of English Linguistics*, 32, 3–30.

Wassink, A. B., Wright, R. A., & Franklin, A. D. (2007) Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech and Lombard speech in Jamaican speakers. *Journal of Phonetics*, 35, 363–79.

Watt, D. & Fabricius, A. (2002) Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1~F2 plane. *Leeds Working Papers in Linguistics and Phonetics*, 9, 159–73.

Watt, D. & Milroy, L. (1999) Patterns of variation and change in three Newcastle vowels: Is this dialect levelling? In P. Foulkes & G. J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 25–46). London: Arnold.

Watt, D. & Smith, J. (2005) Language change. In M. Ball (ed.), *Clinical Sociolinguistics* (pp. 101–19). Oxford: Blackwell.

Watt, D. & Yurkova, J. H. (2007) Voice Onset Time and the Scottish Vowel Length Rule in Aberdeen English. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, pp. 1521–4.

Weinreich, U., Labov, W., & Herzog, M. (1968) Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (eds.), *Directions for Historical Linguistics* (pp. 95–188). Austin: University of Texas Press.

Wells, J. C. (1982) *Accents of English*, 3 vols. Cambridge: Cambridge University Press.

Wells, J. C. (1995) Age grading in English pronunciation preferences. *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 3, 696–9.

Wenker, G. (1895) *Sprachatlas des Deutschen Reichs* [Linguistic Atlas of the German Empire]. Marburg: Elwert.

Whiteside, S. (2001) Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *Journal of the Acoustical Society of America*, 110, 464–78.

Whiteside, S., Henry, L., & Dobbin, R. (2004) Sex differences in Voice Onset Time: A developmental study of phonetic context effects in British English. *Journal of the Acoustical Society of America*, 116, 1179–83.

Wolfram, W. (1969) *A Linguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics.

Wolfram, W. & Schilling-Estes, N. (1998) *American English*. Oxford: Blackwell.

Wolfram, W. & Thomas, E. R. (2002) *The Development of African American English: Evidence from an Isolated Community*. Oxford: Blackwell.

Woolard, K. A. & Schieffelin, B. B. (1994) Language ideology. *Annual Review of Anthropology*, 23, 55–82.

Woolhiser, C. (2005) Political borders and dialect divergence/convergence in Europe. In P. Auer, F. Hinskens, & P. Kerswill (eds.), *Dialect Change: Convergence and Divergence in European Languages* (pp. 236–62). Cambridge: Cambridge University Press.

Wrench, A. A. & Scobbie, J. M. (2003) Categorising vocalisation of English /l/ using EPG, EMA and Ultrasound. In S. Palethorpe & M. Tabain (eds.), *Proceedings of the 6th International Seminar on Speech Production*, Sydney, 314–19.

Wrench, A. A. & Scobbie, J. M. (2006) Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. In H. C. Yehia, D. Demolin, & R. Laboissiere (eds.), *Proceedings of the 7th International Seminar on Speech Production*, Brazil, 451–8.

Wright, S. (1989) The effects of style and speaking rate on /l/-vocalisation in local Cambridge English. *York Papers in Linguistics*, 13, 355–65.

Wright, S. & Kerswill, P. (1989) Electropalatography in the study of connected speech processes. *Clinical Linguistics and Phonetics*, 3, 49–57.

Yaeger-Dror, M. (1994a) Linguistic data solving social psychological questions: The case for (resh) as a measure of ethnic self-identification. *Israel Social Science Research*, 9, 109–60.

Yaeger-Dror, M. (1994b) Phonetic evidence for sound change in Québec French. In P. A. Keating (ed.), *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form* (pp. 267–92). Cambridge: Cambridge University Press.

Yaeger-Dror, M. & Di Paolo, M. (eds.) (2010) *Sociophonetics: A Student's Guide*. London: Routledge.

Zhang, Z., Boyce, S., Espy-Wilson, C., & Tiede, M. (2003) Acoustic strategies for production of American English "retroflex" /r/. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 1125–8.

Part V Speech Technology

20 An Introduction to Signal Processing for Speech

DANIEL P. W. ELLIS

1 Overview

The formal tools of signal processing emerged in the mid twentieth century when electronics gave us the ability to manipulate signals – time-varying measurements – to extract or rearrange various aspects of interest to us, i.e., the *information* in the signal. The core of traditional signal processing is a way of looking at the signals in terms of sinusoidal components of differing frequencies (the Fourier domain), and a set of techniques for modifying signals that are most naturally described in that domain, i.e., filtering. Although originally developed using analog electronics, since the 1970s signal processing has more and more been implemented on computers in the digital domain, leading to some modifications to the theory without changing its essential character. This chapter aims to give a transparent and intuitive introduction to the basic ideas of the Fourier domain and filtering, and connects them to some of the common representations used in speech science, including the spectrogram and cepstral coefficients. We assume the absolute minimum of prior technical background, which will naturally be below the level of many readers; however, there may be some value in taking such a ground-up approach even for those for whom much of the material is review.

2 Resonance

Consider swinging on a child's swing. (It may be a while since you've been on one, but you can probably remember what it was like.) Even without touching the ground, just by shifting your weight at the appropriate points in the cycle, you can build up a considerable swinging motion. The amount of work you put in at each cycle is quite small, but it slowly builds up, until it is enough to lift you high off the ground at each extreme. Building up the swing requires making the right movements at just the right time – it can take a child a while to figure out how to do this. More vigorous movements can build up the swinging more quickly, but the cycle time – the time between two successive instants at the same point in the cycle, for instance the highest point on the back of the swing – is basically unvarying with the amplitude of the swinging, the amount of work you put in. Even the size and weight of the child doesn't have much effect, except in so far as their center of mass gets further from the seat as they get bigger.

The swing is an example of a pendulum, a simple physical system that can exhibit oscillations, or a pattern of motions that repeats with little variation with a fixed repetition time. The conditions that support this kind of oscillation are relatively simple and very common in the physical world, meaning that the simple mathematics describing the relationship between the input (child's weight shifts) and output (swing motion) apply largely unmodified in a wide range of situations. In particular, we are interested in the phenomenon of *resonance*, which refers to the single "best frequency" – the way that the largest swinging amplitude occurs when the swinger injects energy at a single frequency that depends on the physical properties of the oscillating system (which are usually fixed).

The swing is an interesting example because it reveals how small amounts of work input can lead to large amplitude output oscillations, provided the inputs occur at the right frequency, and, importantly, the right point in the cycle, i.e., the correct "phase" relationship to the motion. But another pendulum example can more clearly illustrate the idea of frequency selectivity in resonance: consider a weight – say some keys or a locket – on the end of a chain. By holding the top of the chain, and moving your hand from side to side, you can make the weight move from side to side with the same period. However, the *amplitude* of the weight's motion, for a fixed amplitude of hand motion input, will vary greatly with the cycle period, and for a small range around a particular frequency, the *natural frequency* of the setup, the output motion will be very large, just like the motion of the swinging child. If you deliberately start moving your hand slightly faster or slightly slower than this best frequency, you will still be able to make the weight oscillate with the same period as your hand, but the amplitude will fall rapidly as you move away from the natural frequency. We could make a plot of the amplitude of the side-to-side motion in response to a fixed amplitude input motion as a function of the frequency of that input motion, and it might look something like Figure 20.1. At low frequencies, the ratio between input and output motion is approximately 1, i.e., when moved slowly, the bottom of the pendulum follows the top. Around the natural frequency, the ratio is very large – small motions of the pendulum top lead to wild swinging, at the same frequency, of the weight. At high frequencies, the ratio tends to zero: rapid motion at the top of the pendulum is "lost," leaving the weight almost stationary.

3 Sinusoids

Mathematical equivalents of the pendulum and a few simple variants are remarkably common in the natural world, ranging from the task of trying to rock a trapped

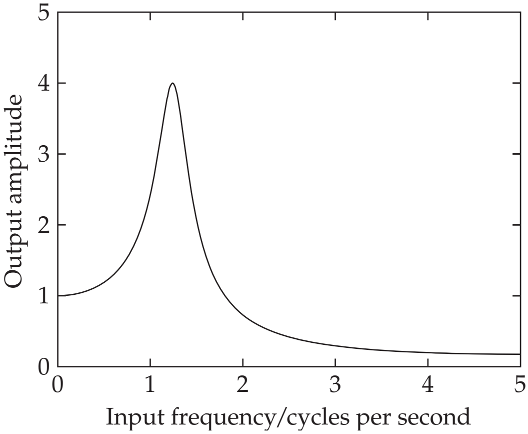


Figure 20.1 Sketch example of the amplitude of pendulum motion in response to a fixed amplitude input of varying frequency.

car out of a snowdrift to the shaking of the earth's crust after an earthquake, all the way to the quartz crystal at the heart of a digital watch or a radio antenna. There are a couple of aspects of this common phenomenon which we should note at this point relating to the shape of the resonant waveform.

Figure 20.2 returns to the swing example, plotting the position of the weight (the child) as a function of time. We see the regular, periodic motion appear as a repeating forward-backward pattern, but the particular shape of this pattern, the smooth alternating peaks of the *sinusoid*, is essential to and characteristic of this behavior. You probably came across sinusoids in trigonometry as the projections of a fixed length (hypotenuse) at a particular angle, but that doesn't seem to provide much insight to their occurrence here. Instead, the interesting property of the sinusoid is that its slope – the rate of change at any given time – is another sinusoid, at the same frequency (of course, since the whole system repeats exactly at that frequency), but shifted in time by one-quarter of a cycle. The slope of a graph showing position as a function of time is *velocity*, the rate at which position is changing, and this is shown to the right of the figure. In the plot, the relative amplitudes of position and velocity are arbitrary, since they depend on the different units used to measure each quantity.

Looking in detail at the figure, there are four time points labeled, with a snapshot of the swing shown for each one. At time A, the beginning of the graph, the swing is hanging straight down, but it is about to move forward, perhaps because the swinger has just pushed off against the ground. Thus the position is at zero, but the velocity is positive and indeed at its maximum value. At point B, the swing has returned to its straight-down position, but now it is traveling backwards, so we see that the velocity is at its maximum backwards (negative value). Conversely, point C finds the swing at its maximum positive position, when the motion is changing from forward to backward, so the swing is still for a split second – i.e.,

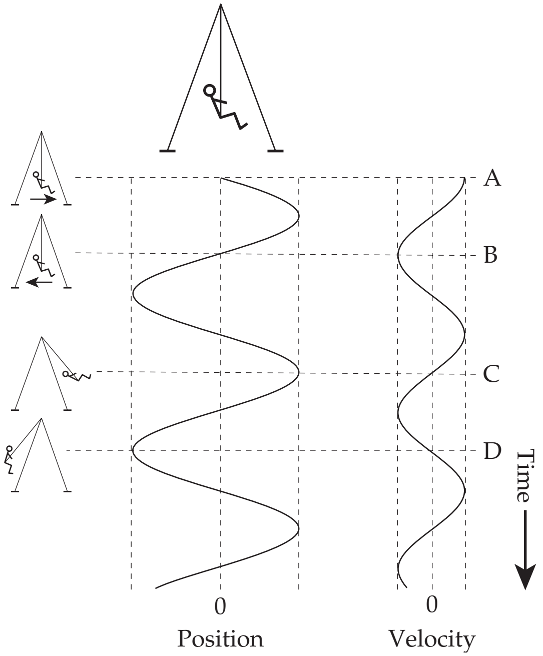


Figure 20.2 Motion of the simple pendulum swing. The slope (derivative) of the plot of position with time gives the velocity, which is shown on the right. Small sketches to the left show the instantaneous configuration of the swing at the times labeled A to D. Note that both position and velocity are sinusoids, a waveform with a "pure" frequency, and that there is a one-quarter cycle phase shift between them.

at zero velocity. D is the symmetric case at the very back of the swing, again a moment of zero velocity.

If we normalize time to be in units of complete cycles, we can equate it to the angles of trigonometry, in which case we speak of the *phase* of the sinusoid, and measure it either in degrees (360° for a complete cycle) or radians (2π rad for a complete cycle). At a given frequency, a time shift is equivalent to a phase shift, and in this case the quarter-cycle difference between position and velocity corresponds to a 90° or $\pi/2$ rad *phase shift*.

Resonant systems of this kind involve the periodic transfer of energy between two forms. In this case, one is the kinetic energy of the motion of the weight, which is proportional to the square of the velocity. At points A and B, all the system in the swing is in the kinetic energy of the swinger's motion – energy that would be violently transferred to someone unlucky enough to step into the path of the swing. The complementary energy form is potential energy, the energy gained by lifting the swinger up against gravity away from the ground due to the arc traced by the swing. At points C and D, when the swing is momentarily stationary (zero kinetic energy), the swinger is also highest from the ground,

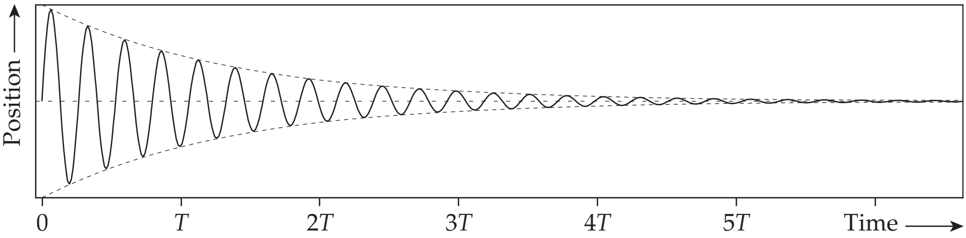


Figure 20.3 Exponential decay. The sinusoid amplitude steadily decreases, losing a fixed *proportion* of its amplitude in a fixed time, i.e., it halves in amplitude every *T* seconds.

corresponding to maximal potential energy. The total of kinetic plus potential energy is constant throughout the cycle, but at all other points it is shared between the two forms. Not only does resonant behavior always involve such an exchange between two energy forms, but the particular 90° phase shift between the two domains is also a common characteristic. It is the constant transfer of energy between these two forms that leads to the visible, dynamic behavior, even without any additional energy input, i.e., if the swinger remained frozen on the seat.

This brings us to a second aspect of resonant motion – *exponential decay*. If the swinger builds up a large-amplitude motion, then stops shifting their weight to inject energy into the system, the sinusoidal motion continues, but its amplitude gradually decays until the swing is essentially still. What has happened is that the energy is being lost, for example to air resistance, and heating up the moving joints of the swing mechanism, so the maxima of kinetic and potential energy are steadily decreasing. Figure 20.3 shows an exponentially-decaying sinusoid, as might occur with undriven swinging. Notice that the amplitude decays by a constant factor for each unit of time, i.e., if it decays to half its initial amplitude by time T, it has decayed to one quarter at time $2T$, one eighth at time $3T$, one sixteenth at time $4T$, etc. This is the hallmark of exponential decay.

Exponential decay of an oscillating, resonant system is also related to its "tuning," i.e., how sharply it responds to its natural frequency. This is visible as the sharpness of the peak in a plot of amplitude versus frequency like Figure 20.1, and actually depends on the rate of energy loss. A highly tuned resonance has a very sharp resonant peak as a function of frequency, and its resonant oscillations die away very slowly. A more strongly "damped" resonance dissipates more energy each cycle, meaning that oscillations die away more rapidly, and the preference for the natural frequency over other oscillation frequencies is less dramatic. As an example, the shock absorbers (also known as dampers) in a car's suspension system are needed to minimize "bouncing" oscillations that would otherwise occur with the spring-mass system at each wheel. When the shock absorbers age or fail they lose their ability to dissipate energy, and the car will develop a tell-tale (and very disconcerting) tendency to oscillate up and down every time it goes over a bump.

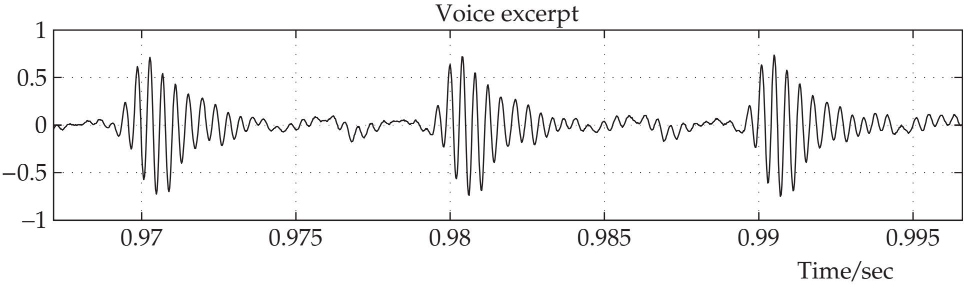


Figure 20.4 Example of vocal waveform during a vowel. We see three cycles of the voice pitch; within each cycle, the waveform looks a lot like an exponentially decaying sinusoid.

A simple resonance always involves sinusoidal motion. Returning to the pendulum being shaken at its top, even if you move your hand in a much less smooth way, for instance shifting almost instantaneously from one extreme to the other for a square-wave input, the motion of the weight will still be essentially a sinusoid with the same period. And if you shake your hand randomly, most often the weight will begin to move at its resonant frequency with a sinusoidal motion – as we will see below, we interpret this as the resonant system "filtering out" a frequency component at the natural frequency from the random motion.

Lest we lose sight of the motivation of this entire discussion, Figure 20.4 shows a brief excerpt of a voice waveform, extracted from a vowel sound. We see a few cycles of the fundamental voice period, but notice how within each period, the waveform looks a lot like an exponentially decaying sinusoid. This is because the vocal tract, shaping the sound, is behaving as a simple resonant system being shocked into motion once every pitch cycle.

4 Linearity

Prior to presenting the central idea of Fourier analysis, there is one more supporting concept to explore: linearity. Very roughly, linearity is the idea that scaling the input to a system will result in scaling the output by the same amount – which was implicit in the choice of using the *ratio* of input to output amplitudes in the graph of Figure 20.1, i.e., the ratio of input to output did not depend on the absolute level of input (at least within reasonable bounds). Linearity is an idealization, but happily it is widely obeyed in nature, particularly if circumstances are restricted to small deviations around some stable equilibrium.

In signal processing, we use "system" to mean any process that takes a signal (e.g., a sound waveform) as input and generates another signal as output. A linear system is one that has the linearity property, and this constitutes a large class of real-world systems including acoustic environments or channels with

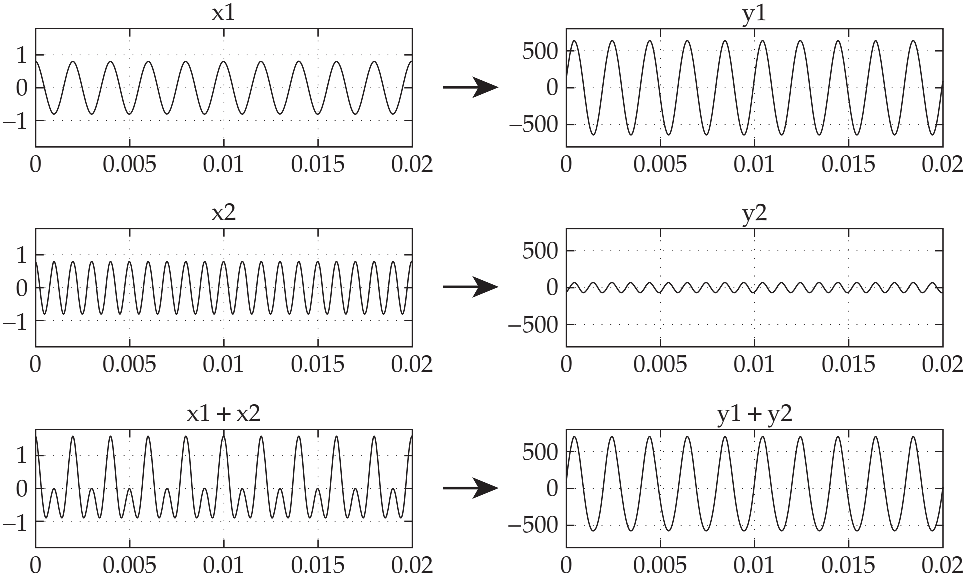


Figure 20.5 Illustration of superposition. Left column shows three inputs to a linear system, where the third is the sum of the preceding two. Right column shows the corresponding outputs; because the first row has a much larger amplitude than the second, the third looks largely the same as the first.

rigid boundaries, as well as other domains including radio waves and mechanical systems consisting of rigid connections, ideal springs, and dampers. Of course, most scenarios of interest also involve some nonlinear components, for example the vocal folds that convert steady air pressure from the lungs into periodic pressure waves in the (largely linear) vocal tract.

Linearity has an important and subtle consequence: *superposition*. The property of superposition means that if you know the outputs of a particular system in response to two different inputs, then the output of the system in response to the *sum* of the two inputs is simply the sum of the two outputs. Figure 20.5 illustrates this. The left columns show inputs, and the right columns show outputs, for a simple resonant system as described in section 2. The first row shows the system response when the input is a sinusoid right around the natural frequency. Notice that the scale on the output graph is much larger, reflecting the increased amplitude of the output sinusoid. The second row is for a sinusoid at double the frequency, outside the resonant peak. The output is still a sinusoid at the same frequency, but its amplitude is much smaller than in the resonant case. Finally, adding the two inputs results in a waveform with the same basic period as the first row, but no longer sinusoidal in shape. For this linear system that obeys superposition, the output is simply the sum of the outputs from the previous two conditions. But because the output in the first row was so much larger than that in the second row, the sum is dominated by the first row, so the contribution of the higher frequency component in the input is practically negligible.

Now we can learn one more very important property of sinusoids: They are the *eigenfunctions* of linear systems. What this means is that if a linear system is fed a sinusoid (with or without an exponential envelope), the output will also be a sinusoid, with the same frequency and the same rate of exponential decay, merely scaled in amplitude and possibly shifted in phase. The scale factor (or *gain*) and shift angles will depend on the sinusoid frequency, but are otherwise constant properties of the system. Any other waveform will in general be modified in a more complex way that cannot be explained by a single, constant factor or phase shift. This property is evident in Figure 20.5, where the output of the resonance to each of the two sinusoids is simply a scaled and delayed version of the input, but the more complex waveform is extensively modified (to look more like a sinusoid). This sinusoid-in, sinusoid-out property, combined with superposition, is the key to the value of Fourier transform, presented in the next section, which describes an arbitrary input as a sum of sinusoids.

One last definition: If a system's property changes – for instance if an acoustic tube changes shape or length – the specific scaling and phase shift values for each sinusoid frequency will likely change too. Much of our analysis assumes *time-invariant* systems, so that we can assume the way in which a signal is modified does not depend on precisely *when* the signal occurs. However, even systems which are not time-invariant (such as the human vocal tract) can be treated as *locally* time invariant, i.e., the modification applied to a given sinusoid will change only relatively slowly and smoothly, so the linearity assumptions can be applied successfully over sufficiently short time scales.

5 Fourier Analysis

The core of signal processing is Fourier analysis, and the core of Fourier analysis is a simple but somewhat surprising fact: Any periodically repeating waveform can be expressed as a sum of sinusoids, each scaled and shifted in time by appropriate constants. Moreover, the only sinusoids required are those whose frequency is an integer multiple of the fundamental frequency of the periodic sequence. These sinusoids are called the *harmonics* of the fundamental frequency.

To get an arbitrarily good approximation to an arbitrary periodic waveform, it may be necessary to include a very large number of sinusoids, i.e., continue up to sinusoids whose frequency is very high. However, it turns out that the scaling of any single sinusoid that gives the best approximation doesn't depend on how many sinusoids are used. Thus, the best approximation using only a few sinusoids can be derived from a higher-order, more accurate approximation simply by dropping some of the harmonics.

It makes sense that the only sinusoids involved are the harmonics (integer multiples of the fundamental frequency), since only these sinusoids will complete an exact, integer number of cycles in the fundamental period; any other sinusoids

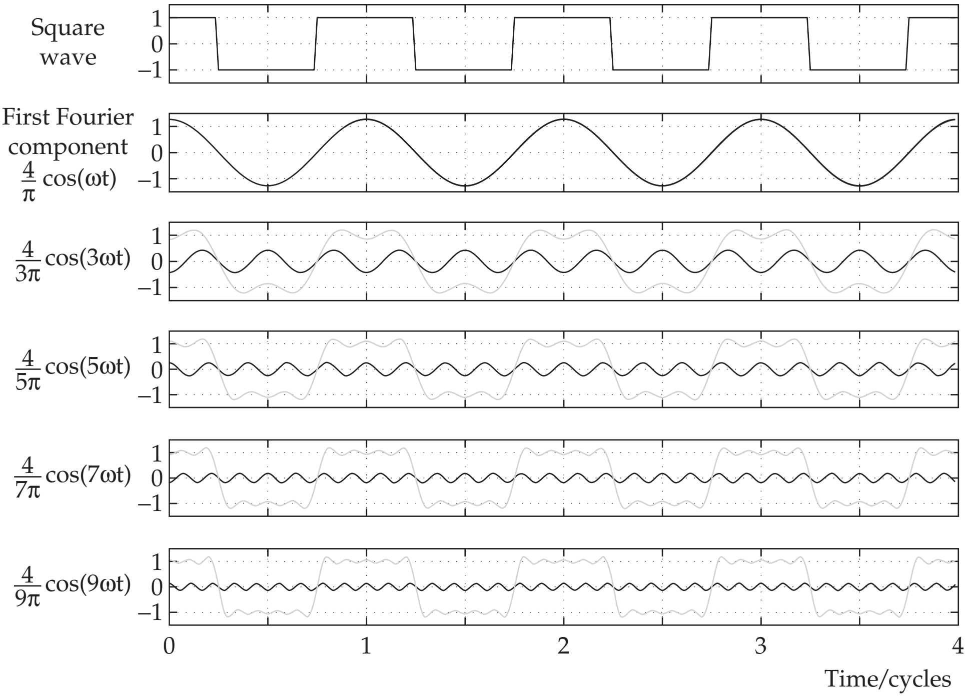


Figure 20.6 Illustration of how any periodic function can be approximated by a sum of harmonics, i.e., sinusoids at integer multiples of the fundamental frequency of the original waveform. Top panel shows the target waveform, a square wave. Next panels show the first five harmonics; in each panel, the dark curve is the sinusoid, and the light curve is the cumulative sum of all harmonics so far, showing how the approximation comes increasingly close to the target signal.

would not repeat exactly in each period of the signal, and thus could not sum to a waveform that was exactly the same every period.

Figure 20.6 illustrates the *Fourier Series* concept. The original periodic waveform is a square wave, i.e., with abrupt transitions from +1 to –1 and back in each cycle. It is particularly surprising that the sum of a series of smooth sinusoid functions can even approximate such a discontinuous function, but the figure illustrates how the first five Fourier components (which in this case consist only of *odd* multiples of the fundamental frequency), when appropriately scaled and aligned, begin to reinforce and cancel to match the piecewise-constant waveform. We also note that in the special case of the square wave, the harmonic amplitudes are inversely proportional to the harmonic number.

It turns out that finding the Fourier series coefficients – the optimal scale constants and phase shifts for each harmonic – is very straightforward: All you have to do is multiply the waveform, point-for-point, with a candidate harmonic, and sum up (i.e., integrate) over a complete cycle; this is known as taking the *inner* *product* between the waveform and the harmonic, and gives the required scale constant for that harmonic. This works because the harmonics are *orthogonal*, meaning that the inner product between different harmonics is exactly zero, so if we assume that the original waveform is a sum of scaled harmonics, only the term involving the candidate harmonic appears in the result of the inner product. Finding the phase requires taking the inner product twice, once with a cosine-phase harmonic and once with the sine-phase harmonic, giving two scaled harmonics that can sum together to give a sinusoid of the corresponding frequency at any amplitude and any phase.

Finding the Fourier series representation is called Fourier analysis; the converse, Fourier synthesis, consists of converting a set of Fourier coefficients into a waveform by explicitly calculating and summing up all the harmonics. A waveform that is created by Fourier synthesis will yield the exact same parameters on a subsequent Fourier analysis, and the two representations – the waveform as a function of time, or the Fourier coefficients as a function of frequency – may be regarded as equally valid descriptions of the function, i.e., together they form a *transform pair*, one in the time domain, and the other in the frequency, or Fourier, domain.

If Fourier analysis could be applied only to strictly periodic signals it would be of limited interest, since a purely periodic signal, which repeats exactly out to infinite time in both directions, is a mathematical abstraction that does not exist in the real world. Consider, however, stretching the period of repetition to be longer and longer. Fourier analysis states that within this very long period we can have any arbitrary and unique waveform, and we will still be able to represent it as accurately as we wish. All that happens is that the "harmonics" of our very long period become more and more closely spaced in frequency (since they are integer multiples of a fundamental frequency which is one divided by the fundamental period, which is becoming very large). Put another way, to capture detail up to a fixed upper frequency limit, we will need to specify more and more harmonics.

Now by letting the fundamental period go to infinity, we end up with a signal that is no longer periodic, since there is only space for a single repetition in the entire real-time axis; at the same time, the spacing between our harmonics goes to zero, meaning that the Fourier series now becomes a continuous function of frequency, not a series of discrete values. However, nothing essentially changes – and, in particular, we can still find the value of the Fourier transform function simply by calculating the inner product integral. Now we have the most general form of the Fourier transform, pairing a continuous, nonrepeating (aperiodic) waveform in time, with a continuous function of frequency. In this form, the symmetry between time and frequency (and the lack of privilege for either domain, from the point of view of mathematics) begins to become apparent.

One kind of aperiodic waveform that might interest us is a finite-length waveform, i.e., a stretch of signal that exists over some limited time range, but is zero everywhere else. Since it never repeats, its Fourier transform is continuous. However, it turns out that the constraint of finite extent in time imposes smoothness in the

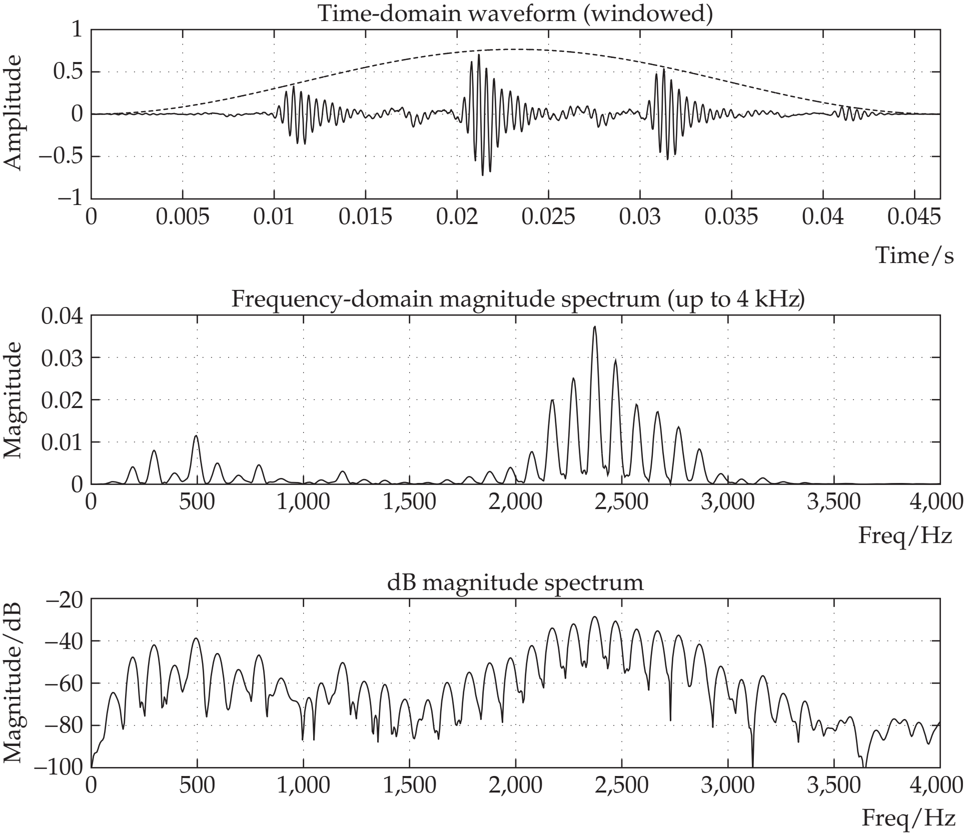


Figure 20.7 Fourier transform pair example. Top panel: time-domain waveform of a brief vowel excerpt, windowed by a raised cosine tapered window (dotted). Middle panel: Fourier transform (spectral) magnitude, up to 4 kHz. Bottom panel: spectral magnitude using a dB (logarithmic) vertical scale.

frequency domain, meaning that we can be sure not to miss any important detail if we only evaluate the Fourier transform at a limited number of regularly spaced frequency points.

As an example, Figure 20.7 shows the brief speech excerpt from Figure 20.4 along with the magnitude of its Fourier transform, up to 4 kHz. (The magnitude is the length of the hypotenuse of the right-angle triangle formed by the sine and cosine coefficients for a particular frequency, and corresponds to the amplitude of the implied sinusoid.) A Fourier transform magnitude plot like this is commonly known as a magnitude *spectrum*, or just spectrum. It is shown in two forms: the middle panel uses a linear magnitude axis, and the bottom panel plots the magnitude in deciBels (dB), a logarithmic scale that reveals more detail in the low-amplitude parts of the spectrum. Note that the 80 dB vertical range in the bottom plot corresponds to a ratio in linear magnitude of 10,000 : 1 between most and least intense values. The time waveform has been scaled by a tapered window (shown dotted) to avoid abrupt transitions to zero at the edges, which would otherwise introduce high-frequency artifacts. The time-domain signal is zero everywhere that is not shown in the image. We notice dense, regularly spaced peaks in the spectrum; these arise because of the pitch-periodicity evident in the waveform (i.e., the repetitions at roughly 10 ms). If the signal were exactly periodic and infinitely repeating, these spectral peaks would become infinitely narrow, existing only at the harmonics of the fundamental frequency; the Fourier transform would become the Fourier series. Superimposed on this fine structure we see a broad peak in the spectrum centered around 2,400 Hz; this is the vocal tract resonance being driven by the pitch pulses, and corresponds to the rapid, decaying oscillations we noticed around each pitch pulse in the time domain. A quick count confirms that these oscillations make around 12 cycles in 5 ms, which indeed corresponds to a frequency of 2,400 Hz. The spectrum has no significant energy above 4 kHz, although the dB-domain plot shows that the energy has not fallen all the way to zero.

6 Filters

Taken alone, the existence of the Fourier transform might be no more than a mathematical curiosity, but in combination with our previous examination of the properties of sinusoids, linearity, and superposition, it becomes extremely powerful. Recall that in section 4 we said that (1) feeding a linear, time-invariant system with a single sinusoid will result in a scaled and phase-shifted sinusoid of the same frequency at the output, and (2) the output of a linear system whose input is the sum of several signals will be the sum (superposition) of the system's output to each of those signals in isolation. The Fourier transform allows us to describe any signal as the sum of a (possibly very large) set of sinusoids, and thus the output of a particular system given that signal as input will be the *same* set of sinusoids, but with their amplitudes and phases shifted by the frequency-dependent values that characterize that system. Thus, by measuring – once – the way in which a system modifies sinusoids of all relevant frequencies, it is a simple matter to predict the Fourier transform (and hence the time-domain waveform) of the system's output in response to any input signal described by its Fourier transform.

In signal processing, a *filter* is essentially any system with an input and an output, but the term implies that the properties of the system are being viewed as emphasizing certain aspects of the signal while reducing or removing certain others. In a linear, time-invariant filter, it is the Fourier components – sinusoids of differing frequencies – that are selected, meaning that they are either amplified (made larger) or attenuated (made smaller). There are infinitely many possible filters, even within this relatively narrow, idealized set, but they are typically categorized according to the broad properties of how their scaling effects vary with frequency: a low-pass filter boosts low frequencies close to zero; high pass does the converse, attenuating lower frequencies; bandpass selects frequencies within a limited range, and band-stop or notch filters remove specific frequency ranges. Note that the simple resonance with which we introduced this chapter in Figure 20.1 is a kind of bandpass filter. One way to make a band-stop filter is to "cancel out" energy at certain frequencies (e.g., by adding it to a negated version of itself, corresponding to a 180° phase shift), leaving the low and high frequencies. Resonances in systems are often called "poles," referring to a specific feature of the mathematical description of the system; the locally attenuating aspects, as found in band-stop filters, are known as "zeros" because they can remove certain sinusoidal or decaying-sinusoidal inputs to give zero output.

Much of the foundation of signal processing involves techniques to design and construct filters to achieve specific goals and characteristics. There are a number of "optimal" design procedures that design filters, for implementation in electronics or software, that do the best possible job in terms of leaving some frequencies unmodified while removing others, subject to various constraints such as cost or complexity. While linear filtering is a relatively limited subset of all possible signal modifications – only slightly more complex than the "treble" and "bass" controls on a hi-fi amplifier – it turns out to be very useful in a wide range of applications, particularly when trying to separate a particular piece of information, such as a particular voice, from the middle of a large amount of background noise.

As we just mentioned, filters can be implemented in a variety of forms: it is possible to build *acoustic* systems with controlled resonant properties, such as an organ pipe which is a very sharply tuned band-pass filter coupled to a nonlinear air-jet oscillator. (You may have noticed that the sound of a pure sinusoid is reminiscent of an organ or a flute.) However, the birth of signal processing occurred when it became possible to represent signals (audio, radio, or others such as television) as electrical voltages and process them using electronic circuits. Much of the theoretical foundation was based on analog electronics, but from the 1960s onwards more and more signal processing has been performed on digital computers using signals represented as sequences of values stored in memory.

This required a modified theoretical foundation, known as discrete-time or digital signal processing (DSP), because whereas an analog voltage can in theory vary at any frequency from the very slow to the extremely fast, a digital representation involves measuring and storing the voltage only at a finite set of discrete instants (usually regularly spaced). DSP systems usually have a fixed sampling rate, which is the number of samples recorded every second, and any variations in the signal which involve significant detail below the timescale of this sampling will not be accurately captured. In fact, it turns out that to store components up to some particular frequency, it is necessary to sample at least double that frequency, thus the highest correctly represented frequency is half the sampling rate, known as the *Nyquist frequency*. For example, in CD digital audio, the sampling rate was chosen as 44.1 kHz to ensure that the highest frequencies perceptible by humans – around 20 kHz – could be adequately represented, with the extra 10 percent

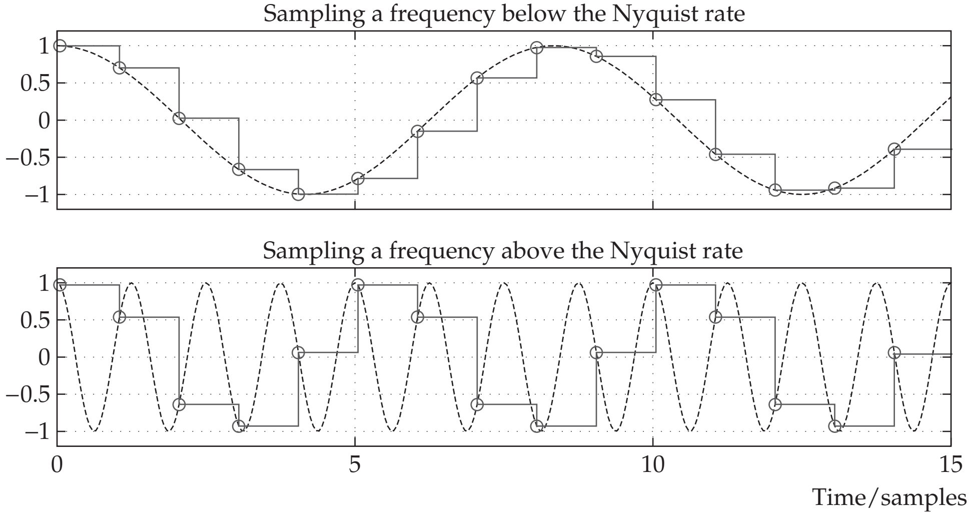


Figure 20.8 Illustration of discrete-time (sampled) representations of sinusoids below (top) and above (bottom) the Nyquist rate. Analog domain signals are shown as dotted, and sampled values are shown with circles, connected by stairstep lines. Note that the discrete, stairstep signal captures the general shape of the below-Nyquist frequency sinusoid, but for the higher frequency, the sampled representation appears to reflect a completely different periodicity.

providing some breathing space to make it easier to construct the digital-to-analog converters required to render the digital representation back into an actual, physical sound for listeners. The problems that arise when sampling a frequency higher than the Nyquist rate are illustrated in Figure 20.8.

7 The Spectrogram

Filters and signal processing turn up in many places in phonetics and speech science, from cleaning up field recordings through to performing data compression on archives, but perhaps their greatest impact is in providing analysis tools that can measure and quantify different acoustic phenomena, and perhaps the most familiar of those is the spectrogram. We can now precisely describe how a spectrogram image is constructed, using the ideas presented so far, but first we must mention one more reason why sinusoids and the frequency domain are so important and relevant for sound: the nature of hearing.

Once air pressure fluctuations have been collected by the outer ear, converted into microscopic force variations by the eardrum, and transferred to the inner ear by the bones of the middle ear, they encounter the single most critical component of auditory perception, the cochlea, which is responsible for converting a one-dimensional time-varying pressure into a sequence of firings on the tens of thousands of nerve fibers of the auditory nerve flowing to the brain. The cochlea is exquisitely sensitive and very complex, but in essence it is just a series of resonators, not so very different from the one described by Figure 20.1. Each resonator responds to energy in a narrow frequency range, causing a certain subset of nerve fibers to fire when there is energy at those frequencies. Thus, the auditory system performs something like a Fourier transform, breaking down the time-domain pressure fluctuations into separate sinusoidal components through a bank of resonant filters.

The representation on the auditory nerve, however, is not a pure frequency domain representation: Although different sets of nerves indicate energy in different frequencies, they also vary in time – whereas a pure Fourier representation would only have a frequency axis. In fact, the transformation performed by the cochlea is closer to a short-time Fourier transform (STFT), which breaks up a longer signal into a succession of smaller fragments, centered around different, specific times by gating the original signal with a sliding window, then calculating the Fourier transform of each of these time-localized pieces to reveal the varying energy in each frequency band as the input signal changes. Note that this is exactly equivalent to constructing a set of bandpass filters, one for each frequency being considered, then calculating the total energy coming out of each band over a succession of short windows.

Although there is rather more to how the ear works than this, the magnitude of the short-time Fourier transform is exactly what we see when we look at a spectrogram, and one of the reasons it is such a valuable tool for visualizing sound content is because of its correspondence, in very broad terms, to the internal representation of sound employed by our brains. A spectrogram is actually a fine grid of cells, indexed by frequency on the vertical axis, and by time on the horizontal axis, where the color (or darkness) of each cell indicates the amount of energy in that frequency band at around that time. The spectrogram typically uses a logarithmic mapping from signal intensity to pixel darkness (i.e., a deciBel scale), corresponding to the nearly-logarithmic mapping from signal intensity to perceived loudness observed in hearing.

The main parameter of a spectrogram is its analysis window length, which in turn determines its spectral resolution: There is a formal relationship of uncertainty between time and frequency (inevitable since they both arise from the same one-dimensional waveform, rather than being in any sense independent quantities). When we are interested in seeing voicing pitch revealed as a set of parallel sinusoidal harmonics, we must use a longer time window of 20 ms or more to give a fine (narrowband) frequency resolution. To see fine timing detail of pitch pulses and stop bursts, we use a much shorter time window of a few ms at which point pitch harmonics blur together and disappear, but the broader spectral variation due to the formant resonances in the vocal tract remain quite visible.

Figure 20.9 illustrates the steps involved in converting a time waveform into a sequence of short-time Fourier transforms and assembling these into a spectrogram image. Although the first versions of this representation used analog bandpass filters, we now always calculate the spectrogram on a computer (meaning, among

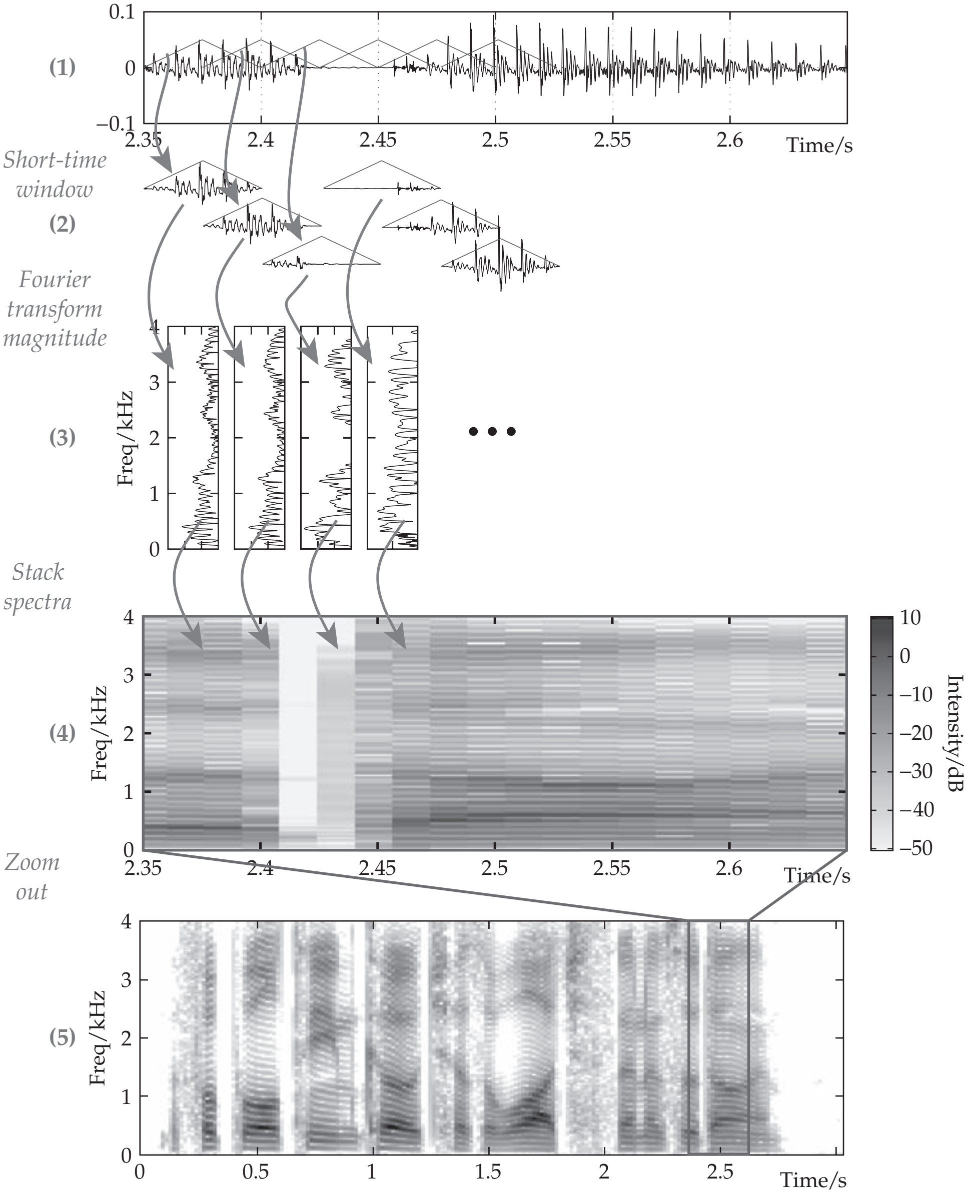


Figure 20.9 Calculation of the spectrogram. Input signal (1) is converted into a sequence of short excerpts by applying a sliding tapered window (2). Each short excerpt is converted to the frequency domain via the Fourier transform (3), then these individual spectra become columns in the spectrogram image (4), with each pixel's color reflecting the log-magnitude at the corresponding frequency value in the Fourier transform. After zooming out, the individual columns of pixels cannot be distinguished, and we have the appearance of a continuous image – the spectrogram (5).

other things, there is always a hard limit on the frequency range at the Nyquist rate). Since each column of pixels involves taking a Fourier transform, the computation involved in creating a spectrogram can be substantial, however it is made significantly more feasible through a special, optimized algorithm for calculating the Fourier transform of discrete-time signals called the Fast Fourier Transform (FFT). The FFT manages to exploit redundancy between the values being employed when taking the inner products against sinusoids of different frequencies to reduce the computation required by a factor that actually improves as the transforms become larger (longer time windows).

8 Linear Prediction

Imagine we have a system and we want to know how to build a copy – an artificial system with the same properties. (This problem is sometimes called *system identification*.) We could measure its gain and phase shift at a set of different frequencies (and in fact we can do this very quickly by feeding it as input the superposition of all frequencies at once, which turns out to be an impulse, the briefest possible click.) We could then build a bank of bandpass filters, adjust the gains of each one to match the measured gains at each frequency, then add the outputs together again, and we'd have a system that performed much like our original. But it would be an approximation, and it would involve a very large amount of computation. By contrast, a direct implementation of the discrete-time, single-pole resonance like Figure 20.1 requires just two multiplications per sample. If we knew that the system we were trying to duplicate consisted of only a few simple resonances, we could in theory create a more accurate and much more efficient duplicate by identifying the parameters of those resonances (their best frequencies and tuning), then implement an equivalent resonant system.

As it happens, there is an efficient and robust procedure for doing just this. In a discrete-time implementation, a resonant filter involves a few delays applied to the output signal, then feeding back these delayed outputs (with particular scaling constants) to the input. (The actual number of delays determines the *order* of the filter, i.e., how many distinct resonances it will have.) In effect, in the absence of inputs, the output at a particular time is a linear combination of a few recent output values, and the process of fitting a resonant filter to a particular signal consists of choosing the scaling constants that do the best job of matching (or predicting) each output sample from its immediate predecessors. For this reason, the technique is known as *Linear Prediction* (LP). It is such a useful and powerful technique that it turns up in various other places under names like all-pole modeling and autoregressive modeling.

The actual mathematics is a little involved, but the net result is that given only a segment of the output of a system, linear predictive analysis finds a simple resonant filter that does the best job of accounting for the spectrum of the signal being analyzed, along with the input signal (called the "residual") which, when fed to the resonant system, would recreate the original signal. The approach minimizes the energy of the residual, which is achieved by making it as close as possible to a purely random (white noise) sequence – but any signal structure that cannot be explained by the resonant filter, either because the model has a lower order than the true filter, or due to the input to the original system, can be left in the residual. This makes the approach particularly robust – it does the best job it can within the limitations of the model, but it is perfectly able to approximate more complex signals and systems.

The biggest limitation of linear prediction is that it can only model systems that consist purely of resonances (poles), whereas very many systems of interest will also include zeros. No approach of comparable simplicity and power exists for modeling systems with zeros – partly because once zeros are introduced it becomes much harder to define the unique, best system to approximate any signal, since similar results can in many cases be achieved with either poles or zeros. However, there is one system of great interest that is well approximated by an all-pole model, namely the vocal tract. Apart from nasals, in which the parallel nasal path gives rise to zeros in the overall spectrum, most speech sounds are well approximated as a spectrally flat input signal – often called the excitation *source*, but equivalent to the LP residual – being shaped by a set of resonances which are generally identified with the *filter* effected by the variable cavities of the vocal tract. In practice, this LP source-filter model leads to usable simulations of voice sounds; it was the key to the first wave of mass-market speech synthesis applications (pioneered by the "Speak and Spell" toy in 1978), and it is at the heart of every speech compression algorithm, including GSM and other cell phone encoding schemes. In these applications, speech is broken up into short (10–30 ms) segments, which are then encoded as a single, fixed, LP filter, plus an excitation signal, which, in the popular Code-Excited Linear Prediction (CELP) scheme, is encoded as an index into a large dictionary known to both encoder and receiver. It is the relatively slowly changing character of the physical shape of the vocal tract, and hence its acoustical properties, that allows it to be described by relatively infrequent model updates, leading to very significant data compression gains.

Figure 20.10 illustrates the kinds of approximations that result from LP modeling. The top trace shows the spectrum of a 30 ms segment of speech, similar to Figure 20.7. Below are the spectra (gains as a function of frequency) for series of LP approximations for models of order 2, 4, 6, 8, 10, and 12. Each resonance actually requires two poles to be modeled, so a 12th-order model (with 12 poles) can reproduce up to six resonant peaks. These relatively small LP filters cannot reproduce the spectral detail of the harmonic peaks, which are provided by the excitation in a complete system. In human voice, harmonic structure comes from the nonlinear oscillations of the vocal folds, not from the resonances of the vocal tract.

9 Speech Features

As we have mentioned, the goals of signal processing can be quite diverse, but they all revolve around the idea of manipulating the information content in signals to

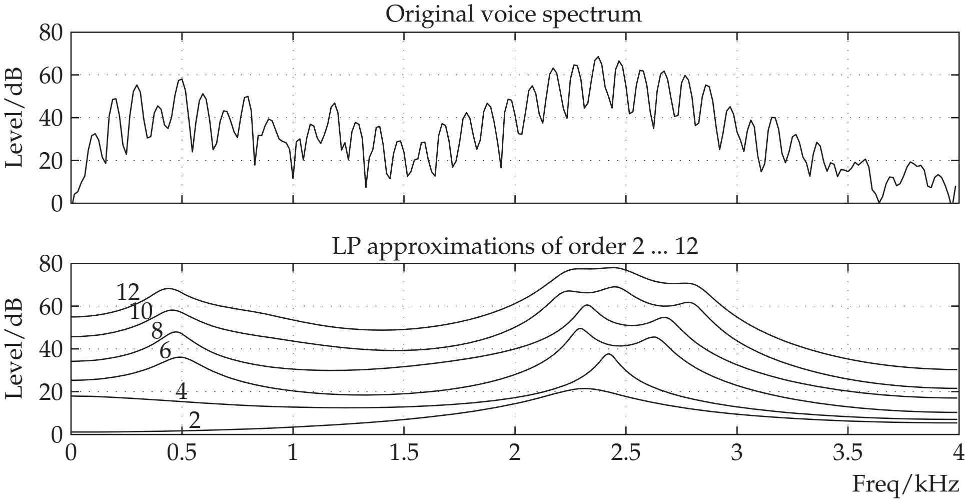


Figure 20.10 Examples of linear predictive models of different orders (lower panel) that are approximating the broad spectral resonances evident in a short speech fragment whose spectrum is shown in the top pane. Adjacent LP spectra are offset by 10 dB to aid visibility.

facilitate some application. One important and illuminating application is automatic speech recognition (ASR), where signal processing is involved at the very beginning to convert the raw speech signals into features that attempt to extract the information from the speech signal most relevant to recognition, while excluding (being invariant to) irrelevant information, and at the same time making the representation as small as possible, to reduce the computational burden. Speech recognition will be described in more detail in the next chapter, but here we will briefly look at the most popular speech features from a signal processing perspective.

9.1 *Spectral features*

Although it is only used directly in fairly rare circumstances, the vast majority of speech recognition features are essentially based on the spectrogram. In particular, speech is first segmented into overlapping short fragments of 20–40 ms, which are given smooth edges with a tapered window, then transformed to the frequency domain to find the magnitude of the energy in each frequency band, while discarding the phase. One reason the unmodified spectrogram is unpopular is that this is still a very large representation, e.g., 256 values per frame, which just means more work and more parameters in the later pattern recognition stages. In fact, the essential challenge of speech recognition is successfully recognizing different instances of the same sound as belonging to the same class – for example, a particular vowel pronounced by different people and at different pitches. Too much spectral detail tends only to make this generalization harder (although it might help in distinguishing two voices from one another).

9.2 LP features

One way to avoid capturing spectral detail that goes beyond that required simply to recognize the phone being pronounced is to fit a low-order, constrained model such as Linear Prediction. As illustrated in Figure 20.10, a low-order LP model will capture the broad spectral shape of the sound but smooth away all the pitch harmonics – an advantage for languages like English where the actual pitch contributes very little phonetic information. By the argument that LP modeling is approximately identifying the resonances arising from the shape and status of the vocal tract (which control exactly how different speech sounds are generated), we even have a feature that directly and compactly describes the nature of the vocal tract configuration. The success of classification tasks often depends on details of how the feature values vary and how well this matches the classifier being used; the mathematical simplicity of LP models opens a wide range of alternative descriptions that carry equivalent information but which have secondary properties making them more suitable for various tasks such as classification, interpolation, and compression. LP models are somewhat vulnerable to background noise, however, since the poles will attempt to model any energy in the original signal, whether or not it comes from noise or interference.

9.3 MFCCs

The most common features used in speech recognition are the *Mel-frequency cepstral coefficients* or MFCCs. Let us explain these two parts separately. The Mel-frequency scale is a nonlinear mapping of the audible frequency range that was proposed in the first half of the twentieth century to account for listeners' judgments about the relative distance between tones – a fixed separation on the Mel axis is supposed to result in pairs of tones that are judged as equally different in pitch. The scale is approximately linear below 1,000 Hz and approximately logarithmic above 1,000 Hz, reflecting the widely supported result that human auditory perception has a bandwidth that increases with frequency – this is even observed in the cochlea, where the resonant structures have broader and broader tuning, and wider spacing, in the higher frequencies. The consequence of this is that a conventional spectrogram, which allocates as many pixels to the spectrum between 0 and 500 Hz as it does to the 3,500 to 4,000 Hz range, seems to be paying too much attention to the higher frequencies at the expense of low-frequency details. There are a number of different auditory frequency scales other than Mel (including Bark), but they all share the property of expanding detail in the low frequencies and compressing it in the high frequencies. In Mel scaling, this can be done by calculating a relatively high-resolution spectrum, then combining together subsets of the frequency values using a weighted average, where the averaging occurs over a wider range of frequencies as the center frequency

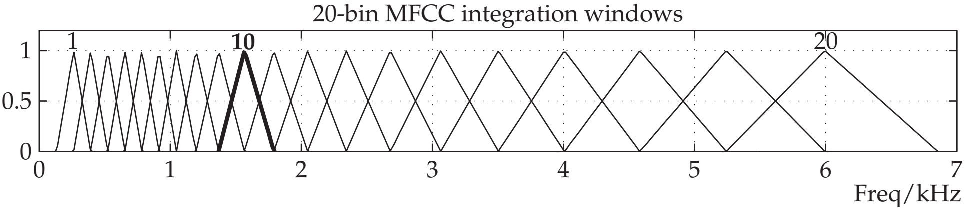


Figure 20.11 Weighting curves to convert a linear frequency spectrum into Mel frequency. The weighting used to construct bin 10 (the middle bin) is shown in bold as an example.

rises. Figure 20.11 illustrates the typical weighting curves used in warping a spectrum to the Mel axis: Each Mel bin combines energy with a triangular weighting scheme, spanning from the center frequencies of its two adjacent bins, with the bin spacing increasing with frequency. For instance, Mel bin number 10, the middle band in this scheme (shown bold in the figure), is composed of a weighted combination of frequencies in the range 1.4 to 1.8 kHz, with the greatest contribution coming from around 1.6 kHz.

The net result of these combinations can be seen in the top two panels of Figure 20.12, which compare a linear frequency spectrogram with the equivalent visualization after the Mel frequency warping is applied. It can be clearly seen how the bottom quarter of the linear-frequency spectrogram has been expanded to fill more than the bottom half in the Mel-scaled version, and the energy above 4 kHz has been squeezed into a small band at the very top of the image, broadly reflecting its relative perceptual importance.

The second part of MFCCs is the Cepstral Coefficients. Cepstra – the name is a play on "spectra," with the idea that the parts have been flipped around – were proposed in the late 1960s as a representation in which the complex effects of filtering on a waveform were converted into simple addition, which could then make them easier to reverse and remove. It amounts to taking a second Fourier transform on the logarithm of the magnitude of the original spectrum (Fourier transform of the time waveform). Because of the symmetry between time and frequency in the basic Fourier mathematics, without the intervening log-magnitude step, taking the Fourier transform of a Fourier transform almost gets you back to the original signal. But taking the magnitude removes any phase (relative timing) information between different frequencies, and applying a logarithm drastically alters the balance between intense and weak components, leading to a very different signal. Cepstra can be calculated on conventional spectra, but the Mel-cepstrum, where the transform is applied to a Mel-warped spectrum, introduces even more changes. As the third panel of Figure 20.12 illustrates, MFCCs are not very useful for visualization, but they are extremely effective as a basis for phonetic classification in speech recognizers. This can be explained by the way that they compactly describe the broad shape of the short-time spectrum using just a few values – and

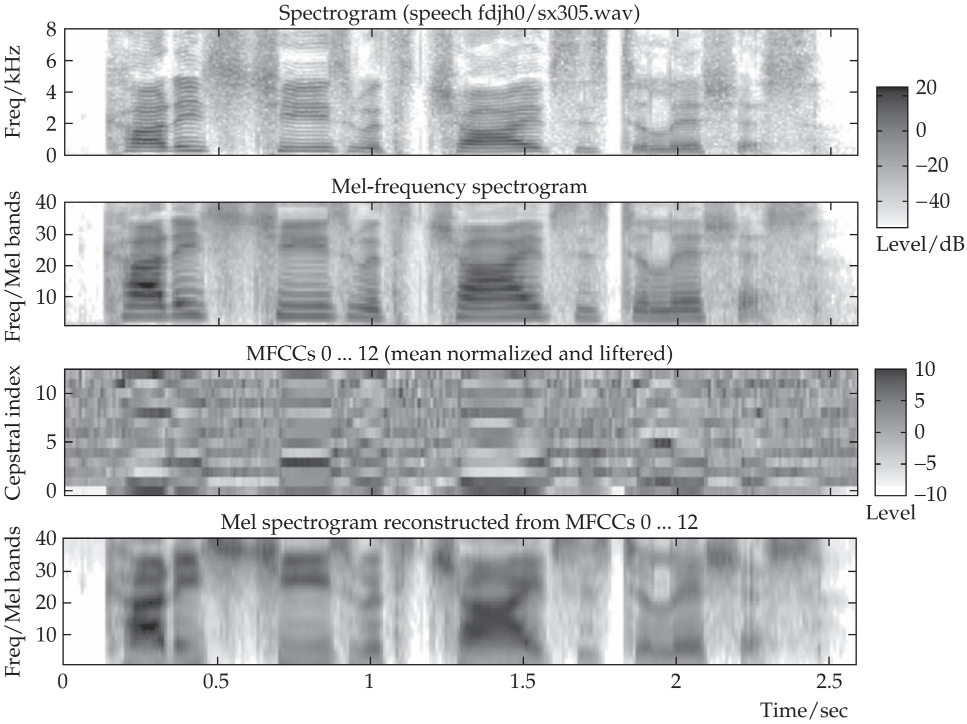


Figure 20.12 Mel-frequency spectra and cepstra. The top panel is a standard, linear-frequency spectrogram; in the second panel, the frequency axis has been warped to the Mel scale by combining each column of the top panel according to a 40-bin version of the weights in Figure 20.11. The third panel shows the first 13 values of the DCT of each column of the Mel spectrogram, which gives the MFCC features most commonly used in speech recognition. ("Liftering" refers to filtering in the cepstral domain.) The final panel shows the effect of inverting those 13 values back into a Mel spectrogram, showing how discarding the higher-order cepstral coefficients has effectively smoothed the spectrum across frequency to remove any pitch-related information or other fine-structure detail.

that these values tend to all be relatively independent, meaning there is little redundant information in the feature vectors. These two properties – compactness and low redundancy – have large practical benefits when building pattern recognition systems, even though the information conveyed is all already present in the original linear-frequency spectrogram.

The final panel of Figure 20.12 shows a reconstruction of the Mel spectrogram based only on the MFCCs. What we see is the smoothing effect that comes from keeping only the first 13 cepstral coefficients. This simply doesn't have enough space to fully describe all 40 Mel values, but it preserves the broad ridges and dips, and discards (smoothes out) finer structure, such as the individual harmonics which are still visible at low frequencies in the original Mel spectrum. It is to be expected that this implicit smoothing is helpful to speech recognizers, since we expect the broad spectral information (i.e., the resonant peaks or formants) to be the relevant information for making phonetic classifications, and that the pitch information would only be a distraction that is better off being discarded.

9.4 Perceptual Linear Prediction (PLP)

Further insight into speech features can be obtained by comparing MFCCs with an alternate, popular representation called *Perceptual Linear Prediction* (PLP; Hermansky, 1990). PLP features often perform comparably to MFCCs, although which feature is superior tends to vary from task to task. PLP features use the Bark auditory scale, and trapezoidal (flat-topped) rather than triangular windows, to create the initial auditory spectrum. Then, rather than smoothing the auditory spectrum by keeping only the low-order cepstral coefficients, linear prediction is used to find a smooth spectrum consisting of only a few resonant peaks (typically 4–6) that matches the Bark spectrum. Although the resulting linear predictor doesn't correspond to any time waveform that has been calculated, it is still possible to perform this fit using a neat piece of mathematics that finds the LP solution starting from only the magnitude spectrum (which we have), rather than the waveform itself. Finally, this smoothed PLP spectrum is again converted to the compact, decorrelated cepstral coefficients via another neat mathematical trick that finds cepstra directly from an LP model – although the same result would arise from calculating the values of the LP model's gain at regularly spaced frequencies, converting to log, and taking the final Fourier transform.

9.5 Other speech feature processing

There are two more steps commonly applied in speech recognition and that can have a significant benefit on recognition accuracy. The first is calculating "delta coefficients," i.e., an estimate of the local slope, along the direction of the time axis, for each frequency or cepstral coefficient. This means that sounds which are better characterized by *changes* in the speech signal than by their instantaneous characteristics – like liquids – can be better recognized in the space of a single frame if their rates of change in different frequency regions are consistent (meaning the delta features will show less spread than the direct spectrum, and will thus be easier for a pattern classifier to identify). The delta slopes are typically calculated by finding the best-fitting straight line over 50–100 ms of signal, to smooth out large variations resulting from noise and other local instability in the voice.

The second commonly applied enhancement is some kind of normalization, most often *Cepstral Mean Normalization* (CMN), in which the average value of each cepstral dimension over an entire segment or utterance is subtracted from that dimension at every time step, making the time-averaged value equal to zero. This additive offset in the cepstral domain is equivalent to an additive offset in the log-spectral domain as well (the inverse Fourier transform of the mean cepstral vector), which is itself equivalent to a constant *multiplicative* factor at each frequency in the linear frequency domain (i.e., a non-time-varying profile of gain as a function of frequency, exactly the effect of a stationary filter). Thus, CMN can largely remove the effect of a large class of fixed, linear filters that might have been applied to the speech, as might arise if the speech has been recorded by different microphones, or passed through a different channel.

10 Conclusions

This chapter has reviewed some aspects of signal processing, starting from a minimum of assumed background, with the aim of giving some additional insight into the properties and meaning of the signal-processing operations and results most often encountered in phonetics. Without any equations, we hope to have supplied some useful, intuitive insights and explanations concerning the operations of speech signal processing. Those seeking greater detail can consult one of the excellent texts on this topic: Lathi (2002) gives a very clear, but mathematically complete, general introduction to signals and systems; Oppenheim et al. (1999) is the most authoritative reference on Digital Signal Processing in general. Finally, Gold and Morgan (2000) (which the current author is at present revising) provides an accessible, wide-ranging, and entertaining overview of speech signal processing and recognition, among other topics.

REFERENCES

Gold, B. & Morgan, N. (2000) *Speech and Audio Signal Processing*. New York: Wiley.

Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis for speech. *Journal of the Acoustical Society of America*, 87, 1738–52.

Lathi, B. P. (2002) *Signal Processing and Linear Systems*. Oxford: Oxford University Press.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999) *Discrete-Time Signal Processing*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

21 Speech Synthesis

ROLF CARLSON AND
BJÖRN GRANSTRÖM

1 Introduction

This chapter will review some of the more popular approaches to speech synthesis, with an emphasis on methods useful in phonetic research. Speech synthesis is not only one of the important applications of speech and language research but, in our opinion, a very valuable tool in the study of phonetics. We will point to some present and future applications of text-to-speech technology and describe some current trends in speech synthesis research.

Speech synthesis, during the last decade, has moved out of the research department and into everyday applications, such as speech-based dialog systems and aids for the disabled. Some of these applications actually employ prerecorded messages. Although a professional phonetician could contribute to creating procedures for optimizing the quality of such services, we will not focus on such methods in this chapter, but concentrate on the general aspects of speech synthesis as used in, for example, text-to-speech systems.

Electronic speech synthesis has developed over the last 50 years. In the publications by Fant (1960), Holmes et al. (1964), Flanagan (1972), Klatt (1976), and Allen et al. (1987), the foundations for speech synthesis based on acoustical or articulatory modeling can be found. The paper by Klatt (1987), gives an extensive review of the developments of speech synthesis techniques at that point in time. A number of textbooks and review papers have since been published. A recent handbook of speech processing (Benesty et al., 2008) covers the field from a technical perspective. The books by Dutoit (1997) and Taylor (2008) cover many of the current challenges on an introductory and detailed level.

2 Speech Synthesis in Text-to-Speech Systems

The most widespread applications of speech synthesis techniques are in text-to-speech systems. Such systems can be thought of as comprehensive models of

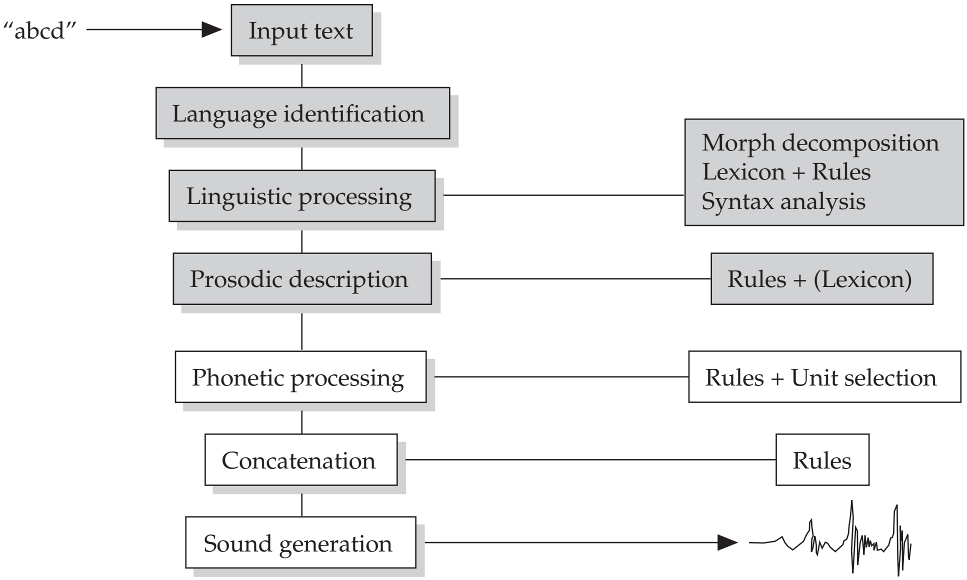


Figure 21.1 A generic text-to-speech system. Shaded modules are not discussed in detail in this chapter.

the process of reading aloud. Advanced versions of text-to-speech systems will hence contain components that are based on more than phonetic knowledge, even with a relatively wide definition of the phonetic sciences. In Figure 21.1, we have outlined a generic text-to-speech system. In its details it does not correspond to any specific system, but contains components found in many systems. In actually implemented systems, still on the research level or commercialized, the developers have put varying emphasis on the different modules and have also found radically different solutions to the posed functional demands.

3 Components of a Generic Text-to-Speech System

Looking at Figure 21.1 from top to bottom, we first see the input text module. This component typically identifies text of different kinds, such as digits, acronyms, and names. This process is generally called text normalization. The input text can normally be mixed with other information, such as phonetic text or special symbols controlling either system functions or linguistic/phonetic processing. Some systems today have a multilingual capability. At least ideally, one would like language switching to take place automatically. Language identification is a research area in its own right and will not be described here, but is obviously useful in, for example, a text-to-speech system used for reading web pages of various origin. Multilingual components have been developed in projects such as ESPRIT/POLYGLOT (Boves, 1991) and can be part of a foreign name pronunciation system (Church, 1986; Carlson et al., 1989; Vitale, 1991; Lindström & Eklund, 2007). In some synthesis systems special phonetic units have been introduced in order to handle special pronunciation needs. The linguistic processing module varies a great deal in complexity and ambition in different systems. The balance between rules and lexicon is due to language structure and implementation constraints. The amount of syntactic analysis varies from simple local phrasing based primarily on function word identification to attempts at complete sentence parsing. The derived information is useful both for disambiguation of homographs and as an input to the prosodic description module. In this module the prosodic phrasing/stress and accent are determined. The components so far described (shaded in Figure 21.1) will not be discussed further in this chapter, but serve as a basis for phonetic processing in the text-to-speech system. The unshaded components will be described in greater detail below.

4 Notations for Rule-Based Parametric Speech Synthesis

Development tools for text-to-speech systems have received considerable attention. Historically the bases of such tools followed the development of phonological theory. The work on generative phonology and especially the publication of *The Sound Pattern of English* by Chomsky and Halle (1968) led to a new kind of synthesis system based on rewrite rules (Carlson & Granström, 1976). Their ideas inspired researchers to create special rule compilers for text-to-speech developments in the early 1970s. The software developed according to this basic principle varies depending on the developer's inclination. It is important to note that crucial decisions are often hidden in the systems – the rules may operate rule-by-rule or segment-by-segment. Other important decisions are based on the following questions: How is the backtrack organized? Are the default values in the phoneme library primarily referred to by labels or by features? These questions might seem trivial, but we see many examples of how the explicit design of a system influences the thinking of the researcher. With the greater emphasis on prosodic modeling and the related development of nonlinear phonology (Pierrehumbert, 1987), synthesis procedures inspired by such theories were created, as in the systems described by Hertz et al. (1985), Hertz (1991), and van Leeuwen and te Lindert (1993). The common feature of these notations was that they keep information on different linguistic levels (tiers) separate in a more explicit way than the essentially linear representation based on generative phonology. This gives potentially higher flexibility, but also more complex notations.

Modeling segmental coarticulation and other phonetic factors is an important part of a text-to-speech system. The control part of a parametric synthesis system calculates the parameter values at each time frame. Two main types of approach can be distinguished: rule-based methods that use an explicit formulation of existing knowledge, and the data-based methods that replace rules by a collection of segment combinations of different unit lengths. Clearly, from a phonetic point of view both approaches have their advantages. Models based on rules force the researcher to understand the underlying principles, and corpus methods bypass this problem to some extent and give much better intelligibility and naturalness. We will discuss corpus-based methods further under the heading "synthesis by concatenation" (section 7 below). If the parametric data is coded in terms of targets and slopes, we need methods to calculate the parameter tracks. The efforts by Holmes et al. (1964) and the filtered square wave approach by Liljencrants (1969) are some classical examples in this context. Some of the problems of this approach will be discussed under the heading "articulatory models" (section 6.4).

5 Prosodic Descriptions and Implementations

Prosodic aspects of speech, primarily segmental duration and fundamental frequency (f_0) contours, are an important topic in speech synthesis research. Different speaking styles, like reading or dialog speech, different dialects, different emotions or attitudes, are all characterized by different kinds of prosody. The general belief is that prosody is the key to the naturalness that is often lacking in current text-to-speech systems. Durational models, often based on the segmental duration model described by Klatt (1979), have been developed for several languages. This work is often based on labeled speech corpora, where model predictions and corpus durations can be matched (van Santen & Olive, 1990). The standard deviation of the prediction error is often found to be on the order of 25 ms (Carlson, 1991). Whether this is mostly due to inherent durational variability or to disregard for some important linguistic/pragmatic factors is not clear. Intonation is inherently more difficult to model. Automatic analysis of f_0 is still not reliable for all voices and substantial variability exists among speakers. Local segmental context often affects the f_0 tracings, often referred to as inherent pitch or microprosody. Speech synthesis, as an alternative to speech analysis, has actually been used as a tool to better understand the important aspects of pitch contours. One important example of such work is the tradition at the Institute for Perception Research (IPO, Eindhoven) on perceptually valid stylization (Collier, 1990). Several other aspects of speech, such as voice source adjustments resulting in amplitude variations or variation in spectral shape, are also prosodically important. General increase of articulatory distinctness in focal positions and relaxation of articulators towards the end of phrases are often observed and are also important to model in speech synthesis.

6 Sound Generation Techniques in Parametric Synthesis

Sound generation in speech synthesis can be divided into three main classes: waveform coding, analysis–synthesis, and synthesis by rule. In waveform coding natural speech is used, digitally coded in a form that makes storage and manipulation efficient. The analysis–synthesis method is defined as a method in which human speech is transformed into parameter sequences, which are stored. The output in such a system is created by a synthesis based on the prestored parameters. In a synthesis-by-rule system, the output is generated with the help of rules which control a synthesis model such as a vocal tract model, a terminal analog, or some kind of coding.

It is not an easy task to place different synthesis methods into unique classes. Some of the common "labels" are often used to characterize a complete system rather than the model it stands for. A rule-based system using waveform coding is a perfectly possible combination, as is speech coding using a terminal analog or a rule-based diphone system using an articulatory model.

The sound-generating part of a synthesis system can be divided into two subclasses depending upon the dimensions in which the model is controlled. A vocal tract model can be controlled by spectral parameters such as frequency and bandwidth, or shape parameters such as size and length. The source model that excites the vocal tract usually has parameters to control the shape of the source waveform. The combination of time-based and frequency-based controls is powerful in the sense that each part of the system is expressed in its most explanatory dimensions. A drawback of the combined approach can be that it makes interaction between the source and the filter difficult. However, the merits seem to outweigh this.

6.1 Voice source models

The traditional voice source model has been a simple or double impulse. This is one reason why different voices produced by parametric text-to-speech systems from the last decade lack naturalness to a great extent. While the male voice sometimes has been regarded as being generally acceptable, an improved glottal source will open the way to more realistic synthesis of child and female voices and also to more naturalness and variation in male voices.

Most source models work in the time domain with different controls to manipulate the pulse shape (Rosenberg, 1971; Rothenberg et al., 1975; Holmes, 1973; Klatt & Klatt, 1990). One influential voice source model is the LF-model (Fant et al., 1985; Gobl, 2003). It has a truncated exponential sinusoid followed by a variable cut-off 6dB/octave low-pass filter modeling the effect of the return phase, i.e., the time from maximum excitation of the vocal tract to complete closure of the vocal folds. Figure 21.2 explains the function of the control parameters. In addition to the amplitude and fundamental frequency control, two parameters influence the amplitudes of the two to three lowest harmonics, and one parameter the high-frequency content of the spectrum. Another vocal source parameter is the diplophonia parameter with which creak, laryngalization, or diplophonia can be simulated (Klatt & Klatt, 1990). This parameter influences the function of the voiced source in such a way that every second pulse is lowered in amplitude and shifted in time.

The acoustic interactions between the glottal source and the vocal tract also have to be considered (Bickley & Stevens, 1986). One of the major factors in this

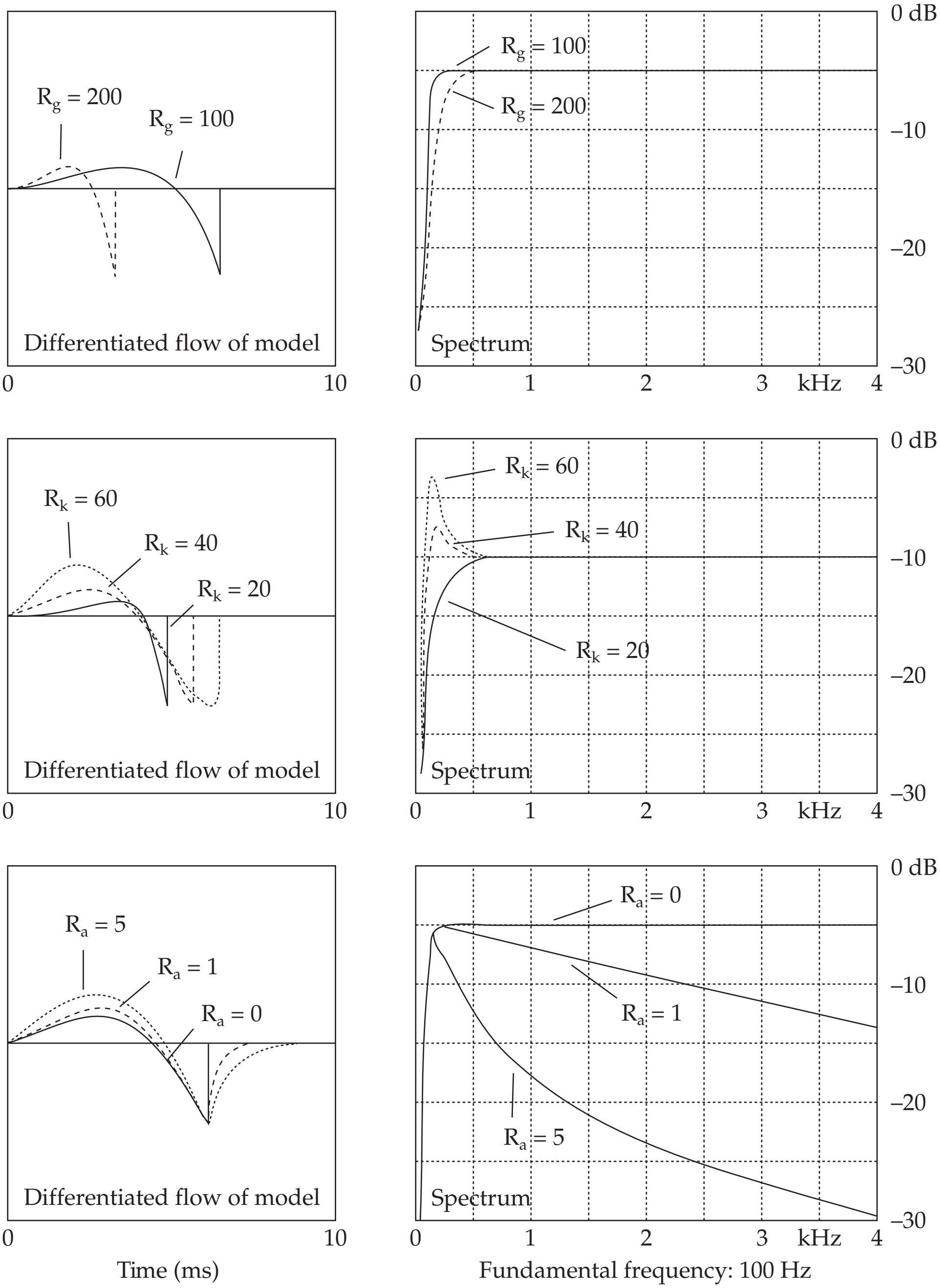


Figure 21.2 The influence of the parameters RG, RK, and FA on the differentiated glottal flow pulse shape, and spectrum. The spectra are pre-emphasized by 6 dB/octave. (After Gobl & Karlsson, 1991)

respect is the varying bandwidth of the formants. This is especially true for the first formant which can be heavily damped during the open phase of the glottal source. However it is not clear to what extent such a variation can be perceived by a listener. Listeners tend to be rather insensitive to bandwidth variation (Flanagan, 1972). When more complex models are to be included, the output from the model has to change from a glottal flow model to a model of the glottal opening. The subglottal cavities can then be included in an articulatory model.

Noise sources have attracted much less research effort compared to the voiced source. However, some aspects have been discussed by Stevens (1971), Shadle (1985), and Badin and Fant (1989). Typically, simple white noise is filtered by resonances which are stationary between each parameter frame. Only a few synthesizers have some interaction between the voice source and the noise source, but the interaction is rather primitive. Realization of transient sounds and aspiration dependent on glottal opening are still under development.

6.2 *Terminal analog formant synthesizers*

The traditional text-to-speech systems use a terminal analog. The ambition with this kind of synthesizer is only that it should be able to produce the sounds (speech spectra) that are found in natural speech. The internal structure is not a model of acoustic speech production in the vocal tract. The basic concept is the combination of sound sources and filters, describing the transfer function. Building on the classical configuration by Klatt (1980), this principle is exemplified in Figure 21.3.

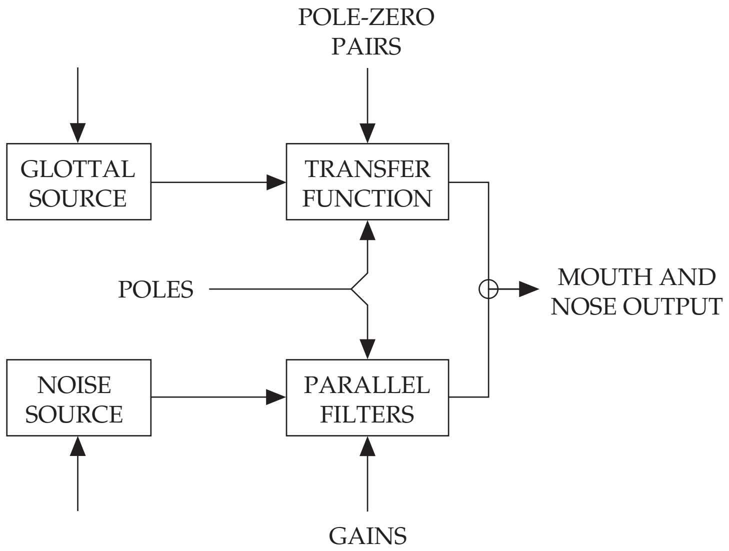


Figure 21.3 Block diagram of the main components of a terminal analog speech synthesizer such as KLSYN88 (Klatt & Klatt, 1990). The vertical arrows on the sides of the boxes indicate arrays of control parameters. (After Stevens & Bickley, 1991)

The vocal tract transfer function is simulated by a sequence of second-order filters in cascade while a parallel structure is used mostly for the synthesis of consonants. One important advantage of a cascade synthesizer is the automatic setting of formant amplitudes. The disadvantage is that it sometimes can be difficult to do detailed spectral matching between natural and synthesized spectra because of the simplified model. Parallel synthesizers such as the one developed by Holmes (1983) do not have this limitation.

The Klatt model has been widely used in research both for general synthesis purposes and for perceptual experiments. A simplified version of this system was used in all commercial products that stem from MIT: MITalk (Allen et al., 1987) and DECtalk.

A formant terminal analog, GLOVE (Carlson, Granström, & Karlsson, 1991), based on the OVE synthesizer (Liljencrants, 1968), has been developed at KTH and was used in text-to-speech modeling (Carlson, Granström, & Hunnicutt, 1982, 1991). In Figure 21.4, the structure of the Glove synthesizer is shown. The controllable parameters are indicated by two-letter symbols. To the left, the two sound sources can be seen. For mixed excitation the sources are connected in two ways. The parameter NM flow-modulates the noise source, typical for voiced fricatives. The parameter NA adds noise to the glottal source, as in breathy or whispered voices. The five parameters above the voice source are the glottal parameters referred to in the voice source section. The sound source signals are fed into the three parallel branches with poles and zeros. All are controlled by amplitude parameters (AN, A0, AH and, AC). The upper branch is primarily used for introducing an extra pole (and zero) in nasals and nasalized sounds, the middle branch is the main branch for sounds produced with glottal excitation, and the lowest branch models

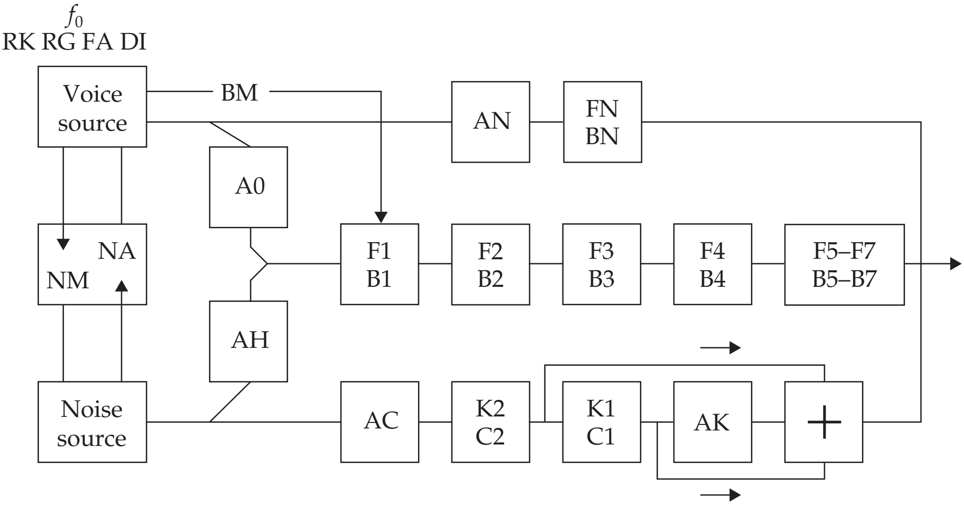


Figure 21.4 Block diagram of the main components of the terminal-analog speech synthesizer GLOVE. (Carlson, Granström, & Karlsson, 1991)

sounds with supraglottal excitation, such as stops and fricatives. This basic configuration can be augmented in several different ways. The interaction between the source and the vocal tract, which can be substantial, is in this case only modeled by the BM parameter that modulates the bandwidth of the first formant, dependent on glottal opening or, more precisely, glottal flow. The main difference between the Klatt and KTH traditions can be found in how the consonants are modeled. In the OVE case, a fricative is filtered by a zero–pole–pole configuration rather than the parallel branch in the Klatt synthesizer.

With the expanded capabilities of the terminal analog synthesizers, it is possible to simulate most human voices, and to replicate an utterance without noticeable quality reduction. However, it is interesting to note that some voices are easier to model than others. Despite the progress, speech quality is not good enough in all applications of text-to-speech. The limited success in formant-based synthesis can largely be explained by incomplete phonetic knowledge – it should be noted that the transfer of knowledge from phonetics to speech technology has not been an easy process. Another reason is that the efforts using formant synthesis have not fully explored alternative control methods to explicit rule-based description.

6.3 Higher-level parameters

Since the control of a formant synthesizer can be a very complex task, some efforts have been made to help the developer. The "higher-level parameters" described by Stevens and Bickley (1991) and Stevens (2002), for example, explore an intermediate level that is more understandable from the developer's point of view compared to the detailed synthesizer specifications. The goal with this approach is to find a synthesis framework to simplify the process and to incorporate the constraints that are known to exist within the process. A formant frequency should not have to be adjusted specifically by the rule developer depending on nasality or glottal opening. This type of adjustment might be better handled automatically according to a well-specified model. The same process should occur with other parameters such as bandwidths and glottal settings. The approach requires detailed understanding of the relation between acoustic and articulatory phonetics.

6.4 Articulatory models

Ultimately an articulatory model will be the most interesting solution for the sound-generating part of text-to-speech systems. Development is going forward in this area, but is still hampered by the lack of reliable articulatory data and appropriate control strategies. One possible solution that has attracted interest is to automatically train neural networks to control such a synthesizer. The work by Rahm et al. (1991) and Bailly et al. (1991) explores such methods. One promise of articulatory synthesis is that the controlling rules should be more simple and natural. Many of the phonetic details observed in natural speech will follow automatically if they depend on articulatory constraints included in the model.

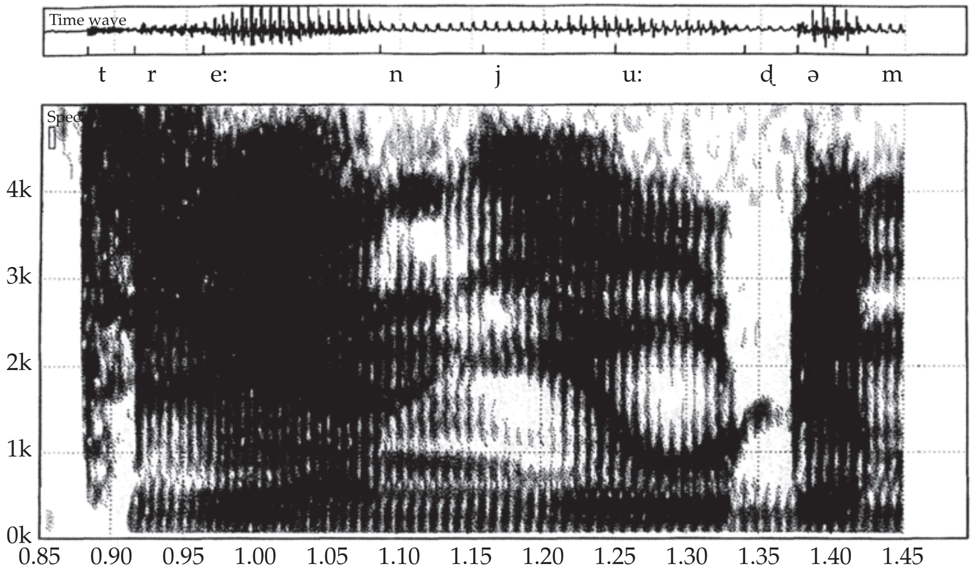


Figure 21.5 Spectrogram of a fragment of the Swedish sentence "Efter arbetet rengjorde målaren . . .". [tre:nju:ɖəm]. Note the changes of formant cavity affiliation and also the first formant change in the nasalized vowel.

One case in point is the modeling of transitions between speech sounds and articulations. In the output of a formant synthesizer it is often observed that the formant transition between phones does not compare very well to natural transitions, even if the segmental target values are well predicted. In the spectrogram in Figure 21.5, taken from a segment of natural speech, such a situation is obvious. In the transition between [j] and [u:], at approximately 1.23 seconds, we can see that an interpolation between the [j] and the [u:] target will look very different in all formants except F1. The reason is that the formants change cavity affiliation. It is, for example, possible to imagine a continuity between F3 of [j] and F2 of [u:].

Articulatory models which are developed today stem from the basic work carried out at laboratories such as Bell Labs, MIT, and KTH more than 40 years ago. In these models an approximation of the vocal tract is used either to calculate the corresponding transfer filter or to filter a source waveform directly. Different vocal tract models have been used based on varying assumptions and simplifications. The models by Flanagan et al. (1975), Coker (1976), Mermelstein (1973), and Maeda (1990) have been studied by many researchers in the development of current articulatory synthesis.

The term "articulatory modeling" is often used in a rather loose way. The situation is explained in Figure 21.6. Often a so-called articulatory model only models a simplified area function, rather than describing the movements of articulators.

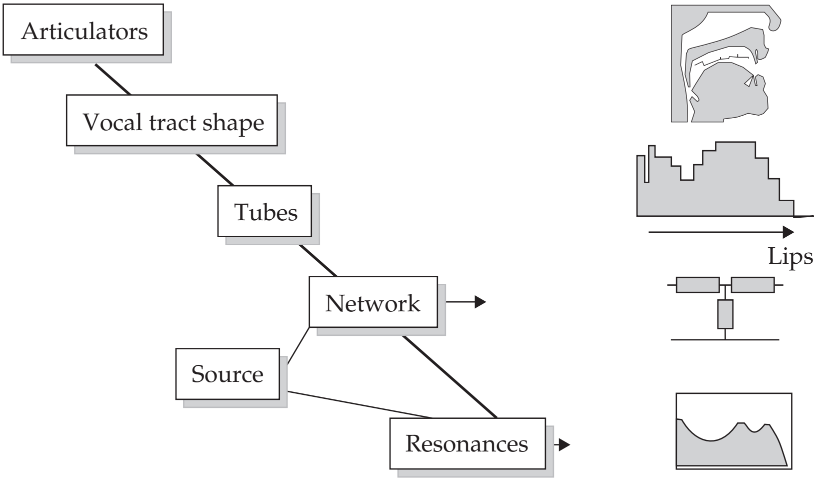


Figure 21.6 Different levels of the representation in an articulatory model for text-to-speech.

Also, the distinction between static and dynamic models must be kept in mind when a synthesis approach is discussed. A complete model has to include several transformations from the control signal to the actual speech output. The relation between an articulatory gesture and a sequence of vocal tract shapes has to be modeled. Each shape should be transformed into some kind of tube model which has its acoustic characteristics. The vocal tract is then modeled in terms of an electronic network. At this point, the developer can choose to use the network as such to filter the source signal. Alternatively, the acoustics of the network can be expressed in terms of resonances which can control a formant-based synthesizer. The main difference is the domain, time or frequency, in which the acoustic events are simulated.

The developer has to choose at which level the controlling part of the synthesis system should connect to the synthesis model. All levels are possible and many have been used. One of the pioneering efforts using articulatory synthesis as part of a text-to-speech system was that of Bell Labs (Coker, 1976). Lip, jaw, and tongue positions were controlled by rule. The final synthesis step was done by a formant-based terminal analog. Efforts by Lin and Fant (1992) used a parallel synthesizer with parameters derived from an articulatory model.

In the development of articulatory modeling for text-to-speech we can take advantage of parallel work on speech coding based on articulatory modeling (Sondhi & Schroeter, 1987). This work not only focuses on synthesizing speech but also on how to extract appropriate vocal tract configurations. Thus, it will also help us to get articulatory data through an analysis–synthesis procedure. Another noninvasive technique to obtain articulatory data is the magnetic resonance imaging (MRI) technique that has replaced X-ray for many medical investigations.

With this technique it is possible to get a three-dimensional representation of the vocal tract without the health hazards of X-ray. The "exposure time" is still several seconds, implying that true dynamic registrations are not yet possible (see also Stone, this volume). Combination with rapid motion capture techniques has introduced dynamic aspects in articulatory synthesis (Engwall, 2002). With a few exceptions, articulatory models have been only two-dimensional. Progress in computing power and three-dimensional techniques are now being exploited also in more comprehensive articulatory modeling (see section 8.3 on multimodal synthesis below). In this section we have not dealt with the important work carried out to model speech production in terms of volumes, masses, and airflow. The inclusion of such models still lies in the future beyond the next generation of text-to-speech systems, but the results of these experiments may well improve the current articulatory and terminal analog models.

6.5 Analysis–synthesis systems

Synthesis systems based on coding have as long a history as the vocoder. The underlying idea is that natural speech can be analyzed and stored in such a way that it can be assembled into new utterances. Synthesizers such as the systems from AT&T (Olive, 1977, 1990; Olive & Liberman, 1985), NTT (Nakajima & Hamada, 1988; Hakoda et al., 1990), and ATR (Sagisaka, 1988; Sagisaka et al., 1992) are based on the source-filter technique where the filter is represented in terms of linear predictive coding (LPC) or similar parameters. This filter is excited by a source model that can be of the same kind as the one used in terminal analog systems. The source must be able to handle all types of sounds: voiced, aspirative, and fricative, as well as sounds with combined excitation sources.

6.6 Corpus-based methods for parametric synthesis

Traditionally, speech synthesis has been based on very labor-intensive optimization work. The notion "analysis by synthesis" has not been explored except by manual comparisons between hand-tuned spectral slices and reference spectra. When increasing our ambitions to multilingual, multispeaker, and multistyle synthesis it is obvious that we want to find at least semi-automatic methods to collect the necessary information, using speech and language corpora. The work by Holmes and Pearce (1990) is a good example of how to speed up this process. With the help of a synthesis model, the spectra are automatically matched against analyzed speech. In Hertz (2002) and Carlson and Granström (2005) models are discussed that take advantage of both rule-based and corpus-based methods. Automatic techniques such as these will probably also play an important role in making speaker-dependent adjustments. One advantage with these methods is that the optimization is done in the same framework as that to be used in the production. The synthesizer constraints are thus already imposed in the initial state.

Methods for pitch-synchronous analysis will be of major importance in this context. Experiments such as the one presented by Talkin and Rowley (1990) will lead to better estimates of pitch and vocal tract shape. Several efforts on formant tracking techniques can be found in a review by O'Shaughnessy (2008). Minkyu et al. (2005) recently presented a method for formant tracking using context-dependent phonemic information. These automatic procedures will, in the future, make it possible to gather a large amount of data. Lack of glottal source data is currently a major obstacle for the development of parametric speech synthesis with improved naturalness.

A collection of parameter data from analyzed speech corpora provides a good basis to look for coarticulation rules and context-dependent variations. The collection of speech corpora also facilitates possibilities to test duration and intonation models.

7 Synthesis by Concatenation

The commercial introduction of rule-based synthesis paved the way for several applications especially as aids for the functionally handicapped. At the same time it was clear that the quality needed to be improved to be accepted by the general public.

The most radical solution to the synthesizer problem is simply to have a set of prerecorded messages stored for reproduction. Simple coding of the speech wave might be performed in order to reduce the amount of memory needed. The quality is high, but the usage is limited to applications with few messages. If units smaller than sentences are used, the quality degenerates because of the problem of connecting the pieces without distortion and of overcoming prosodic inconsistencies. One important, and often forgotten, aspect in this context is that a vocabulary change can be an expensive and time-consuming process, since the same speaker and recording conditions have to be used as with the original material. The whole system might have to be completely rebuilt in order to maintain equal quality of the speech segments. We will not discuss these methods further in this chapter.

A solution to improve the speech quality compared to parametric speech was to use fragments of natural speech as building blocks, units, and to concatenate them with or without signal processing. Initially units of equal length were selected representing a sequence of two phonemes, called diphones. Multiple diphones could be collected covering variations due to stress, duration, or other contextual variations. Considerable success was achieved by systems that based sound generation on concatenation of such units (Moulines et al., 1990). Sophisticated techniques have been developed to manipulate these units. The PSOLA (Carpentier & Moulines, 1990) methods are based on a pitch-synchronous overlap-add approach for concatenating waveform pieces. The frequency domain approach, FD-PSOLA, is used to modify the spectral characteristics of the signal; the time domain approach, TD-PSOLA, provides efficient solutions for real-time implementation of synthesis systems.

The importance of PSOLA in phonetic research lies in its possibility of manipulating prosodically interesting parameters (duration, fundamental frequency, and intensity) of natural speech without losing much of the original sound quality. In this respect it is quite superior to the speech coding techniques such as LPC that have previously been used for the same purpose. It must be stressed that it is a nontrivial task to perform optimal PSOLA synthesis. Many groups have been stimulated by the excellent pioneering work of Carpentier and Moulines (1990) and have demonstrated much less convincing results. One of the key problems lies in precisely defining each glottal pulse in time, something that is theoretically difficult to do for natural voices.

The MBROLA technique (Dutoit, 1997) became a successful solution for text-to-speech systems in several products. In this method special care is taken to model the phase of each unit to reduce distortion, and the need for accurate pich period estimation is reduced. As a service for the research community the MBROLA project was initiated by the TCTS Lab of the Faculté Polytechnique de Mons (http://tcts.fpms.ac.be/synthesis/mbrola.html). The goal is to obtain a set of speech synthesizers for as many languages as possible, and provide them free for non-commercial applications.

One of the major problems in concatenative synthesis is to make the best selection of units and to describe how to combine them. Two major factors create problems: distortion because of segmental and prosodic discontinuities at the connecting points, and distortion because of the limited size of the unit set. Systems using elements of different lengths depending on the target phoneme and its function have been explored by several research groups. In a paper by Olive (1990), a method was described to concatenate "acoustic inventory elements" of different sizes. The system developed at ATR is also based on nonuniform units (Sagisaka et al., 1992).

Special methods to generate a unit inventory have been proposed by the research group at NTT in Japan (Nakajima and Hamada, 1988; Hakoda et al., 1990). The synthesis allophones are selected with the help of the context-oriented clustering method, COC. The COC searches for the phoneme sequences of different sizes that best describe the phoneme realization.

The COC approach is a good illustration of a current trend in speech synthesis: automatic methods based on corpora. The studies are concerned with much wider phonetic contexts than before. (It might be appropriate to remind the reader of similar trends in speech recognition.) It is not possible to take into account all possible coarticulation effects by simply increasing the number of units, as at some point the total number might be too high or some units might be based on a very few observations. In this case a normalization of data might be a good solution before the actual unit is chosen. Thus the system contains rules.

As the synthesis approach to some extent is turned into a search problem, the system gets a close similarity to speech recognition. Similar theories can be used for both research fields. A major step forward was made by Hunt and Black (1996) when a general method for unit selection was presented. A general search algorithm was presented taking into account both the target cost depending on the unit itself and the join cost depending on the specific context. The method also opens up the possibility of including features on different levels spanning everything from acoustic realization to linguistic analysis. An interesting aspect is that in some cases the classical parameters such as formants can be explored as descriptive features. Many research efforts have recently been focused on how to gather and how to select optimal units (e.g., Black & Lenzo, 2003).

Since the unit-selection technique in principle is built on collected exemplars, we will always reach a limit for coverage due to realistic corpus size, memory size, and search time. Thus we find a new trend in speech synthesis where we again return to a parametric representation such as a spectral and source description. However, this time the synthesis is described in a statistical model, mostly the Hidden Markov Model (HMM) also used in speech recognition (Tokuda et al., 2000, 2002). In these approaches each phoneme is represented by a number of states (mostly three) and the transitions between each state are dependent on a trained probability. Contextual variation is handled by context-dependent models, mostly triphone models. Phonetic features can be used to group these into shared models. In some cases (e.g., Acero, 1999; Deng et al., 2003), formant tracking is used to optimize the selection. The system can be built using traditional speech recognition methods.

8 Trends in Speech Synthesis

During the last few decades speech synthesis has developed from manually controlled research tools to advanced text-to-speech products. The change is strongly related to our increased understanding of the speech production process, but perhaps to an even greater extent to developments in computer technology and signal-processing techniques. In Figure 21.7 we indicate some of the more important trends in these developments. It is interesting to note that these developments are incremental rather than revolutionary – any of the old methods and problem areas indicated in the figure are still valid.

8.1 Multilingual synthesis

Many societies in the world are increasingly multilingual. The situation in Europe is an especially striking example of this, most of the population being in touch

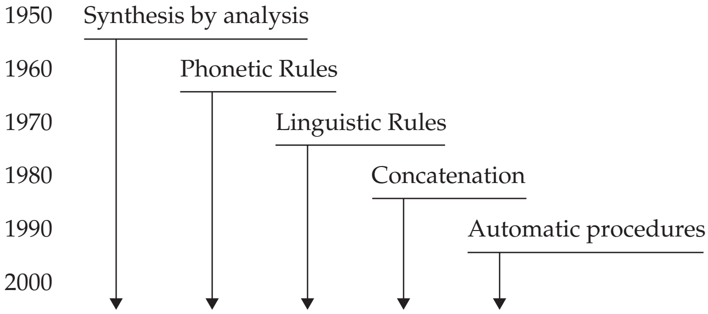


Figure 21.7 Important trends in speech synthesis developments.

with more than one language. This is natural in multilingual societies such as Switzerland and Belgium. Most schools in Europe have foreign languages on their mandatory curriculum from the early years. With the opening of the borders in Europe, more and more people will be in direct contact with several languages on an almost daily basis. For this reason, text-to-speech devices, whether they are used professionally or not, ought to have a multilingual capability.

Based on this understanding, many synthesis efforts are multilingual in nature. The Polyglot project supported by the European ESPRIT program was an early joint effort by several laboratories in several countries. The common software in this project was, to a great extent, language independent and the language-specific features were specified by rules, lexica, and definitions rather than in the software itself. This was the key to the multilingual effort at KTH. The synthesis developments pursued in companies are now frequently multilingual, serving the global market. As examples of noncommercial, freely available multilingual speech synthesis we refer to the already mentioned MBROLA project and to Festival (http://www.cstr.ed.ac.uk/projects/festival).

8.2 Style and personality

Currently available text-to-speech systems are not characterized by a great amount of flexibility, especially not when it comes to varying the voice or speaking style. On the contrary, the emphasis has been on a neutral way of reading, modeled after reading of nonrelated sentences. There is, however, a very practical need for different speaking styles in text-to-speech systems. Such systems are now used in a variety of applications and many more are projected as the quality is developed. The range of applications demands a variation close to that found in human speakers. General use in reading stock quotations, weather reports, email or warning messages are examples in which humans would choose rather different ways of reading. The speaking style of a spoken dialog system is expected to be very different from reading a text. Different voices are also important in speech prostheses so that nonspeaking persons in the same environment have different voices. Apart from these practical needs in text-to-speech systems, there is the scientific interest in formulating our understanding of human speech variability in explicit models.

The current ambition in speech synthesis research is to model natural speech on a global level, allowing changes of speaker characteristics and speaking style. One obvious reason is the limited success in enhancing general speech quality by only improving the segmental models. The speaker-specific aspects are regarded as playing a very important role in the acceptability of synthetic speech.

One interesting effort to include speaker characteristics in a complex system has been reported by the ATR group in Japan. The basic concept is to preserve speaker characteristics in interpreting systems (Abe et al., 1990). Their proposed voice conversion technique consists of two steps: mapping code-book generation of LPC parameters, and a conversion synthesis using the mapping code book. The method has been extended from a frame-by-frame transformation to a segment-by-segment transformation (Abe, 1991). A voice conversion system has been proposed that combines the PSOLA technique for modifying prosody with a source-filter decomposition which enables spectral transformations (Valbret et al., 1992). Other solutions to this problem are discussed in, for example, Ye and Young (2004) and in Toda et al. (2007).

One concern with this type of effort is that the speaker characteristics are specified through training without a specific higher-level model of the speaker. It would be helpful if the speaker characteristics could be modeled by a limited number of parameters. Only a small number of sentences might in this case be needed to adjust the synthesis to one specific speaker. The needs in both speech synthesis and speech recognition are very similar in this respect.

For many applications of speech synthesis, paralinguistic information, such as attitudes and emotions, is important. Experiments using DECtalk have been reported by Cahn (1990) in which a special "affect editor" was developed to control the synthesizer. Its success in generating recognizable affects was confirmed in an experiment in which the affect intended was perceived as such for the majority of the presentations. Similar efforts have been reported by Murray et al. (1991). The HAMLET system was developed for use in speech prostheses for the non-vocal, and was designed for incorporation into communication systems. The system uses DECtalk as an output device just as in the experiments by Cahn. Any of six emotions can be selected from a menu. The corresponding rules then operate on the phonemes and the voice quality settings, which are sent to the text-to-speech system. Recently emotion synthesis has attracted new interest and approaches have been suggested for corpus-based synthesis (Iida et al., 2003).

8.3 Multimodal synthesis

As we interact with others, we routinely make use of several of our sensory modalities in the process of communicating and exchanging information. A full account of the speech communication process must therefore include multiple modalities. The visible articulatory movements are mainly those of the lips, jaw, and tongue. However, these are not the only visual information carriers in the face during speech. Much prosodic information related to prominence and phrasing, as well as communicative information such as signals for feedback, turn-taking, emotions and attitudes, can be conveyed by, for example, nodding of the head, raising and shaping of the eyebrows, eye movements and blinks. As in acoustic speech synthesis, there are different ways of implementing "talking heads." One method is 2D picture manipulation/concatenation similar to concatenation techniques in acoustic synthesis. The Video Rewrite system is an early example of this (Bregler et al., 1997). More flexible and apt to experimentation are the 3D model approaches where facial appearance can be manipulated by parameters. A few of these models are very close to physiological speech production models, based on, for example, muscular deformations (e.g., Waters, 1987). More mainstream are the models based on 3D meshes. Many are based on the original work of Parke (1982). One well-known example is Baldy (Cohen & Massaro, 1993). An excellent introduction to the topic is given in Beskow (2003), where both the basic technology and some

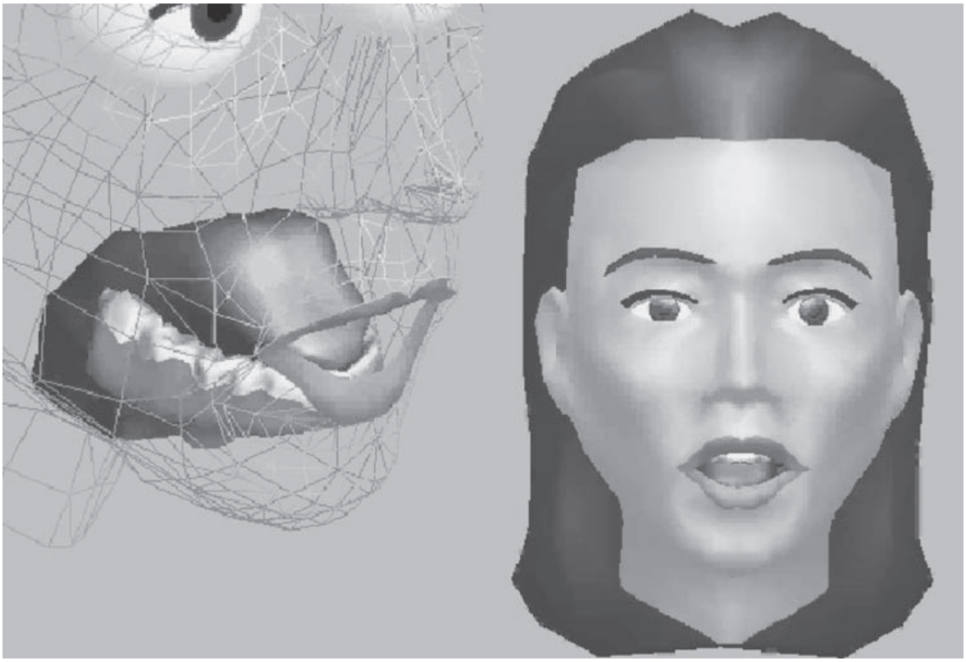


Figure 21.8 Example of the KTH wire-mesh-based articulatory talking head model.

applications are described. One example of the KTH talking head can be seen in Figure 21.8. To some extent the multimodal synthesis comes quite close to articulatory synthesis. While focusing on the visible movements, many include an elaborate 3D description of the entire vocal tract. Data for these models are acquired by a combination of static (e.g., MRI) and dynamic techniques (ultrasound, EMA, and different motion capture techniques; Beskow et al., 2003). These models have been used in a multitude of studies, pointing to the visual impact on segment and prosody perception. In spoken dialog systems, where a computer takes the role of both speaker and listener, a talking head can ease the interaction by better showing both the phonetic and nonlinguistic aspects of communication. Some examples from our own work can be found in Granström and House (2007).

9 Concluding Remarks

In this review we have touched upon a number of different synthesis methods and research goals to improve current text-to-speech systems. It might be germane to remind the reader that nearly all methods are based on historic development, where new knowledge has been added piece by piece to old knowledge, rather than by a dramatic change of approach. Perhaps the most dramatic change is in the field of tools rather than in the understanding of the "speech code." However, considerable progress can be seen in terms of improved speech synthesis quality. Today, speech synthesis is a common facility outside the research world, especially in speaking aids for persons with disabilities and in telephone services for the general public. New synthesis techniques under development in speech research laboratories will play a key role in future multimodal spoken dialog systems, but will also continue to play an important role as a tool in phonetic research.

REFERENCES

Abe, M., Shikano, K., & Kuwabara, H. (1990) Voice conversion for an interpreting telephone. In W. N. Campbell (ed.), *ETRW on Speaker Characterization in Speech Technology*, Edinburgh, UK.

Abe, M. (1991) A segment-based approach to voice conversion. *Proceedings of ICASSP '91*.

Acero, A. (1999) Formant analysis and synthesis using hidden markov models. In *Proceedings of Eurospeech '99*, pp. 1047–50.

Allen, J., Hunnicutt, M. S., & Klatt, D. (1987) *From Text to Speech: The MITalk System*. Cambridge: Cambridge University Press.

Badin, P. & Fant, G. (1989) Fricative modeling: Some essentials. *Proceedings of the European Conference on Speech Technology*, Paris.

Bailly, G., Laboissière, R., & Schwartz, J. L. (1991) Formant trajectories as audible gestures: An alternative for speech synthesis, *Journal of Phonetics*, 19, 9–23.

Benesty, J., Sondhi, M., & Huang, Y. (2008) *The Springer Handbook of Speech Processing*. Berlin: Springer.

Beskow, J. (2003) Talking heads: Models and applications for multimodal speech synthesis. Doctoral dissertation, KTH.

Beskow, J., Engwall, O., & Granström, B. (2003) Resynthesis of facial and intraoral articulation from simultaneous measurements. In M. J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 431–4). Barcelona/Australia: Causal Productions.

Bickley, C. & Stevens, K. (1986) Effects of the vocal tract constriction on the glottal source: Experimental and modelling studies. *Journal of Phonetics*, 14, 373–82.

Black, A. & Lenzo, K. (2003) Optimal utterance selection for unit selection speech synthesis databases. *International Journal of Speech Technology*, 6, 357–63.

Boves, L. (1991) Considerations in the design of a multi-lingual text-to-speech system. *Journal of Phonetics*, 19.

Bregler, C., Covell, M., & Slaney, M. (1997) Video rewrite: Driving visual speech with audio. *Proceedings of SIGGRAPH '97*, Los Angeles, 353–60.

Cahn, J. E. (1990) The generation of affect in synthesized speech. *Journal of the American Voice Input/Output Society*, 8, 1–19.

Carlson, R. (1991) Duration models in use. *Proceedings of the International Congress of Phonetic Sciences*, Aix-en-Provence, 4, 278–81.

Carlson, R. & Granström, B. (1976) A text-to-speech system based entirely on rules. *Proceeding of ICASSP '76*, Philadelphia, PA.

Carlson, R. & Granström, B. (2005) Data-driven multimodal synthesis. *Speech Communication*, 47, 182–93.

Carlson, R., Granström, B., & Hunnicutt, S. (1982) A multi-language text-to-speech module. *Proceedings of ICASSP '82*, Paris, 3, 1604–7.

Carlson, R., Granström, B., & Hunnicutt, S. (1991) Multilingual text-to-speech development and applications. In A. W. Ainsworth (ed.), *Advances in Speech, Hearing and Language Processing*. London: JAI Press.

Carlson, R., Granström, B., & Karlsson, I. (1991) Experiments with voice modelling in speech synthesis. *Speech Communication*, 10, 481–9.

Carlson, R., Granström, B., & Lindström, A. (1989) Predicting name pronunciation for a reverse directory service. *Proceedings of the European Conference on Speech Communication and Technology*, Paris, 1, 113–16.

Carpentier, F. & Moulines, E. (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 453–67.

Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York: Harper & Row.

Church, K. (1986) Stress assignment in letter to sound rules for speech synthesis. *Proceedings of ICASSP '86*, 4, 2423–6.

Cohen, M. M. & Massaro, D. W. (1993) Modelling coarticulation in synthetic visual speech. In N. Magnenat-Thalmann & D. Thalmann (eds.), *Models and Techniques in Computer Animation* (pp. 139–56). Tokyo: Springer.

Coker, C. H. (1976) A model for articulatory dynamics and control. *Proceedings of IEEE*, 64, 452–60.

Collier, R. (1990) Multi-language intonation synthesis. *Journal of Phonetics*, 19.

Deng, L., Bazzi, I., & Acero, A. (2003) Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint. In *Proceedings of Eurospeech*, 73–6.

Dutoit, T. (1997) *An Introduction to Text-To-Speech Synthesis*. Dordrecth: Kluwer.

Engwall, O. (2002) Tongue talking: Studies in intraoral speech synthesis. Doctoral dissertation, KTH.

Fant, G. (1960) *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G., Liljencrants, J., & Lin, Q. (1985) A four parameter model of glottal flow. *Speech Transmission Laboratory Quarterly and Status Report STL-QPSR* 4.

Flanagan, J. L. (1972) *Speech Analysis, Synthesis and Perception*. Berlin: Springer.

Flanagan, J. L., Ishizaka, K., & Shipley, K. L. (1975) Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *Bell System Technical Journal*, 54, 485–506.

Gobl, C. (2003) The voice source in speech communication: Production and perception experiments involving inverse filtering and synthesis. Doctoral dissertation, KTH.

Gobl, C. & Karlsson, I. (1991) Male and female voice source dynamics. In J. Gauffin & B. Hammarberg (eds.), *Proceedings of the Vocal Fold Physiology Conference*. San Diego: Singular Publishing Group.

Granström, B. & House, D. (2007). Inside out: Acoustic and visual aspects of verbal and non-verbal communication. In J. Trouvain & W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 11–18.

Hakoda, K, Nakajima, S., Hirokawa, T., & Mizuno, H. (1990) A new Japanese text-to-speech synthesizer based on COC synthesis method. *Proceedings of ICSLP '90*, Kobe.

Hertz, S. R. (1991) Streams, phones, and transitions: Toward a new phonological and phonetic model of formant timing. *Journal of Phonetics*, 19.

Hertz, S. R. (2002) Integration of rule-based formant synthesis and waveform concatenation: A hybrid approach to text-to-speech synthesis. In *Proceedings of the IEEE 2002 Workshop on Speech Synthesis*, Santa Monica.

Hertz, S. R., Kadin, J., & Karplus, K. J. (1985) The Delta rule development system for speech synthesis from text. *Proceedings of IEEE*, 73.

Holmes, J. (1973) Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE Transactions on Audio and Electroacoustics*, AU-21, 298–305.

Holmes, J. (1983) Formant synthesizers, Cascade or Parallel. *Speech Communication*, 2, 251–73.

Holmes, J., Mattingly, I. G., & Shearme, J. N. (1964) Speech synthesis by rule. *Language and Speech*, 7, 127–43.

Holmes, W. J. & Pearce, D. J. B. (1990) Automatic derivation of segment models for synthesis-by-rule. *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans.

Hunt, A. & Black, A. (1996) Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of ICASSP '96*, Atlanta, 1, 373–6.

Iida A., Campbell, N., Higuchi, F., & Yasumuram M. (2003) A corpus-based speech synthesis system with emotion. *Speech Communication*, 40, 161–87.

Klatt, D. K. (1976) Structure of a phonological rule component for a synthesis-by-rule program. *IEEE Transactions*, ASSP-24.

Klatt, D. K. (1979) Synthesis by rule of segmental durations in English sentences. In B. Lindblom & S. Öhman (eds.), *Frontiers in Speech Communication Research* (pp. 287–300). New York: Academic Press.

Klatt, D. K. (1980) Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67, 971–55.

Klatt, D. K. (1987) Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America*, 82, 737–93.

Klatt, D. K. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, 820–57.

Leeuwen, H. C. van & Lindert, E. te (1991) Speech Maker: Text-to-speech synthesis based on a multi-level, synchronized data structure. *Proceedings ICASSP '91*, Toronto, 2, 781–4.

Liljencrants, J. (1968) The OVE III speech synthesizer. *IEEE Transactions on Audio and Electroacoustics*, AU-16, 137–40.

Liljencrants, J. (1969) Speech synthesizer control by smoothed step functions. *Speech Transmission Laboratory, Quarterly Progress and Status Report (STL-QPSR)*, 4, 43–50.

Lin, Q. & Fant, G. (1992) An articulatory speech synthesizer based on a frequency domain simulation of the vocal tract. *Proceedings of ICASSP '92*.

Lindström, A. & Eklund, R. (2007) The integration of foreign items: A corpus-based study of cross-lingual influence with examples from Swedish. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: An International Handbook*. Berlin: Mouton.

Maeda, S. (1990) Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. J., Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling* (pp. 131–49). Dordrecht: Kluwer.

Mermelstein, P. (1973) Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53, 1070–82.

Minkyu, L., Santen, J. van, Möbius, B., & Olive, J. (2005) Formant tracking using context-dependent phonemic information. *IEEE Transactions on Speech and Audio Processing*, 13, 741–50.

Moulines, E., Emerard, F., & Larreur, D. et al. (1990) A real-time French text-to-speech system generating high quality synthetic speech. *Proceedings of ICASSP '90*, 1, 309–12.

Murray, I. R., Arnott, J. L., Alm, N., & Newell, A. F. (1991) A communication system for the disabled with emotional synthetic speech produced by rule. *Proceedings of the European Conference on Speech Communication and Technology*, Genoa, 311–14.

Nakajima, S. & Hamada, H. (1988) Automatic generation of synthesis units based on context oriented clustering. *Proceedings of ICASSP '88*.

Olive, J. P. (1977) Rule synthesis of speech from dyadic units. *Proceedings of ICASSP '77*, 568–70.

Olive, J. P. (1990) A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans.

Olive, J. P. & Liberman, M. Y. (1985) Text-to-speech: An overview. *Journal of the Acoustical Society of America*, suppl. 1, 78, S6.

O'Shaughnessy, D. (2008) Formant estimation and tracking. In J. Benesty, M. Sondhi, & Y. Huang (eds.), *The Springer Handbook of Speech Processing* (pp. 213–27). Berlin: Springer.

Parke, F. I. (1982) Parametrized models for facial animation. *IEEE Computer Graphics*, 2, 61–8.

Pierrehumbert, J. B. (1987) *The Phonetics of English Intonation*. Bloomington: Indiana University Linguistics Club.

Rahm, M., Kleijn, B., & Schroeter, J. (1991) Acoustic to articulatory parameter mapping using an assembly of neural networks. *Proceedings of ICASSP '91*.

Rosenberg, A. E. (1971) Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 53, 1632–45.

Rothenberg, M., Carlson, R., Granström, B., & Lindqvist-Gauffin, J. (1975) A three-parameter voice source for speech synthesis. In G. Fant (ed.), *Proceedings of the Speech Communication Seminar, Stockholm, 1974*, vol. 2 (pp. 235–43). Stockholm: Almqvist & Wiksell.

Sagisaka, Y. (1988) Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proceedings of ICASSP '88*.

Sagisaka, Y., Kaiki, N., Iwahashi, N., & Mimura, K. (1992) ATR v-TALK speech synthesis system. *Proceedings of ICSLP '92*, Banff, 483–6.

Santen, J. van & Olive, J. P. (1990) The analysis of segmental effect on segmental duration. *Computer Speech and Language*, 4.

Shadle, C. H. (1985) The acoustics of fricative consonants. Doctoral dissertation, MIT.

Sondhi, M. M. & Schroeter, J. (1987) A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 35.

Stevens, K. N. (1971) Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, 50, 1180–92.

Stevens, K. N. (2002) Toward formant synthesis with articulatory controls. *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, 67–72.

Stevens, K. N. & Bickley, C. (1991) Constraints among parameters simplify control of Klatt formant synthesizer. *Journal of Phonetics*, 19.

Talkin, D. & Rowley, M. (1990) Pitch-synchronous analysis and synthesis for TTS systems. *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans.

Taylor, P. (2008) *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

Toda, T., Black A., & Tokuda, K. (2007) Voice conversion based on maximum likelihood estimation of speech parameter trajectory. *IEEE Transactions on Acoustics, Speech and Language Processing*, 15, 2222–35.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000) Speech parameter generation algorithms for HMM-based speech synthesis. *Proceedings of ICASSP*, 1315–18.

Tokuda, K., Zen, H., & Black, A. (2002) An HMM-based speech synthesis system applied to English. *2002 IEEE Speech Synthesis Workshop*, Santa Monica, California.

Valbret, H., Moulines, E., & Tubach, J. P. (1992) Voice transformation using PSOLA technique. *Proceedings of ICASSP '92*, San Fransisco, 1, 145–8.

Vitale, T. (1991) An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics*, 17, 257–76.

Waters, K. (1987) A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH '87)*, 21, 17–2.

Ye, H. & Young, S. (2004) High quality voice morphing. *International Conference on Acoustics Speech and Signal Processing*, Montreal.

22 Automatic Speech Recognition

STEVE RENALS AND SIMON KING

1 Introduction

Speech recognition – the transcription of an acoustic speech signal into a string of words – is a hard problem owing to several cumulative sources of variation. Specifying the units of speech, such as phonemes, words, or syllables, is not straightforward and, whatever units are chosen, identifying the boundaries between them is challenging. Furthermore the relation between the acoustic speech signal and a symbolic sequence of units is complex due to phenomena such as varying rates of speech, context-dependences such as coarticulation, and prosodic effects.

We can factor this variation across four principal dimensions. First, the "size" of the problem may be measured in terms of the number of words that may be recognized. One of the first recognizers to be widely commercially deployed (Lennig, 1990) had a recognition vocabulary of just two words ("yes" or "no," in response to a question), but was required to be robust to uncontrolled user responses. Other small vocabulary tasks include the recognition of digit sequences, or simple command and control applications. However, for there to be any degree of richness in a spoken interaction, much larger vocabularies are required: for example, the recognition of human-to-human (or, indeed, human-to-computer) conversations will typically require a vocabulary with about 50,000 words (or many more in an agglutinative language, such as Finnish). The size of the problem is also measured by the estimated *perplexity* of the language, a statistical measure of the number of words that may be expected to follow any word in the language. For example, a travel information system may have a very large vocabulary, due to the number of place names, but at most points in the interaction with such a system the expected number of possible words is small, except when a place name is likely to be spoken – hence the overall perplexity will be relatively low.

The second axis of variability is concerned with the speaking style. Early systems such as the yes/no recognizer mentioned above, or the DragonDictate office dictation system (Baker, 1989), were *isolated word* (discrete utterance) recognizers, in which pauses were left between each word, thus eliminating the problem of word segmentation, and reducing coarticulation. Although such systems are still appropriate for platforms with low computational resources (such as phones or embedded devices), most research is focused on *continuous speech* recognition, a much more difficult task since word boundaries may no longer be assumed to coincide with pauses. Thus the task of a continuous speech recognizer includes segmenting the stream into words as well as recognizing the individual words. These two tasks are usually treated jointly. Until the mid 1990s, virtually all research in large vocabulary speech recognition was focused on read speech or office dictation. Since then, research in large vocabulary speech recognition has increasingly focused on unplanned, conversational or spontaneous speech, such as transcription of a business meeting or a telephone conversation. This change in style has a dramatic effect on the word error rate (see section 8): a state-of-the-art system transcribing someone reading a newspaper has a typical word error rate of 5–10 percent; transcription of a human–human conversation has a word error rate in the range 10–30 percent (Chen et al., 2006; Hain et al., 2007). Other aspects of speaking style that have a significant effect on speech recognition accuracy include the rate of speech, and variability arising from semantic or emotional context.

A third dimension of variability arises from speaker characteristics and accent. It is the case that a *speaker-dependent* recognizer will make fewer errors than a more general *speaker-independent* system. Although speaker adaptation algorithms are extremely effective in improving accuracy (section 4.4), there is still a stark contrast between the adaptability and robustness of human speech recognition, compared with automatic systems. Furthermore there has been relatively little work in adapting speech recognition systems to children or older users. In addition to speaker characteristics arising from an individual's anatomy or physiology, there are also systematic variations based on a speaker's accent.

Finally, the difficulty of the problem depends on the acoustic characteristics of the speaker's environment and the transmission channel. Commercially available dictation systems assume that the speaker is wearing a close-talking microphone in a quiet room with low reverberation; much academic research has also assumed these acoustically benign conditions. Acoustic sources additional to the speaker make the task of speech recognition more difficult: the additional sources may be regarded as "noise" in some circumstances, but in other circumstances they may be competing talkers. Transmission channel variations, such as those that arise owing to the relative motion of the talker's head with respect to the microphone, or due to the telephone handset, also have a significant impact on speech recognition accuracy.

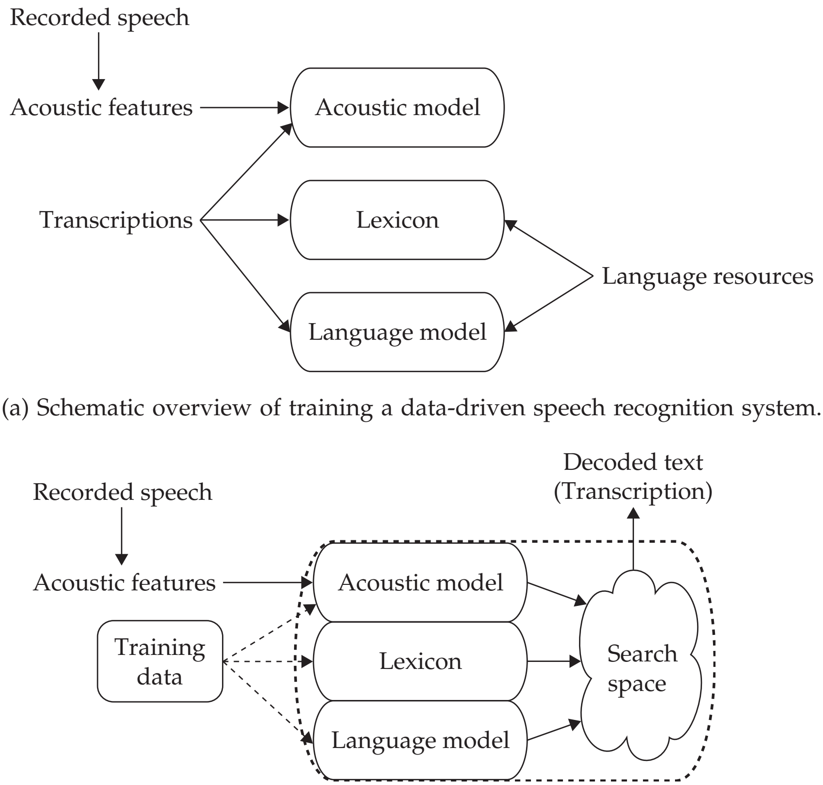
It is the case that automatic speech recognition (ASR) in "natural" conditions (such as recognizing the speech of an unknown talker at a cocktail party) has considerably lower accuracy than human speech recognition. However, considerable progress has been made over the past three decades: taking into account the increasing difficulty of tasks attempted, the state-of-the-art in speech recognition has improved at an average rate of 10 percent reduction in word error rate per year (Deng & Huang, 2004). These improvements have come about through a combination of algorithmic and modeling improvements, increased computational resources, and the availability of ever-larger orthographically transcribed speech corpora.

The state-of-the-art of speech recognition is based on statistical and data-driven methods, building on fundamental research carried out in the 1970s by researchers at Carnegie Mellon University (Baker, 1975) and IBM (Jelinek, 1976; Bahl et al., 1983). This approach is based on the *noisy channel* model of information theory, in which the human speech production system is regarded as a noisy channel between the sequence of words in the speaker's mind and the observed speech waveform. The function of the speech recognizer is to decode the speech signal in order to obtain the correct word sequence. In this approach, outlined further in section 2, the task of decoding the speech signal is decomposed into four interacting modules: *acoustic feature extraction* from the speech waveform (section 3); the *acoustic model*, which relates the acoustic features to the basic unit of speech employed, such as the phone (section 4); the *dictionary* (section 5); and the *language model* (section 6). An important aspect of this approach is that the acoustic and language models, in particular, may be learned from a corpus of training data. Once these models have been trained then the task of speech recognition may be expressed as a search problem in which the most probable transcription for an acoustic signal is found given the models (section 7). This process is illustrated in Figure 22.1.

Statistical speech recognition seems far from linguistics, and it is indeed the case that modern speech recognition systems make little explicit use of linguistic or phonetic knowledge. The acoustic models embody a simplistic "beads-on-a-string" view of speech processing (Ostendorf, 1999), in which speech is viewed as made up of a sequence of phonemes. However, the overall framework permits richer acoustic models, and there is a growing body of research focused on the development of acoustic models exploiting speech production knowledge (King et al., 2007). Likewise, although language models are usually based on flat *n*-gram models in which a word is predicted based on only two or three words of context, there is considerable work in the development of richer, more structured models (Rosenfeld, 2000). The reason that the state-of-the-art is still represented by these simple, linguistically impoverished models is because such models can be efficiently trained from extremely large amounts of data. Both hidden Markov acoustic models and *n*-gram language models are able to scale to massive amounts of training data, making ever more detailed models – with a concomitant increase in the number of parameters.

2 A Data-Driven Approach

The decomposition of the speech recognition problem into an acoustic model and a language model is very well matched to a statistical view of the problem. If **X** is a sequence of acoustic feature vectors (extracted from the recorded waveform),



(a) Schematic overview of training a data-driven speech recognition system.

(b) Schematic overview of speech-to-text transcription employing a trained data-driven speech recognition system.

Figure 22.1 Schematic overview of training (a) and recognition (b) in an ASR system.

then the aim of speech recognition – in the language of probability – is to recover the most probable sequence of words **W*** given **X**. We may write this as follows:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}), \tag{1}$$

that is, transcribe the spoken utterance using the word sequence **W*** that has the greatest posterior probability $P(\mathbf{W} \mid \mathbf{X})$. The art of speech recognition engineering involves finding ways to efficiently express, approximate, and evaluate posterior probabilities of this form. A first step, involving no approximations, is to re-express this conditional probability using Bayes' theorem:

$$P(\mathbf{W} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{W})P(\mathbf{W})}{P(\mathbf{X})}. \tag{2}$$

Now, observe that finding $\mathbf{W}^*$ in (1) does not demand that the exact value of the posterior probability is calculated for each word sequence. What is required is to be able to accurately *compare* the posterior probabilities of different word sequences given the same acoustics $\mathbf{X}$. Thus factors that are common to all word sequences do not need to be considered. Since the denominator in (2), $P(\mathbf{X})$, does not depend on $\mathbf{W}$ we may write:

$$P(\mathbf{W} \mid \mathbf{X}) \propto P(\mathbf{X} \mid \mathbf{W})P(\mathbf{W}) \tag{3}$$

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \underbrace{P(\mathbf{X} \mid \mathbf{W})}_{\text{acoustic model}} \underbrace{P(\mathbf{W})}_{\text{language model}} . \tag{4}$$

This decomposes the problem into two parts: an *acoustic model* which supplies $P(\mathbf{X} \mid \mathbf{W})$ and the *language model, $P(\mathbf{W})$*. The acoustic model is typically estimated using a corpus of transcribed speech; the language model, which is independent of the acoustics, may be estimated from a text corpus.

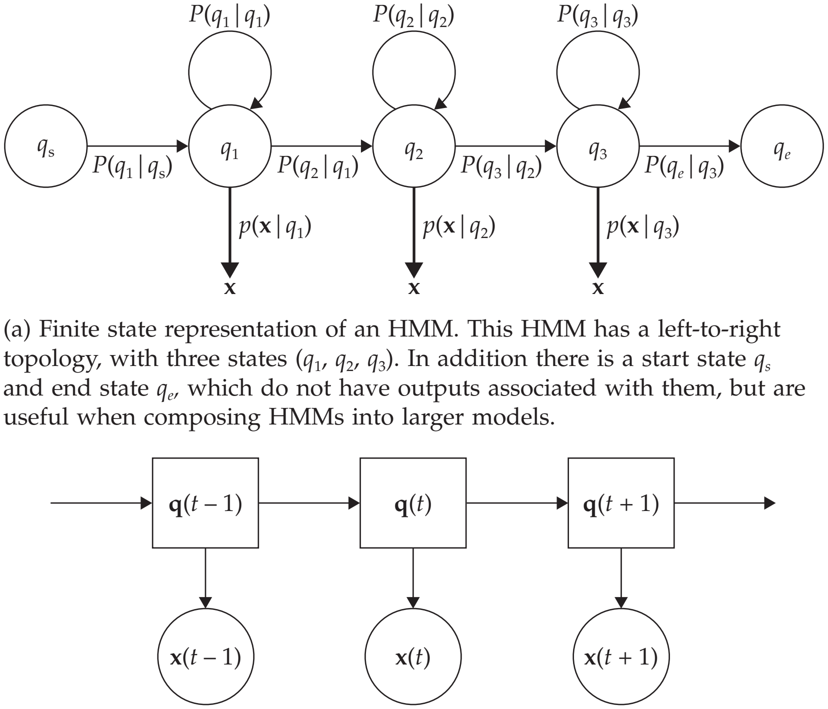
Decomposing the problem in this way is an example *of generative modeling,* a powerful technique used in statistical pattern recognition and machine learning (Bishop, 2006). In this approach it is assumed that there is an underlying model, $\mathcal{M}$, which *generates* the observed speech as represented by the sequence of acoustic feature vectors $\mathbf{X}$.[1] The acoustic model provides a continuous probability distribution over the space of sequences of acoustic feature vectors, giving the likelihood of a particular feature vector sequence being generated from a particular word sequence. The language model may be regarded as supplying a *prior probability* for each word sequence.

Generative modeling is conceptually simple, although at first it appears that the model is doing the opposite of what is expected. In this approach we do not think of the speech waveform going into the model and the recognized word sequence coming out. Instead, a model is constructed for each word sequence that we want to recognize. To do recognition, we ask the model for each word sequence in turn to generate the acoustics. Whichever model can generate the observed pattern with the highest probability is the winner.

2.1 Hidden Markov models

We do not have access to the true generative model – for speech, this would include components such as an accurate model of the speech articulators, of the air flow in the vocal tract, and of the linguistic plans used. In practice we use models that can supply a likelihood of the form $P(\mathbf{X} \mid \mathbf{W})$, are mathematically tractable, and may be trained or estimated from data. In addition the models must be made up of some basic units, since it is not possible to have a separately estimated model for each word sequence (of which there are an infinite number!).

Acoustic models for speech recognition are generally based on *hidden Markov models* (HMMs) (Baker, 1975; Poritz, 1988; Rabiner, 1989; Jelinek, 1998). HMMs are probabilistic finite state machines – more precisely, they are probabilistic finite



(a) Finite state representation of an HMM. This HMM has a left-to-right topology, with three states (q_1, q_2, q_3). In addition there is a start state q_s and end state q_e, which do not have outputs associated with them, but are useful when composing HMMs into larger models.

(b) Probabilistic dependency representation of an HMM, showing the relation between the state q and observation $\mathbf{x}$ variables.

Figure 22.2 Two representations of an HMM, emphasizing the finite state nature of the model (a) and the probabilistic dependences between the variables (b). Both figures illustrate the two assumptions underlying the use of HMMs: the first-order Markov process and the conditional independence of observation vectors given the current state.

state generators. An HMM consists of a set of states, connected by transitions. To generate an acoustic sequence from an HMM we first generate a state sequence by traversing a path from state to state, following the transitions. At each timestep a single feature vector is generated by an output (or emission) process that belongs to that state. By visiting a sequence of states in a particular order, we can generate a sequence of feature vectors. Thus HMMs are well suited to the generation of sequential data, such as speech. The *topology* of the HMM (i.e., the way that the transitions connect the states) can be used to provide a constraint on the type of sequences that the model can generate. This is illustrated in Figure 22.2a.

HMMs may be thought of as a "doubly stochastic" process. The generated state sequence is governed by the model topology and the transition probabilities associated with each state transition. The second stochastic process is concerned with the generation of acoustic features at each state. This takes the form of a multidimensional probability distribution over the space of acoustic feature vectors. The output probability distribution is the most important part of the hidden Markov model approach to speech recognition, and exists at the signal–symbol interface, linking the continuous space of acoustic vectors with the discrete space of HMM states. A simple form for the output probability distribution is a multidimensional Gaussian. In this case, given an HMM state q, the probability density of a d-dimension acoustic feature vector $\mathbf{x}$ may be written as:

$$p(\mathbf{x} \mid q) = \frac{1}{(2\pi)^{d/2} \,|\sum_q|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_q)^T \sum\nolimits_q^{-1} (\mathbf{x} - \boldsymbol{\mu}_q) \right), \tag{5}$$

where $\boldsymbol{\mu}_q$ is the mean vector and Σ_q is the covariance matrix for the Gaussian attached to state q. In practice, the required dimensionality for acoustic modeling is relatively high. For example, acoustic feature vectors consisting of 12 mel frequency cepstral coefficients (see section 3) and delta log energy would result in a 13-dimension Gaussian; if first and second derivatives of acoustic feature vectors are added, as is usual, this results in a 39-dimensional vector.

The use of hidden Markov models for acoustic modeling embeds two principal assumptions. First, it is assumed that the state sequence is a (first-order) Markov process: the probability of being in a particular state depends only on the previous one. Second, observations are assumed to be dependent on the current state only: given the current state an observed acoustic feature vector is *conditionally independent* of all past (and future) observations. We can rephrase this second assumption by saying that all information about the history of previous observed acoustic feature vectors is given by the current state. These assumptions are shown clearly if an HMM is represented in terms of probabilistic dependences (Figure 22.2b).

These assumptions are rather unrealistic, and on the face of it do not provide a good basis to model speech. However, they enable the model to become computationally and mathematically tractable and allow training from very large amounts of data. Much acoustic modeling research (discussed in section 4) is concerned with alleviating the effects of these assumptions.

2.2 Hierarchical modeling

There seems to be a large gap between the simple HMM illustrated in Figure 22.2 and the model of a word sequence required by equations (1–4). However, by using a hierarchical modeling approach it is possible to build word sequence models out of these simple HMM building blocks. To do this, we specify some fundamental units of speech, typically phone models. In the TIMIT corpus (Fisher et al., 1986) there are 61 phone classes in the hand transcriptions: for automatic speech recognition, this set is often reduced to 48 or 39 phone classes (Lee, 1989). Each phone class is represented by a hidden Markov model, typically with a three-state, left-to-right topology, as illustrated in Figure 22.2a. Word models are constructed from phone models by concatenating a sequence of HMMs, and word

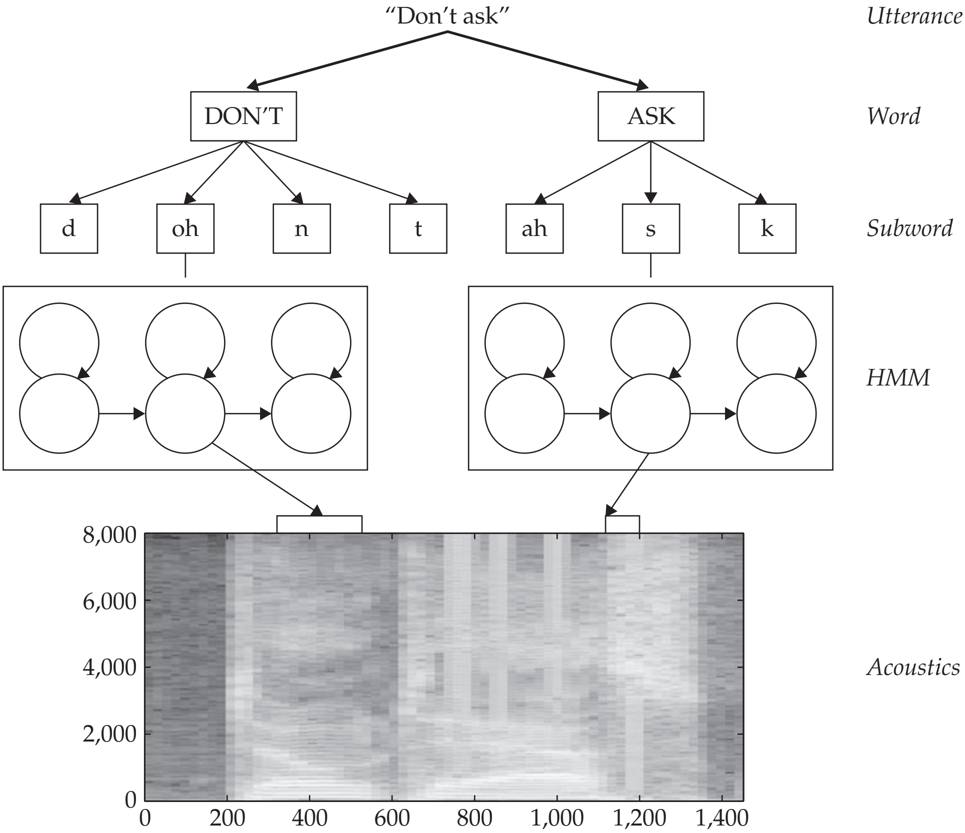


Figure 22.3 Hierarchical modeling in HMM speech recognition – constructing a word sequence model from basic phone models, and generation of speech acoustics.

sequences are made by concatenating word models. Thus a word sequence is represented by a large left-to-right HMM, illustrated in Figure 22.3. Note that the number of distinct HMM states in the system is determined by the size of the inventory of basic units, and is not dependent on the overall size of the vocabulary.

2.3 *Algorithms for HMMs*

Three basic algorithms are required to use HMMs for data-driven speech recognition.

1 *Alignment*: The HMM alignment problem is as follows. Given a hidden Markov model derived from a known sequence of words **W**, and a sequence of acoustic feature vectors **X**, what is the sequence of states (i.e., the path through the model) that is most likely to have generated the observed acoustics **X**?

2 *Decoding*: In the HMM decoding problem a sequence of acoustic feature vectors **X** is observed, but the word sequence is unknown. The problem is to decode the sequence of words **W*** most likely to have generated the observed speech acoustics.

3 *Training*: The HMM training problem is to find the parameters of the models – the transition probabilities and the parameters of the output distributions (e.g., Gaussian mean vectors and covariance matrices) – given a training set. A training set usually consists of a set of acoustic feature vectors and their corresponding orthographic transcriptions.

These three algorithms are closely related and take advantage of the HMM assumptions, in particular the Markov property whereby the state, q_t, at timestep t depends only on the state at the previous timestep, q_{t-1}.

To solve the alignment problem we need to find the HMM state sequence that generates the observed acoustics with the highest probability. Conceptually this involves enumerating all possible state sequences and obtaining the probability of each one generating the observed acoustics. Since the number of state sequences scales as the factorial of the number of timesteps, performing such calculations directly is infeasible for all but very short sequences. However, it is possible to obtain the most probable state sequence in an efficient manner, without making any approximation, using a version of dynamic programming referred to as the *Viterbi algorithm* (Forney Jr., 1973).

The Viterbi algorithm exploits the first-order Markov property in hidden Markov models. Consider two paths at state c at the same time: a, b, c and x, y, c. Owing to the Markov property, it is clear that if a, b, c has a higher probability than x, y, c, then any future paths with the prefix a, b, c will have a higher probability than paths with the prefix x, y, c. Therefore at each timestep only the most probable path need be kept at each state. This results in a huge saving in computation, and is illustrated in Figure 22.4.

Decoding is also based on the Viterbi algorithm. However, the problem is more complex than alignment since the word sequence is unknown: in the case of continuous speech recognition this means that each word may start at every

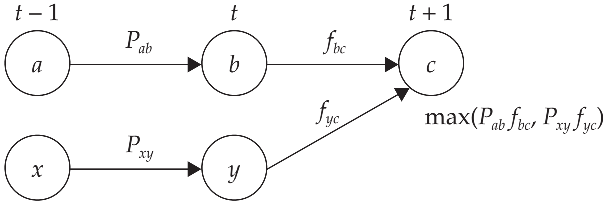


Figure 22.4 Illustration of dynamic programming: At state c the probabilities of paths a, b, c and x, y, c are compared and only the path with the highest probability need be retained.

timestep, leading to a large search problem that requires approximation, or *pruning,* for vocabularies of greater than a few hundred words. This is the search problem in speech recognition, and is discussed in more detail in section 7.

The ability to train HMM systems with millions of parameters from hundreds of hours of training data is the single most important factor in the success of HMMs for speech recognition. A vital aspect of HMM training is that a simple orthographic transcription is all that is required. Time alignments or phonetic transcriptions are not needed. In the training process an HMM is constructed for each word sequence, as described above. If the alignment of the speech acoustics with the state sequence was known, then training would be straightforward, since statistics could be collected at each state enabling the HMM parameters to be estimated (e.g., Gaussian mean vector estimated as the mean of the acoustic vectors aligned with that state). However, the speech–state alignment is not known. In this case an iterative algorithm known as the expectation maximization (EM) algorithm (Dempster et al., 1977) must be employed. This algorithm was originally developed by Baum (1972) in the context of HMMs, and is referred to as the Baum-Welch (or forward-backward) algorithm. In Baum-Welch training of HMMs a key quantity to be estimated is the posterior probability of an HMM being in a particular state at a particular time, given the sequence of observed acoustic vectors. These posterior probabilities, which may be computed recursively using two processes – structurally similar to the Viterbi algorithm – known as the forward and backward recursions, are then used to estimate a "soft" speech–state alignment. Each acoustic vector is probabilistically assigned to the HMM states according to these probabilities, and the HMM parameters are updated according to this soft alignment. Baum (1972) and Dempster et al. (1977) showed that this algorithm is guaranteed to increase the likelihood of the HMMs generating the observed data at each iteration. For this reason, HMM training in this manner is referred to as an example of *maximum likelihood* parameter estimation.

3 Acoustic Features

As discussed in Chapter 20, the time-amplitude waveform undergoes some form of signal processing, referred to as acoustic feature extraction, before modeling by an HMM. (Autoregressive hidden filter models are a theoretically interesting HMM variant that model speech at the waveform level (Poritz, 1988).) The main objective of feature extraction is to derive a representation which emphasizes those aspects of the speech signal that are relevant to ASR, and to remove those that are not important. The most widely used feature representations for speech recognition are mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) cepstral coefficients, both of which are discussed in Chapter 20. A frame rate of about 10 ms is typically used, resulting in a second of speech being characterized by a sequence of about 100 acoustic feature vectors. Each feature vector may also contain the first and second temporal derivatives of the extracted features, to provide some information about the local temporal dynamics of the speech signal (Furui, 1986), resulting in a typical feature vector size of 39 coefficients if the basic processing results in 12 cepstral coefficients plus delta energy. The preferred feature representations for ASR have been developed in conjunction with the development of HMMs using multivariate Gaussian or Gaussian mixture (section 4.1) output probability density functions. In particular, to make modeling using Gaussians easier, the features should not be correlated, and to save computation and reduce the number of model parameters that must be trained (from limited data), the features should be as few in number as possible.

The salient features are those that most directly relate to the *phoneme* being spoken: i.e., the *vocal tract shape.* Typically, the only salient source feature is the overall energy of the speech signal, although in tone languages the fundamental frequency also carries segmental information so may be used as a feature in systems designed for such languages. Section 6.3 mentions the use of suprasegmental information in ASR, but this is not typical of mainstream ASR systems.

4 Acoustic Modeling

Section 2 outlined the fundamentals of statistical speech recognition based on a hierarchical HMM structure. However, this basic approach alone does not result in an accurate speech recognition model. In practice, systems based on simple HMM phone models with Gaussian output distributions will result in high speech recognition error rates on all but the simplest tasks. Over the past two decades, several modeling approaches have been introduced that, combined, result in significantly more accurate speech recognition.

In this chapter we review advances in some important areas of acoustic modeling. First, we discuss how the output probability distributions attached to each state can be made significantly more flexible. Second, we outline how approaches such as context-dependent phone modeling can be used to expand the state space of the model and restrict the variability required to be modeled by each basic unit. Third, we examine alternative training approaches which aim to directly minimize the speech recognition error, by penalizing the incorrect models as well as improving the correct models during training. Fourth, we outline a number of techniques that enable a speaker-independent speech recognition system to be be tuned to the voice of a particular talker (or class of talkers). Finally, we discuss acoustic modeling approaches that are designed to deal with cluttered acoustic environments in which there is background noise and other competing acoustic signals.

4.1 Gaussian mixture models

Although Gaussian distributions are mathematically convenient and straightforward to estimate from data, they are constrained in the distributions that they are able to represent. In particular they are unable to represent distributions with more than a single mode, and have a limited ability to model distributions with heavy tails. A powerful way to address these limitations is via a technique known as a *mixture modeling.* A mixture model is a linear combination of a number of component density functions. For example if a k-component *Gaussian mixture model* (GMM) $p_{\text{mix}}(\mathbf{x} \mid q)$ is used to model the output density of a state q, then we have:

$$p_{\text{mix}}(\mathbf{x} \mid q) = \sum_{j=1}^{k} p_{\text{comp}}(\mathbf{x} \mid j, q) P(j \mid q) \tag{6}$$

where $p_{\text{comp}}(\mathbf{x} \mid j, q)$ are the k Gaussian components that make up the mixture model. In addition to the mean vectors and covariance matrices for each Gaussian component, a mixture model has another set of parameters, the mixture weights $P(j \mid q)$, which sum to one.

GMMs significantly relax the distributional assumptions imposed by Gaussians: indeed, given enough mixture components, any probability density function can be realized by a GMM. Additionally, GMMs may be trained using the EM algorithm (Dempster et al., 1977; Redner & Walker, 1984), in a manner analogous to HMMs: in this case the hidden variable is the mixture component rather than the state. Juang et al. (1986) showed that the EM algorithm may be extended in a straightforward manner to estimate the parameters for an HMM whose state output distributions are given by GMMs. This is now the core acoustic model used in most speech recognition systems. In a typical large vocabulary speech recognition system, each output distribution will have 16 or more mixture components, the exact number being chosen by a data-driven process (e.g., Young & Woodland, 1994). To reduce the number of parameters to be estimated it is usual for each Gaussian component to have a *diagonal* covariance matrix, in which the off-diagonal terms are zero. This greatly reduces the number of parameters to estimate for each Gaussian (from $(d^2 + 3d)/2$ to $2d$ for a d-dimension Gaussian).

4.2 Phone models

Hierarchical modeling depends on a set of basic speech units. In all but the smallest vocabulary speech recognition systems, such as digit recognizers, direct modeling of words is infeasible since the training data will not supply enough examples of all words in the vocabulary to reliably estimate whole word models. Thus speech recognition requires *subword models,* which may be used as the building blocks for word and phrase modeling. Although units such as syllables are well motivated linguistically, and reduce coarticulation effects across units, estimating models for such units from training data is difficult due to both data sparsity and the fact that syllable units display a high degree of spectral and temporal variability.

Most speech recognition systems are based on phone models. This reduces the problem of data sparsity since even moderate amounts of training data will provide plenty of examples for each phone model. Additionally phone models are well matched to pronunciation dictionaries, which are usually written by human experts. However, dependence on the surrounding phonetic context, as well as factors such as speaking rate, means that a phone model must cope with substantial variability. The most important way to improve the modeling power of a hidden Markov model system is to extend the state space. This may be loosely interpreted as meaning that each HMM state needs to model a smaller partition of the acoustic feature space. Rather than directly increasing the complexity of the individual phone HMMs, the usual way to extend the HMM state space is by using multiple *context-dependent* models for each phone, where each model is dependent on the surrounding phonetic context.

Context-dependent phone modeling is a divide and conquer approach, in which the number of different contexts is determined by the data – as the amount of training data increases, so more contexts may be used for each phone, and each individual context-dependent phone HMM is required to cover less of the acoustic space. The simplest context-dependent phone model is usually referred to as a triphone model (Schwartz et al., 1985; Lee, 1989). In basic triphone modeling there is a separate context-dependent phone HMM for each phonetic context of a particular phone, based on the left and right neighboring phones.[2] In general, context-dependent phone modeling includes contexts resulting from neighboring words. This approach provides much more detailed phonetic modeling by hugely expanding the state space, but, without modification, is difficult to estimate from data since the number of potential models is very large: for a system with 40 phones, this leads to $40^2 = 1{,}600$ triphone models per phone, or 64,000 models in total. Most of these will not be observed and hence cannot be estimated from the training data. One solution to this data sparsity problem is to "back off" to context-independent models if enough examples of a phone in a given context have not been observed.

Backed-off context-dependent modeling is an improvement over context-independent modeling, but fails to take advantage of the fact that similar acoustic contexts may be clustered to provide a more robust context-dependent modeling. This approach was pioneered by Lee (1989) who developed the approach of generalized triphones in which the total number of triphones was limited by clustering acoustic contexts. Young et al. (1994) developed an improved approach, in which automatically constructed decision trees were used to infer generalized phonetic contexts from the training data. Observed contexts in the training data are progressively refined and specialized so long as the training data is able to support a context. The approach relies on a predefined set of acoustic phonetic "questions" about the context. At each point in the decision tree, an information theoretic criterion is used to choose a question to partition the data, leading to more contexts being available as more training data is available. The idea of sharing model components based on the training data was further developed into a technique referred to as generalized tying (Young & Woodland, 1994). State-of-the-art speech recognition systems, trained on huge amounts of data, can use models with a large amount of phonetic context, such as the septaphone models developed by Chen et al. (2006). Alternative approaches to increasing the modeling power use triphone models with GMM distributions with a larger number of components (Hain et al., 2007) or a richer covariance structure (Olsen & Gopinath, 2004).

4.3 *Discrimination*

Generative modeling is based on the idea that the model parameters should be estimated such that the model reproduces the observed training data with the greatest probability: maximum likelihood estimation. However, this criterion is only indirectly related to the goal of speech recognition which is to reduce the word error rate – or, at a phonetic level, to classify each frame or segment with the correct phone label. Nadas (1983) showed that maximum likelihood training does indeed result in a minimum classification error if certain conditions are met. However, these conditions – which include a guarantee that the generative model is indeed correct, and that there is an infinite amount of training data – are never met in practice, and so maximum likelihood training of HMMs cannot be assumed to result in models that are optimal in terms of classification error. In this case a training criterion that is directly related to the posterior probability $P(\mathbf{W} \mid \mathbf{X})$ may be preferred, and such a criterion, conditional maximum likelihood, was proposed by Nadas.

In an influential paper, Bahl et al. (1986) extended this approach and proposed training HMM parameters using a related criterion, maximum mutual information (MMI). In MMI training, parameter values are chosen to maximize the mutual information between the acoustic observation sequence and the corresponding word sequence. This involves improving the correct models and penalizing incorrect models, or improving the models' *discrimination*. This approach did indeed significantly improve speech recognition accuracy, but Bahl et al.'s formulation of MMI training of HMMs was based on an optimization procedure known as gradient descent which is much less efficient than the Baum-Welch algorithm used for maximum likelihood training. This placed a considerable computational limitation on the problems to which discriminative training could be applied.

Normandin (1996) presented the extended Baum-Welch algorithm, a much more efficient way to train HMMs using MMI. However, extended Baum-Welch training is still expensive compared with maximum likelihood training, since parameter re-estimation in MMI training involves calculating the ratio of the probability of the HMM generating the correct transcription against the probability of it generating any word sequence. Obtaining the latter is computationally very expensive, since it involves a sum over all possible decodings. To apply this criterion to large vocabulary speech recognition required a number of approximations (Woodland & Povey, 2002). State-of-the-art large vocabulary speech recognition systems now employ these discriminative training techniques which offer a significant improvement in accuracy compared with maximum likelihood trained systems.

Other directly discriminative models have been proposed including connectionist approaches based on multilayer perceptrons (Bourlard & Wellekens, 1989; Morgan & Bourlard, 1995; Renals et al., 1994) and recurrent neural networks (Robinson, 1994; Robinson et al., 1996). These approaches directly estimate the posterior probabilities of an HMM state given the data $p(q \mid \mathbf{x})$ – and in the case of recurrent neural networks are able to incorporate a large amount of acoustic context in these probability estimates. These approaches do not define a generative model – there is no way to recover $p(\mathbf{x})$ – and training is expensive, based on gradient descent. In large vocabulary tasks with limited training data these models performed very well (Young et al., 1997), but they suffer limitations due to the expensive training (which is also more difficult to parallelize over multiple computers), and from weaker speaker adaptation algorithms.

More recently, there have been efforts to develop discriminative approaches to acoustic modeling, based on advances in machine learning, including approaches based on support vector machines (Venkataramani et al., 2007) and on conditional random fields (Gunawardana et al., 2005). These have been somewhat successful, but a major barrier has been scaling up these more complex discriminative methods to the large amounts of training data typically employed in large vocabulary speech recognition.

Related work on the development of *discriminative features* has been very successful. For example, feature sets derived from local posterior probability estimates (e.g., frame-wise phone posteriors) (Morgan et al., 2005), or derived from the set of Gaussians used in the acoustic model (Povey et al., 2005), increased the accuracy in large vocabulary speech recognition, in combination with conventional acoustic features.

4.4 Adaptation

Speaker-dependent ASR systems, trained to recognize the speech of a specific known speaker, can have significantly lower word error rates compared with speaker-independent ASR systems. *Speaker adaptation* is the process by which a speaker-independent system is tuned to the speech of a target speaker, and usually refers to processes that adjust the parameters of the acoustic model, or transform the acoustic features. *Speaker normalization* approaches transform the acoustic features of the target speaker to be closer to those of an "average" speaker. *Model-based* approaches transform the parameters of the set of speaker-independent acoustic models to better model the speech of the target speaker. *Speaker space* approaches estimate multiple sets of acoustic models, characterizing a target speaker in terms of this model set. Speaker adaptation is well reviewed by Woodland (2001).

Speaker adaptation may operate in either a *supervised* or *unsupervised* fashion. In supervised speaker adaptation the system is adapted to the voice of the target speaker using data for which the correct transcript is available – in the case of commercial dictation systems this is achieved by the user reading a predefined adaptation script. In the case of unsupervised speaker adaptation, the correct orthographic transcription is not available, and thus it is not possible to construct an HMM for each utterance which may be used to perform operations based on HMM alignment or training. In this case, the (erroneous) output of the unadapted recognizer may be used, possibly augmented with confidence measures.

Speaker normalization involves inferring a transform for the acoustic feature vectors to make the target speaker appear closer to an average speaker. The length of the vocal tract has a substantial effect on the spectrum – for a female adult the observed formant frequencies are, on average, 20 percent higher than those for a male adult. Cohen et al. (1995) reported that a linear warping of the frequency axis could compensate for differences in vocal tract length between speakers. Vocal tract length normalization (VTLN) estimates a frequency warping for a target speaker and has proven to be a very effective speaker adaptation technique. In cases where there are several tens of seconds of data for a target speaker, for example in conversational telephone speech, VTLN has consistently resulted in significantly reduced word error rates (Lee & Rose, 1996; Wegmann et al., 1996; Hain et al., 1999; Welling et al., 2002). Two main approaches have been employed to compute the VTLN frequency warping parameter. One approach explicitly estimates formant positions (Eide & Gish, 1996; Wegmann et al., 1996); a second treats the warping factor similarly to the other parameters in the acoustic model and optimizes it so as to maximize the likelihood of the HMMs generating the observed data (Lee & Rose, 1996; Hain et al., 1999; Welling et al., 2002). In the maximum likelihood case, rather than optimizing the models to better fit the data, we can view the process as warping the observed features to better fit the models. After estimating the warp factors the acoustic models are retrained, and the process may be iterated. VTLN does not directly estimate vocal tract size, and it is the case that the optimal value of the warping factor is influenced by other factors, such as the fundamental frequency. Faria and Gelbart (2005) present a method which exploits the relation to the fundamental to enable VTLN from very small amounts of adaptation data. Irino and Patterson (2002) present a signal processing technique that directly segregates vocal tract size from the acoustic signal.

Model-based approaches address a similar problem to acoustic model training: how to adapt the HMM parameters to fit the observed data from the target speaker. The major contrast with HMM training is that the amount of adaptation data is typically orders of magnitude less than the amount of training data. There are two principal approaches to model-based speaker adaptation, maximum a-posteriori (MAP) adaptation and the linear transformation of parameters. Both use the adaptation data to adjust the speaker-independent acoustic model parameters to better fit the target speaker.

MAP adaptation uses the speaker-independent acoustic models to construct a prior probability distribution over the model parameters, an approach first suggested by Brown et al. (1983) and presented in detail by Gauvain and Lee (1994). The advantage of MAP adaptation is that it is a theoretically well-motivated approach to incorporating the prior knowledge that is inherent in the parameters of the speaker-independent system. Furthermore, the update equations for the HMM parameters have a clear relation to those used for maximum likelihood HMM training. When the amount of training data is small, then the speaker-independent parameters dominate; as the amount of adaptation data increases, so the parameters begin to more strongly reflect the target speaker. The main drawback to the MAP approach is that it is local. Only the parameters belonging to models observed in the adaptation data will be updated. In large vocabulary systems, which may have tens of thousands of component Gaussians, this means that most parameters will not be adapted even in the case of relatively large amounts of adaptation data.

A set of speaker adaptation algorithms have been developed to address the local nature of MAP, building on the intuition that adaptations in acoustic model parameters for a particular speaker are systematic across phone models. Observing how the acoustics generated by a particular speaker are transformed for one phone model tells us something about how they may be transformed for other phone models. In particular, a family of speaker adaptation techniques has been developed in which the HMM parameters undergo a linear transformation (Cox & Bridle, 1989; Digalakis et al., 1995; Leggetter & Woodland, 1995; Gales & Woodland, 1996; Gales, 1998;) – maximum likelihood linear regression (MLLR). In MLLR, speaker adaptation consists of estimating the parameters for these linear transforms. In order to use a relatively small amount of training data, the same transformation may be shared across multiple phone models. Leggetter and Woodland (1995), shared a transform across the set of context-dependent models corresponding to the same (context-independent) phone – around 40 separate adaptation transforms in total. However, more recent research in large vocabulary continuous speech recognition has indicated that better performance can be obtained using fewer transforms, even just a single linear adaptation transform for all speech models corresponding to the target speaker (e.g., Chen et al., 2006). It should be noted that such linear transforms can account for any systematic (linear) variation from the speaker-independent models, for example those caused by channel effects.

A third approach to speaker adaptation involves characterizing a speaker in terms of a "speaker space." In these approaches, rather than constructing a single speaker-independent acoustic model, a set of "canonical" models is constructed, based on a speaker clustering. The process of speaker adaptation involves computing the appropriate interpolation between canonical models to best represent the target speaker. One way of viewing this is by regarding each canonical model as forming a dimension of a speaker space. In an approach called *eigenvoices* (Kuhn et al., 2000), principal component analysis is used to construct such a space from the parameters of a set of speaker-dependent HMMs. However, this technique is computationally intensive and does not scale well to large vocabulary systems. A related technique, *cluster adaptive training* (CAT; Gales, 2000), explicitly clusters speakers resulting in a set of speaker cluster acoustic models. Again, adaptation involves interpolating between the speaker cluster models, but in this case since a vector space is not explicitly constructed, application to large-scale problems is still possible.

Speaker adaptation has been one of the most intensively researched areas of acoustic modeling since the mid 1990s. Progress has been substantial, particularly for systems where over a minute of adaptation material is available, and many variants of the methods described above have been proposed, including combinations of MAP and linear transformation approaches, e.g., structured MAP linear regression (SMAPLR; Siohan et al., 2002). Some of the techniques are complementary and give additive improvements in performance, for example a feature space transform such as VTLN, followed by a linear model parameter transform from the MLLR family. Additionally there are close mathematical relationships between all three categories of speaker adaptation (e.g., Gales, 1998).

4.5 Robust speech recognition

The acoustic modeling approaches discussed above are typically applied in situations where background noise, reverberation, or competing talkers are not present. These systems typically fail when forced to operate in more challenging acoustic environments – environments which are characteristic of everyday life, and which do not present problems to human speech recognition. *Robust* speech recognition has become a major research area. A large proportion of the work in this area has been concerned with the development of robust acoustic feature representations (e.g., Hermansky & Morgan, 1994; Liu et al., 1994).

Model-based approaches use the acoustic model to extract the parts of the speech signal that correspond to the target speaker. Varga and Moore (1990) introduced a technique called *HMM decomposition* in which parallel HMMs were constructed to account for both speech and noise. During recognition an inference is performed so that the speech and noise models cooperatively account for the observed signal: the recognized speech is then based on the state sequence of the speech models alone. This techniques was extended by Gales and Young (1996), and termed *parallel model combination,* and was used to separate and recognize the speech of two overlapping talkers by Kristjansson et al. (2006) (although these experiments made use of substantial language model restrictions, and acoustic information about the competing speaker). The main disadvantage of these approaches is that they can be computationally costly, and they have not yet been applied to large problems.

An alternative approach, taking inspiration from computational auditory scene analysis (Cooke & Ellis, 2001), makes the assumption that each location in the time-frequency map of an acoustic signal resulting from two acoustic sources (speech + speech, or speech + noise) is dominated by one of the sources. *Missing feature theory* attempts to identify those regions of reliable data for the target speaker, and to perform recognition based on this partial information (Cooke et al., 2001; Raj & Stern, 2005).

5 Pronunciation Modeling

A pronunciation model is a key component of the hierarchical approach to speech recognition outlined in section 2.2. The role of the pronunciation model is to specify how words are composed from subword units, typically context-dependent phone models (section 4.2). By using a pronunciation model to map from a sequence of words to an utterance-level HMM, formed by concatenating subword models, it is possible to train a speech recognition system from an orthographic transcription, rather than a phonetic transcription.

In English (the most widely used language in ASR research), the pronunciation model is nothing more than a listing of all words along with their pronunciation in terms of a sequence of subword units: a pronunciation dictionary. This dictionary is written by human experts. The subword unit used in a typical ASR dictionary is essentially the phoneme, but the set of phonemes is considerably smaller than the IPA: it is specific to one language and is sufficient only to distinguish all the words in the dictionary (apart from genuine homophones). Within-word phonological phenomena, such as assimilation, are represented directly (e.g., "hand" + "bag" becomes "hambag").

It is common for some words to have more than one possible pronunciation, as described in terms of the subword units. In these cases, the dictionary can contain multiple entries for those words. It has been found that including a large number of pronunciation variants in the dictionary can lead to worse performance (Hain, 2002). This is because an increase in the number of pronunciation variants causes the system to become more flexible, with many possible state sequences able to match the observed acoustics. In typical large vocabulary ASR systems, the great majority of words will therefore have only a single pronunciation listed in the dictionary, with an overall average of about 1.1 pronunciations per word.

Describing speech, especially spontaneous speech, as "beads-on-a-string" (Ostendorf, 1999), where the beads are phonemes, must be seen as a consistent, rather than faithful, representation. Consistency – transcribing the same word always with the same phonemic representation – has the advantage that training speech material only needs a word transcription, which can be automatically converted to a phonemic transcription in order to train HMMs of subword units. A faithful phonetic transcription of speech is far more expensive to obtain, and would be impractical for the large speech corpora in use today.

Any variation present in the data that is not modeled in the dictionary must instead be absorbed by the acoustic model. HMMs, the GMM probability density functions they use, and the many sophisticated training algorithms available, have proven to be better mechanisms for learning this variation than using larger inventories of subword units, or listing more variants in the dictionary (Hain, 2002). Effects present in connected speech (e.g., assimilation across word boundaries) cannot be represented in the dictionary and must also be left to the acoustic model.

Although the subword unit inventory used in the dictionary is quite small, the acoustic models of these units will usually be context dependent (e.g., triphones) and therefore much finer grained. As discussed in section 4.2, automatic procedures are used to share models in similar acoustic contexts, so the effective number of subword unit types is actually learned from the data.

6 Language Modeling

In equation (4), the probability of the word sequence, $P(\mathbf{W})$, is calculated by the *language model*. The probability $P(\mathbf{W})$ is called a *prior* because it can be calculated

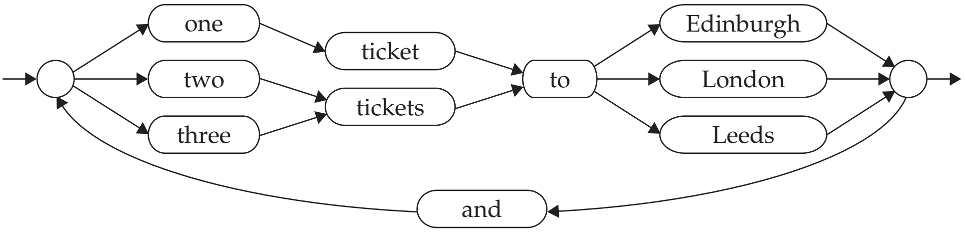


Figure 22.5 A finite state language model for a limited domain: booking a train ticket. Note the special start and end of sentence states.

before the speech signal **X** is observed – it is a model of prior beliefs about how probable various possible word sequences are. The language model will help disambiguate acoustically similar words. For example, the two phrases "the nudist play" and "the new display" may have similar pronunciations, leading to identical utterance-level HMMs. In this case, the acoustic model in (4) will return equal values for the likelihoods $P(\mathbf{X}\,|\,\text{“the nudist play”})$ and $P(\mathbf{X}\,|\,\text{“the new display”})$. Prior knowledge must be used to decide which is the most likely: e.g., the language model may estimate that "the new display" is a more likely phrase. The final result of recognition depends on both the acoustic evidence, and on prior expectations, just as in human speech recognition.

The language model describes how probable any given word sequence is, and this is an essential component of most speech recognizers. The more accurately the language model can predict the word sequence, the higher the expected accuracy of the whole system will be. Without a language model, speech recognition accuracy is generally very poor.

In limited application domains, a simple finite state network can be written by hand to describe all allowable sentences. Figure 22.5 shows an example network which can generate (or accept) sentences about booking a train ticket. It is not difficult to think of a sentence that this network cannot handle (e.g., "one Edinburgh ticket"). However, if the input speech conforms precisely to this word network, then we can expect high accuracy ASR, because the network provides very strong constraints. The art in crafting such a network by hand, is to allow all the sentences necessary for the application, whilst allowing as few other unnecessary sentences as possible. Such an approach can work very well in limited domains but is not useful for applications such as dictation, or transcribing a business meeting. Such applications require robust language models, with wide coverage.

Traditional grammars consist of sets of rules, describing the way valid sentences can be built from constituents such as noun phrases, verbs, clauses, and so on. These are typically written by hand. Creating a rich enough set of rules to cover a wide variety of natural sentences is very time consuming. Hand-crafted grammars do not assign probabilities, they simply accept or reject sentences. To be usable in ASR, a grammar must be probabilistic, have wide coverage, and be compatible with left-to-right search algorithms (see section 7). Although modern statistical parsers, which use wide coverage grammars learned from large amounts of text data, meet the first two requirements, it is only recently that the third requirement has been satisfied, and these powerful models of language have been successfully applied to ASR (Chelba & Jelinek, 2000).

6.1 **n**-*gram models*

Although wide-coverage probabilistic grammars are available, they do not represent the state of the art in large vocabulary ASR. *Statistical language modeling* for ASR is based on *n*-gram models, a linguistically implausible model with a short finite context (typically the previous two words). Such a language model may seem far less sophisticated than even the simplest theory of syntax, but there are good reasons for this. Statistical language models must be able to cover "ungrammatical" spoken utterances, they must be easy to learn from huge amounts of data, they must be computationally efficient to use, and they must be able to assign a probability to a fragment of an utterance (not just a whole sentence).

The *n*-gram model meets all these requirements. This model estimates $P(\mathbf{W})$ by using an approximation: that the probability of a word depends only on the identity of that word and of the preceding $n - 1$ words. These short sequences of n words are called "*n*-grams." If $n = 2$, the model is known as a *bigram,* and it estimates $P(\mathbf{W})$ as follows:

$$\mathbf{W} = \{W_1, W_2, \ldots W_M\}$$

$$P(\mathbf{W}) = P(W_1)\, P(W_2 \mid W_1)\, P(W_3 \mid W_1, W_2) \ldots P(W_M \mid W_{1,} W_2, \ldots W_{M-1}) \quad (7)$$

$$\approx P(W_1)\, P(W_2 \mid W_1)\, P(W_3 \mid W_2) \ldots P(W_M \mid W_{M-1}) \quad (8)$$

Equation (7) is exact, but is not usually practical. Equation (8) is an approximation using bigrams. The approximation is better (closer to (7)) for larger *n*. *n*-grams with $n = 3$ are called trigrams. They are the most common language model used currently in ASR, and have been for some time, despite many efforts to find better models (Jelinek, 1991).

The parameters of an *n*-gram model are the conditional probabilities in (8), for all possible sequences of *n* words constructed from the vocabulary. These may be estimated from some training data (which need only be text, not speech). The probability $P(W_B \mid W_A)$ can be estimated from data as the number of times that the sequence W_AW_B occurs, divided by the number of times that W_A occurs:

$$P(W_B \mid W_A) = \frac{C(W_A, W_B)}{C(W_A)} \quad (9)$$

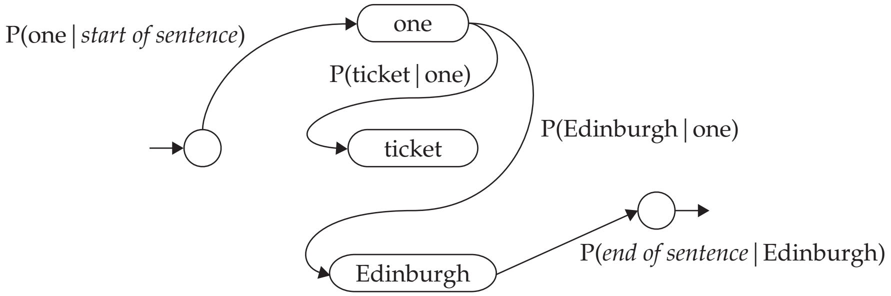

P(one | *start of sentence*)
one
P(ticket | one)
ticket
P(Edinburgh | one)
Edinburgh
P(*end of sentence* | Edinburgh)

Figure 22.6 A bigram can be written as a weighted finite state network in which the arcs have probabilities associated with them. Note the special start and end of sentence probabilities. Only some of the arcs are shown, for clarity: the full model has an arc from every word to every other word, including itself.

where $C(\cdot)$ is a count. When word W_A is used in a sentence, it must be followed by some other word in the vocabulary. This means that the total probability, summed across all the words that can follow W_A must be 1:

$$\sum_{W_B} P(W_B \mid W_A) = 1 \tag{10}$$

where the sum is over all the words in the vocabulary. The counting method in (9) ensures that (10) is satisfied.

An important property of *n*-gram models is that they can be written as finite state networks, as in Figure 22.6. Why this is important will become clear when we consider search in section 7. Although the language model in Figure 22.5 has some advantages over the bigram in Figure 22.6 (e.g., it does not allow ungrammatical sentences such as "three ticket Edinburgh"), the bigram language model in Figure 22.6 is more flexible: it allows all possible sentences (including "one Edinburgh ticket"), but each is weighted by a probability. This is often more appropriate for speech recognition, because people say unexpected or ungrammatical things!

If we use the *n*-gram language model to calculate the total likelihood of the training data, the estimate of $P(W_B \mid W_A)$ given by (9) is the value that maximizes this likelihood. The estimate of $P(W_B \mid W_A)$ in (9) is therefore called the maximum likelihood estimate. However, the language model will not usually be applied to the training data – it will be used on new, unseen data. It must therefore *generalize*, and maximum likelihood estimation may not lead to the best generalization performance. If only a few examples of W_AW_B occur, then the maximum likelihood estimate may be unreliable. In particular, if there is no example of W_AW_B in the training data, then (9) will assign a probability of zero to $P(W_B \mid W_A)$.

To better capture the patterns of natural language requires *n* to be as large as possible. But, as *n* gets larger, the number of parameters to estimate grows very quickly and the simple counting method fails, because there will be very many *n*-word sequences that do not occur in the data. Yet, even for large *n*, there will still be *some* commonly occurring sequences of *n* words. The challenge for language modeling is to estimate accurate probabilities for frequent word sequences *and* for rare or unseen ones. The various solutions to this problem are known as "smoothing" because they "even out" the probabilities in the model: in particular, *n*-grams that did not happen to occur in the training data will not simply be assigned a zero probability.

6.2 Smoothing

Estimating the probability of an *n*-gram by counting occurrences in the training data will fail for *n*-grams with zero counts, and will give unreliable estimates for *n*-grams seen only a small number of times. One trivial way to avoid zero counts is to add 1 to all the $C(W_A, W_B)$ counts in (9) and then renormalize so that (10) is still satisfied. This so-called "add-1 smoothing" makes the probability distributions more uniform, because the high probabilities are slightly reduced and the low or zero probabilities are increased. Another way to view add-1 smoothing is therefore as a *discounting* of the higher counts, and a distribution of that "spare" probability mass to the smaller counts. Add-1 smoothing does not work very well in practice: the amount added to each count is the same and is chosen arbitrarily, and it has the effect of allocating too much probability mass to unseen *n*-grams (Manning & Schutze, 1999). However, redistributing the probability mass obtained by discounting is an important concept which is used in more sophisticated smoothing schemes for *n*-gram models. Chen and Goodman (1996) offer a comparison of several of the many smoothing techniques that have been proposed.

The most common approach to *n*-gram smoothing estimates the probabilities for *n*-grams that are unseen (or rarely seen) in the training data using a simpler model: a lower-order *n*-gram. Several orders of *n*-gram model can be combined, either by interpolating their probability estimates (Jelinek & Mercer, 1980) or through a technique known as *back off* (Katz, 1987). In a backed-off trigram language model, if the trigram probability $P(W_C \mid W_A, W_B)$ cannot be reliably estimated, the bigram probability $P(W_C \mid W_B)$ is used instead, or even just the unigram probability $P(W_C)$. A back-off model gradually drops the amount of left context, trying simpler and simpler models until the probability of W_C can be reliably estimated. Backed-off *n*-gram models must use discounting, and are re-normalized, so that (10) is still satisfied:

$$\begin{aligned}
&\text{if the bigram probability estimated by counting is zero: } P(W_B \mid W_A) = 0 \\
&\qquad\qquad\text{then use the unigram instead } P(W_B \mid W_A) \approx P(W_B) \\
&\qquad\qquad\text{but unfortunately that means } \sum_W P(W \mid W_A) = 1 + P(W_B)
\end{aligned}$$

This problem must be rectified by discounting the higher bigram counts, in order to "free up" some probability mass to redistribute to the low/zero frequency bigrams (whose probabilities will be computed using unigrams):

$$P(W_B \mid W_A) = \frac{C(W_A, W_B) - D}{C(W_A)} \quad \text{if} \quad C(W_A, W_B) > c \tag{11}$$

$$= P(W_B)b_{W_A} \qquad \text{otherwise} \tag{12}$$

where c is the minimum count required for the bigram probability estimate to be considered reliable (e.g., $c = 2$), D is a fixed discount (e.g., $D = 0.1$), and b_{W_A} is the back-off weight which is set so that (10) is satisfied. This form of discounting is still rather ad hoc: the discount D and threshold c must be chosen rather than learned from data, and they are the same for all n-grams.

There are more sophisticated ways to redistribute the probability mass than using a fixed discount D. In the training data, a very large number of n-gram types will never be seen: they have a count of zero. Some smaller number of types will have a frequency of 1; an even smaller number will be seen twice, and so on. *Good-Turing discounting* (Nadas, 1985) makes the assumption that this distribution will be smooth. This is reasonable: we would expect the number of n-gram types occurring exactly four times to be larger than the number that occur exactly five times, etc. By smoothing the original raw counts using a simple function, *corrected counts* are obtained. The smoothing function is designed so that the corrected count for n-gram types which had a frequency of zero in the training data will be a little greater than zero, and the corrected counts for all other types will be a little less than their original values: probability mass has been redistributed, and the model has been smoothed. The corrected counts could be used directly in (9), but are more commonly used to calculate the discount amounts for a backed-off n-gram model. *Kneser-Ney smoothing* (Kneser & Ney, 1995) and its interpolated and modified variants (Goodman, 2001) are probably the best-performing smoothing methods currently available for n-grams used in ASR; these methods are also based on prior assumptions about the distributions of n-gram types.

6.3 *Language models specifically designed for* **speech**

All the probabilistic language models discussed so far are equally applicable to text or speech. In fact, they are invariably trained on written language corpora (usually from newspapers or web pages) simply because there is not enough transcribed spoken language to train these models. As a consequence, the models cannot specifically deal with phenomena that only occur in speech and never in text, such as disfluency, and neither can they make use of additional information available in speech, such as prosody. Some initial attempts have been made to deal with disfluency (e.g., Stolcke & Shriberg, 1996) and to use prosodic information to reduce word error rate (e.g., Stolcke et al., 1999). These methods have produced only small improvements in word error rates, so are not widely used in speech recognition systems. The relationship between prosodic features and word recognition is governed by the *meaning* of the utterance, which limits the ability of the fairly shallow models typically used in ASR to use prosody to aid word recognition. Prosody can also be used to assist recognition of other properties of utterances – so-called "rich transcription" (see section 9).

7 Search

Once the components of a typical speech recognizer based on HMMs of phones, a pronunciation dictionary, and an *n*-gram language model have been assembled and trained, all that remains is to consider exactly how equation (4) is used to find the most likely word sequence, **W***. This is the job of the *decoder.* As discussed in section 2.3, the Markov, or finite state, nature of HMMs and of *n*-gram language models can be exploited, and efficient algorithms exist to find the most probable word sequence, **W***.

The key insight behind all search algorithms for ASR is that computation can be shared: when evaluating $P(\mathbf{X}\,|\,\mathbf{W})P(\mathbf{W})$ for two different word sequences, it is possible to share some of the computation if the word sequences share some subsequences. This sharing of computation is possible because both the HMM and the *n*-gram only use local computations: the probabilities that they calculate only require knowledge of very local context, and not the entire sentence. The HMM is an extreme case: to compute the probability of generating a particular observation requires only knowing what state of the model is generating that observation, and nothing else. To reduce the computation required to evaluate $P(\mathbf{X}\,|\,\mathbf{W})$, it is usual only to consider the single most likely state sequence, which can be very efficiently found – by sharing computation – using the Viterbi algorithm. In an *n*-gram model, only the current and preceding $n - 1$ words need be specified; any history further back than that (or indeed the future) has no effect. Because the language model uses an amount of context dependent on *n*, increasing *n* will have a significant impact on the computational cost of the search for **W***. This effectively limits the order of language model that can be used in ASR, although there are techniques to work around this problem.

To illustrate this sharing of computation, consider a simple practical implementation of search for ASR, illustrated in Figure 22.7. The figure shows the process of "compiling" the language model with the dictionary and the acoustic model, by first replacing every word instance with its phoneme sequence and then replacing each phoneme instance with the corresponding HMM. The result is a network of HMM states. **W*** corresponds to the path through this network that generates the observation sequence with the highest likelihood, and it is the task of the search algorithm to find this path.

In the full compiled network, consider these two possible paths:

one ticket to	Edinburgh
one ticket to	Leeds

Now trace these two paths on Figure 22.5, doing so *time synchronously* – that means that the paths are explored in parallel, working through the observation

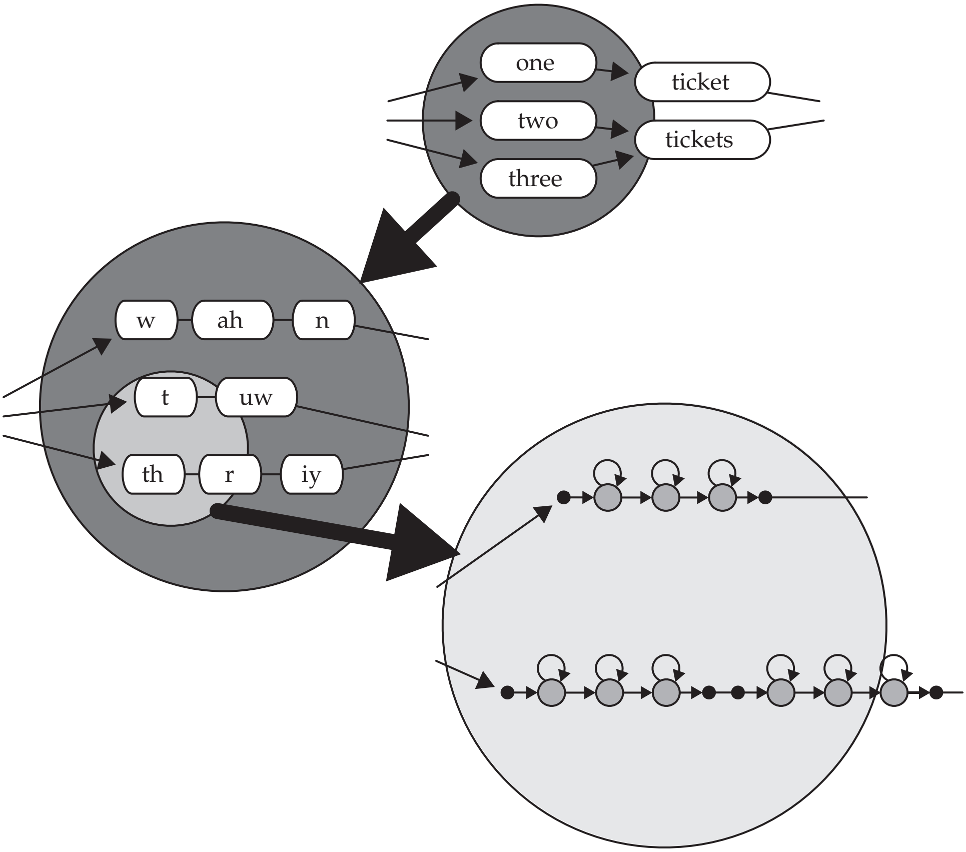


Figure 22.7 The process of compiling a recognition network. At the top of the figure is a fragment of the language model from Figure 22.5. Then, each word is replaced by its phonetic pronunciation (provided by the dictionary) and finally each phoneme is replaced by the appropriate HMM. The result is a network of HMM states, which is therefore also an HMM, so all HMM algorithms can be applied to it, such as Viterbi search.

sequence one observation at a time. Clearly, computation can be shared for the first three words. But even more savings in computation can be made, illustrated by this example:

two	tickets . . .
three	tickets . . .

where the start time of the word "tickets" is the same in each case. Because of the finite state property of the compiled network, all the computation involved in continuing the first sentence (i.e., searching all possible word sequences) can be shared by the second sentence. The probability of each possible continuation W' will be the same in the two cases because $P(W' \mid \text{two tickets}) = P(W' \mid \text{three tickets}) = P(W' \mid \text{tickets})$ for all possible W'. If the network represented an n-gram language model of order N, the paths would need to have the last $N - 1$ words in common in order for them to be merged.

The simple search algorithm described above is exact: it will find the same word sequence as the naive exhaustive search which considers each possible sentence one by one. However, it will be still too slow for practical use in medium and large vocabulary ASR systems.

In order to speed things up further, an approximation has to be made: to discard some parts of the search space before they have been fully explored. This is called *pruning.* As a path through the network is extended, its cumulative probability so far is known. If this probability falls too far below that of the current most likely path, then we can guess that the less likely path is unlikely to turn out to be the eventual "winner" and can be abandoned immediately. This is called *beam pruning.* The computational cost of this algorithm can be lowered (e.g., to make the system run in "real time") by reducing the *beam* width: that is, the maximum allowed difference between a path's likelihood and that of the current best path. However, the narrower the beam, the more chance of accidentally pruning the path that would have been the eventual winner. This will generally decrease the accuracy of the system, by introducing *search errors.*

For large vocabularies and long span language models, explicitly building the network in Figure 22.7 is no longer practical because it may not fit in the computer's memory. Additionally, when using beam pruning, for any given utterance to be recognized, only a tiny part of the network will be explored. Therefore, large vocabulary decoders use more complex implementations which, for example, do not explicitly compile this network (e.g., Ney & Ortmanns, 2000).

Since it is the language model that has the largest effect on the search space and the memory and computational requirements, another strategy for reducing computation cost (and memory requirements) is to use a *rescoring* technique. The concept is the same as pruning: to quickly discard unlikely parts of the search space. This can be done by first performing decoding using a simple language model (e.g., a bigram) and only retaining the likely parts of the search space (i.e., a set of sentences which hopefully contains the correct sentence). This reduced search space can be represented as a list of sentences (an *N-best list*) or, more compactly, as a *lattice.* The sentences in this list or lattice are rescored using a more complex language model, to arrive at the final most likely word sequence.

8 Evaluation

Speech transcription is straightforward to evaluate since it is usually possible to agree on a reference orthographic transcription. The output of a speech recognizer may then be *aligned* to the reference transcription. The quality of the speech recognizer output may then be measured in terms of errors relative to the reference transcription. Comparisons of this type are referred to as the *string edit distance* in computer science. There are three possible types of error: *substitutions, insertions*, and *deletions.* For example if the utterance *handbook of phonetic sciences* was transcribed as *handbook frenetic sign says*, this would usually be measured as four errors in total: two substitutions (*phonetic/frenetic; sciences/sign*), 1 deletion (*of*), and 1 insertion (*says*). Although this is probably the most "natural" alignment, an alternative alignment (with three errors, all substitutions) is also possible. The alignment is carried out using a dynamic programming algorithm, and the transition costs implicit in such an algorithm enable a preferred alignment to be found. It is also possible to incorporate phonetic, pronunciation, and timing information in the alignment process (Fisher & Fiscus, 1993). In practice, researchers use alignment procedures with agreed transition costs, and with freely available implementations such as the US National Institute of Standards and Technology (NIST) sclite software,[3] or the HResults tool in the HTK speech recognition toolkit.[4]

Although it is possible to measure the accuracy of a speech recognizer using the percent of words recognized correctly, this measure does not take account of insertion errors. The most usually employed metric is the word error rate (WER), which is obtained by summing the three different types of error. If there are N words in the reference transcript, and alignment with the speech recognition output results in S substitutions, D deletions, and I insertions, the word error rate is defined as:

$$\text{WER} = 100 \cdot \frac{(S + D + I)}{N}\% \tag{13}$$

$$\text{Accuracy} = (100 - \text{WER})\% \tag{14}$$

In the case of a high number of insertions, it is possible for the WER to be above 100 percent.

In the late 1980s the TIMIT corpus (Fisher et al., 1986) was collected and made widely available to speech researchers. This is a significant point in the history of speech recognition research: TIMIT was a carefully recorded and phonetically transcribed corpus including over 600 North American speakers. Importantly, the data was divided into *training* and *test* sets. Nearly all research on the TIMIT corpus has used this split of data, therefore enabling precise comparisons between different approaches, since models developed by different researchers are trained on precisely the same data, and tested using the same test data. Since then, the development of speech corpora for speech recognition has had a very close relation to evaluation: many corpora have now been released with corresponding evaluation protocols. Since the late 1980s there have been regular benchmark evaluations of ASR systems, in domains such as conversational telephone speech, broadcast news, and multiparty meetings, using standardized data and evaluation protocols. This cycle of evaluation, primarily led by NIST, has given an objective basis to speech recognition research and resulted in consistent improvements in accuracy. On the other hand, it has been argued that those techniques that have been demonstrated to incrementally reduce word error rate are quickly adopted across the research community at the expense of reduced innovation and adventure in research (Bourlard et al., 1996).

9 Discussion

Automatic speech recognition involves transcribing the acoustic signal as a sequence of words: it takes no account of the meaning of an utterance, of the fact that speech is usually part of a conversation. Rather than trying to perform a complete semantic interpretation of spoken utterances, there has been a recent emphasis on *rich transcription* or the automatic extraction of semantic content from speech (Koumpis & Renals, 2005), including the automatic identification of entities such as names (Makhoul et al., 2000), development of more readable transcriptions through automatic sentence segmentation and capitalization (Kim & Woodland, 2003), and automatic speech summarization (Hori & Furui, 2003). Spoken dialogue research – which includes speech generation and dialogue planning – is an active research area, and many domain-specific systems have been constructed (McTear, 2002), with a trend towards the development of systems based on statistical machine learning (Levin et al., 2000).

In this chapter we have reviewed the techniques, models, and algorithms that form the state-of-the-art of ASR. The approaches which dominate the field incorporate only very shallow linguistic or phonetic knowledge. The extent of knowledge derived from human speech processing in current systems includes the nonlinear scaling of the frequency axis, and the use of phonemes as basic units of speech. Many researchers view speech recognition as an engineering challenge, in which advances will come through further investigation of statistical machine learning and digital signal processing. However, the performance gap between human and automatic speech recognition is considerable, especially when one considers the flexibility and robustness of human speech recognition (Lippmann, 1997). Allen (1994) and Hermansky (1998) have argued that current speech recognizers have an underlying acoustic model that is an order of magnitude less accurate than humans (in terms of phoneme recognition accuracy) and that the investigation of human speech recognition is an important area for for automatic speech recognition research. Considering human and automatic speech recognition jointly has become an area of some interest, and is reviewed by Scharenborg (2007); however, compelling automatic speech recognition results using models inspired by human speech recognition have not yet been obtained.

10 Further Reading

This chapter provides an overview of statistical, data-driven speech recognition, outlining the main concepts, and giving a flavor of some of the necessary details that are essential to develop a state-of-the-art system. Gold and Morgan (2000) give an excellent historical perspective on ASR and, together with Huang et al. (2001) and Jurafsky and Martin (2008), provide a modern and in-depth treatment of the area. Jelinek (1998) gives a treatment focused on hidden Markov acoustic models and *n*-gram language models.

Rabiner (1989) is a classic tutorial about hidden Markov models, with a treatment based on that of Ferguson (1980), which is rather difficult to obtain. Alternative treatments of hidden Markov models, and their associated training algorithms, are provided by Durbin et al. (1998) and Neal and Hinton (1998).

Gauvain and Lamel (2000) provide a review of large vocabulary continuous speech recognition, and Hain et al. (2005) and Chen et al. (2006) are examples of current, state-of-the-art systems.

Several freely available (for noncommercial use) open-source software packages are available, in particular HTK, the HMM toolkit, developed at Cambridge University (http://htk.eng.cam.ac.uk), and SRILM, the SRI language modelling toolkit (www.speech.sri.com/projects/srilm).

NOTES

[1] Strictly speaking we should write the acoustic and language models as conditioned on the form of the generative model: $P(\mathbf{X}\,|\,\mathbf{W}, \mathcal{M})$ and $P(\mathbf{W}\,|\,\mathcal{M})$.

[2] "Triphone," although standard terminology, is a misnomer since a triphone models a single phone (and not three phones) in a given left and right context.

[3] www.nist.gov/speech/tools.

[4] http://htk.eng.cam.ac.uk.

REFERENCES

Allen, J. B. (1994) How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2, 567–77.

Bahl, L., Brown, P., de Souza, P. & Mercer, R. (1986) Maximum mutual information estimation of hidden Markov model parameters for speech recognition. *In Proceedings of IEEE ICASSP '86*, 11, 49–52.

Bahl, L., Jelinek, F., & Mercer, R. (1983) A maximum likelihood approach to speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5, 179–90.

Baker, J. (1975) The DRAGON system: An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23, 24–9.

Baker, J. M. (1989) DragonDictate – 30K: Natural language speech recognition with 30,000 words. *In Proceedings of Eurospeech '89*.

Baum, L. E. (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3, 1–8.

Bishop, C. M. (1989) *Pattern Recognition and Machine Learning*. New York: Springer.

Bourlard, H., Hermansky, H., & Morgan, N. (1996) Towards increasing speech recognition error rates. *Speech Communication*, 18, 205–31.

Bourlard, H. & Wellekens, C. J. (1989) Speech pattern discrimination and multilayer perceptrons. *Computer Speech and Language*, 3, 1–19.

Brown, P., Lee, C.-H., & Spohrer, J. (1983) Bayesian adaptation in speech recognition. In *Proceedings of IEEE ICASSP '83*, 8, 761–4.

Chelba, C. & Jelinek, F. (2000) Structured language modeling. *Computer Speech and Language*, 14, 283–332.

Chen, S. F. & Goodman, J. (1996) An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (pp. 310–18). Morristown, NJ: Association for Computational Linguistics.

Chen, S. F., Kingsbury, B., Mangu, L. et al. (2006) Advances in speech transcription at IBM under the DARPA EARS program. *IEEE Transactions on Audio, Speech and Language Processing*, 14, 1596–1608.

Cohen, J., Kamm, T., & Andreou, A. (1995) Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *Journal of the Acoustical Society of America*, 97, 3246–7.

Cooke, M. & Ellis, D. P. W. (2001) The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35 141–77.

Cooke, M., Green, P., Josifovski, L., & Vizinho, A. (2001) Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34, 267–85.

Cox, S. J. & Bridle, J. S. (1989) Unsupervised speaker adaptation by probabilistic spectrum fitting. In *Proceedings of IEEE ICASSP '89*, 1, 294–7.

Dempster, A., Laird, N., & Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Deng, L. & Huang, X. (2004) Challenges in adopting speech recognition. *Communications of the ACM*, 47, 69–75.

Digalakis, V. V., Rtischev, D., & Neumeyer, L. G. (1995) Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3, 357–66.

Durbin, R., Krogh, A., Mitchison, G., & Eddy, S. R. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Eide, E. & Gish, H. (1996) A parametric approach to vocal tract length normalization. In *Proceedings of IEEE ICASSP '96*, 1, 346–8.

Faria, A. & Gelbart, D. (2005) Efficient pitch-based estimation of VTLN warp factors. In *Proceedings of Interspeech '05.*

Ferguson, J. D. (1980) Hidden Markov analysis: An introduction. In J. D. Ferguson (ed.), *Hidden Markov Models for Speech*. Princeton, NJ: IDA-CRD.

Fisher, W. M., Doddington, G. R., & Goudie-Marshall, K. M. (1986) The DARPA speech recognition research database: Specifications and status. In *Proceedings of DARPA Workshop on Speech Recognition*, 93–9.

Fisher, W. M. & Fiscus, J. G. (1993) Better alignment procedures for speech recognition evaluation. In *Proceedings of IEEE ICASSP '93*.

Forney, G. D., Jr. (1973) The Viterbi algorithm. *Proceedings of the IEEE*, 61, 268–78.

Furui, S. (1986) Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34, 52–9.

Gales, M. J. F. (1998) Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12, 75–98.

Gales, M. J. F. (2000) Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 8, 417–28. doi: 10.1109/89.848223.

Gales, M. J. F. & Woodland, P. C. (1996) Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10, 249–64.

Gales, M. J. F. & Young, S. J. (1996) Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, 4, 352–9. doi: 10.1109/89.536929.

Gauvain, J.-L. & Lamel, L. (2000) Large-vocabulary continuous speech recognition: Advances and applications. *Proceedings of the IEEE*, 88, 1181–200.

Gauvain, J.-L. & Lee, C.-H. (1994) Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2, 291–8. doi: 10.1109/89.279278.

Gold, B. & Morgan, N. (2000) *Speech and Audio Signal Processing*. New York: Wiley.

Goodman, J. T. (2001) A bit of progress in language modeling. *Computer Speech and Language*, 15, 403–34.

Gunawardana, A., Mahajan, M., Acero, A., & Platt, J. (2005) Hidden conditional random fields for phone classification. In *Proceedings of Interspeech '05*, Lisbon, Portugal, 1117–20.

Hain, T. (2002) Implicit pronunciation modelling in ASR. In *Proceedings of the ISCA Pronunciation Modeling Workshop*.

Hain, T., Burget, L., Dines, J., et al. (2007) The AMI system for the transcription of speech in meetings. In *Proceedings of IEEE ICASSP '07*.

Hain, T., Woodland, P. C., Evermann, G., et al. (2005) Automatic transcription of conversational telephone speech. *IEEE Transactions on Speech and Audio Processing*, 13, 1173–85. doi: 10.1109/TSA.2005.852999.

Hain, T., Woodland, P. C., Niesler, T. R., & E. Whittaker, W. D. (1999) The 1998 HTK system for transcription of conversational telephone speech. In *Proceedings of IEEE ICASSP '99*, 57–60.

Hermansky, H. (1998) Should recognizers have ears? *Speech Communication*, 25, 3–27.

Hermansky, H. & Morgan, N. (1994) RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2, 578–89. doi: 10.1109/89.326616.

Hori, C. & Furui, S. (2003) A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 5, 368–78.

Huang, X., Acero, A., & Hon, H.-W. (2001) *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ: Prentice-Hall.

Irino, T. & Patterson, R. (2002) Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilized wavelet-Mellin transform. *Speech Communication*, 36, 181–203.

Jelinek, F. (1976) Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64, 532–56.

Jelinek, F. (1991) Up from trigrams! The struggle for improved language models. In *Proceedings of Eurospeech* (pp. 1037–40). Genoa, Italy.

Jelinek, F. (1998) *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Jelinek, F. & Mercer, R. L. (1980) Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop Pattern Recognition in Practice* (pp. 381–97). Amsterdam: North-Holland.

Juang, B.-H., Levinson, S., & Sondhi, M. (1986) Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, 32, 307–9.

Jurafsky, D. & Martin, J. H. (2008) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

Katz, S. (1987) Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35, 400–1.

Kim, J.-H. & Woodland, P. C. (2003) A combined punctuation generation and speech recognition system and its performance enhancement using prosody. *Speech Communication*, 41, 563–77.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007) Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, 121, 723–43.

Kneser, R. & Ney, H. (1995) Improved backing-off for M-gram language modeling. In *Proceedings of IEEE ICASSP '95*.

Koumpis, K. & Renals, S. (2005) Content-based access to spoken audio. *IEEE Signal Processing Magazine*, 22, 61–9.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., & Gopinath, R. (2006) Super-human multi-talker speech recognition: The IBM 2006 Speech Separation Challenge system. In *Procedings of Interspeech '06*, 97–100.

Kuhn, R., Junqua, J. C., Nguyen, P., & Niedzielski, N. (2000) Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8, 695–707.

Lee, K.-F. (1989) *Automatic Speech Recognition: The Development of the SPHINX system*. Dordrecht: Kluwer.

Lee, L. & Rose, R. C. (1996) Speaker normalization using efficient frequency warping procedures. In *Proceedings of IEEE ICASSP '96*, 1, 353–6.

Leggetter, C. J. & Woodland, P. C. (1995) Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9, 171–85.

Lennig, M. (1990) Putting speech recognition to work in the telephone network. *Computer*, 23, 35–41.

Levin, E., Pieraccini, R., & Eckert, W. (2000) A stochastic model of human–machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8, 11–23.

Lippmann, R. P. (1997) Speech recognition by machines and humans. *Speech Communication*, 22, 1–15.

Liu, F., Stern, R., Acero, A., & Moreno, P. (1994) Environment normalization for robust speech recognition using direct cepstral comparison. In *Proceedings of IEEE ICASSP '94*.

Makhoul, J., Kubala, F., Leek, T., et al. (2000) Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, 88, 1338–53.

Manning, C. D. & Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McTear, M. (2002) Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys*, 34, 90–169.

Morgan, N. & Bourlard, H. A. (1995) Neural networks for statistical recognition of continuous speech. *Proceedings of the IEEE*, 83, 742–72. doi: 10.1109/5.381844.

Morgan, N., Zhu, Q., Stolcke, A., et al. (2005) Pushing the envelope – aside. *IEEE Signal Processing Magazine*, 22, 81–8.

Nadas, A. (1983) A decision theorectic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31, 814–17.

Nadas, A. (1985) On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33, 1414–16.

Neal, R. M. & Hinton, G. E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (ed.), *Learning in Graphical Models* (pp. 355–68). Dordrecht: Kluwer.

Ney, H. & Ortmanns, S. (2000) Progress in dynamic programming search for LVCSR. *Proceedings of the IEEE*, 88, 1224–40.

Normandin, Y. (1996) Maximum mutual information estimation of hidden Markov models. In C.-H. Lee, F. K. Soong, & K. K. Paliwal (eds.), *Automatic Speech and Speaker Recognition* (pp. 58–81.) Dordrecht: Kluwer.

Olsen, P. A. & Gopinath, R. A. (2004) Modeling inverse covariance matrices by basis expansion. *IEEE Transactions on Speech and Audio Processing*, 12, 37–46.

Ostendorf, M. (1999) Moving beyond the "beads-on-a-string" model of speech. In *Proceedings of the IEEE ASRU Workshop*.

Poritz, A. B. (1988) Hidden Markov models: a guided tour. In *Proceedings of IEEE ICASSP '88*, 7–13.

Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., & Zweig, G. (2005) fMPE: Discriminatively trained features for speech recognition. In *Proceedings of IEEE ICASSP 2005.*

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 257–86.

Raj, B. & R. Stern, M. (2005) Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, 22, 101–16.

Redner, R. A. & Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26, 195–239.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., & Franco, H. (1994) Connectionist probability estimators in HMM speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2, 161–74.

Robinson, A. J. (1994) An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5, 298–305.

Robinson, A., Hochberg, M., and Renals, S. (1996) The use of recurrent networks in continuous speech recognition. In C. H. Lee, K. K. Paliwal, & F. K. Soong (eds.), *Automatic Speech and Speaker Recognition: Advanced Topics* (pp. 233–58). Dordrecht: Kluwer.

Rosenfeld, R. (2000) Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88, 1270–8.

Scharenborg, O. (2007) Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49, 336–47.

Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., & Makhoul, J. (1985) Context-dependent modeling for acoustic-phonetic recognition of continuous speech. In *Proceedings of IEEE ICASSP '85*, 10, 1205–8.

Siohan, O., Myrvoll, T. A., & Lee, C.-H. (2002) Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer Speech and Language*, 16, 5–24.

Stolcke, A. & Shriberg, E. (1996) Statistical language modeling for speech disfluencies. In *Proceedings of IEEE ICASSP '96*, 1, 405–8.

Stolcke, A., Shriberg, E., Hakkani-Tür, E., & Tür, G. (1999) Modeling the prosody of hidden events for improved word recognition. In *Proceedings of Eurospeech '99*, Budapest, 311–14.

Varga, A. & Moore, R. (1990) Hidden Markov model decomposition of speech and noise. In *Proceedings of IEEE ICASSP '90*, 845–48.

Venkataramani, V., Chakrabartty, S., & Byrne, W. (2007) Ginisupport vector machines for segmental minimum Bayes risk decoding of continuous speech. *Computer Speech and Language*, 21, 423–42.

Wegmann, S., McAllaster, D., Orlo, J., & Peskin, B. (1996) Speaker normalization on conversational telephone speech. In *Proceedings of IEEE ICASSP '96*, 1, 339–41.

Welling, L., Ney, H., & Kanthak, S. (2002) Speaker adaptive modeling by vocal tract normalization. *IEEE Transactions on Speech and Audio Processing*, 10, 415–26.

Woodland, P. C. (2001) Speaker adaptation for continuous density HMMs: A review. In *Proceedings of the Isca Workshop on Adaptation Methods for Speech Recognition*, 11–19.

Woodland, P. C. & Povey, D. (2002) Large scale discriminative training of hidden Markov models for speech recognition. *Computer Speech and Language*, 16, 25–47.

Young, S. J., Adda-Dekker, M., Aubert, X., et al. (1997) The European SQALE project. *Computer Speech and Language*, 11, 73–89.

Young, S. J., Odell, J. J., & Woodland, P. C. (1994) Tree-based state tying for high accuracy acoustic modelling. *In Proceedings of the Workshop on Human Language Technology*, 307–12.

Young, S. J. & Woodland, P. C. (1994) State clustering in hidden Markov model-based continuous speech recognition. *Computer Speech and Language*, 8, 369–83.

Index

References to notes, figures and tables are entered as (respectively) 21n, 21*f* or 21*t*.