Functional Phonology

Functional Phonology

Formalizing the interactions between
articulatory and perceptual drives

Academisch proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof. dr. J.J.M. Franse
ten overstaan van een door het
college voor promoties ingestelde commissie
in het openbaar te verdedigen in de Aula der Universiteit

op maandag 14 september 1998 te 13.00 uur

door

Paulus Petrus Gerardus Boersma
geboren te Sint Nicolaasga

Faculteit der Geesteswetenschappen
Universiteit van Amsterdam

Promotor: prof. dr. ir. L.C.W. Pols
Overige commissieleden: prof. dr. C.H.M. Gussenhoven
dr. V.J. van Heuven
prof. dr. G.E. Booij
prof. dr. A.M. Bolkestein
dr. N.S.H. Smith

Uitgave van
Holland Academic Graphics
P.O. Box 53292
2505 AG The Hague
The Netherlands

phone: +31 70 4480203
fax: +31 70 4480177
e-mail: mail@hag.nl
http: www.hag.nl

ISBN 90-5569-54-6
NUGI 941

Printed in the Netherlands.

Contents

Acknowledgements

I wish to thank all those who discussed parts of this book with me: Louis Pols, Mirjam Ernestus, Ton Wempe, David Weenink, Robert Kirchner, Markus Walther, Youngshik Hwangbo, Mark Ellison, Yishai Tobin, Bruce Hayes, Antony Dubach-Green, Carlos Gussenhoven, Benji Wald, Vincent van Heuven, and Jan van Dijk. Most of the time, however, I was engaged in imaginary discussions with the authors of various books and articles. I thank them for their writings; their names are in the References section of this book. Finally, I wish to thank the members of the Institute of Phonetic Sciences for their support and friendship; and Jacqueline, Jelle, and Jildou for being there.

Introduction

This book is concerned with the use of sound in language. It develops a theory that uses general capabilities of human motor behaviour and perception to *explain* as well as *describe* the data of the languages of the world. We can predict as well as clarify generalizations about the organization of human speech and solve many outstanding controversial phonological issues, just by separating the roles of the ***articulation*** and the ***audition*** of speech sounds. Providing a synthesis between the “phonetic” and “phonological” standpoints, the theory of functional phonology expresses explanatory functional principles like the minimization of articulatory effort and the minimization of perceptual confusion directly in a descriptive formal grammar, and proves to be a typologically and empirically adequate replacement for generative theories of autosegmental phonology and feature geometry.

After making explicit (in Part I of this book) some of the dichotomies and relations between articulation and perception, I will answer (in Part II) the question of what segmental phonology would look like if it adhered to functional principles of speech production and perception. In Part III, I will assess the empirical adequacy of such a theory in various subareas of phonology, by confronting it with data from the languages of the world.

0.1 Articulatory and perceptual representations

Part I (“Representations”, chs. 1-5) treats some of the entities involved in the organization of spoken language. Chapter 1 stresses the contrasting roles of articulatory and perceptual features (as opposed to the traditional hybrid representations of generative phonology), and proposes a rigorous division of labour between perceptual input specifications, articulatory implementations, and perceptual output representations.

To corroborate the functional explanations proposed in later chapters, which are often stated in terms of articulation-perception interactions, I decided to make use of a computer-simulation model of speech production and perception. I developed a new comprehensive model of the speech apparatus (ch. 2), together with a numerical simulation of its aerodynamics and myoelastics (ch. 3). With the help of some simple perceptually-based analysis methods (ch. 4), we can check the suitability of the articulation model for simulating several speech-like events (ch. 5).

0.2 Functional principles and constraints of articulation and perception

Part II ("Constraints", chs. 6-13) treats some of the relations between the representations identified in chapter 1, and develops a functional theory about the subject matter of autosegmental phonology.

The functional hypothesis for linguistics maintains that the primary function of a language is communication. The aim for efficient and effective communication can be expressed in a number of ***functional principles***, which were first formulated in explanations for sound change. According to Passy (1891: 229; my translations), sound changes have the same cause that motivates the existence of language itself: "**we speak in order to be understood**".

0.2.1 Functional principles of speech production

Passy (1891: 227) states the *principle of economy*: "languages tend to get rid of anything that is superfluous", and the *principle of emphasis*: "languages tend to stress or exaggerate anything that is necessary". These principles are of a composite nature: the use of the terms *superfluous* and *necessary* expresses the idea that articulatorily motivated constraints may be honoured unless stronger perceptually motivated constraints are violated. We can, therefore, disentangle Passy's two principles into a more fundamental speaker-oriented principle of ***minimization of articulatory effort*** and an equally basic listener-oriented principle of ***minimization of perceptual confusion***.

0.2.2 Functional principle of the communication channel

One of the aspects of Passy's principle of economy translates into the principle of the ***maximization of information flow***, which we could phrase as "put as many bits of information in every second of speech as you can".

0.2.3 Functional principles of speech perception

Passy ignored any functional principles on the part of the listener. In order to accomplish an adequate understanding of phonology, we will have to include some.

First, we have ***maximization of recognition***: the listener will try to make maximum use of the available acoustic information, because that will help her reconstruct the meaning of the utterance.

Second, there is ***minimization of categorization***: in a world of large variations between and within speakers, the disambiguation of an utterance is facilitated by having large perceptual classes into which the acoustic input can be analysed: it is easier to divide a perceptual continuum into two categories than it is to divide it into five.

0.2.4 The functional hypothesis for phonology

Thus, I will maintain that historical sound changes, synchronic phonological processes, and the structure of sound inventories are built in such a way that the following natural drives will be honoured:

(a) The speaker will minimize her articulatory and organizational effort, i.e., she will try to minimize the number and complexity of her gestures and coordinations.
(b) The speaker will minimize the perceptual confusion between utterances with different meanings.
(c) The listener will minimize the effort needed for classification, i.e., she will use as few perceptual categories as possible.
(d) The listener will minimize the number of mistakes in recognition, i.e., she will try to use the maximum amount of acoustic information.
(e) The speaker and the listener will maximize the information flow.

0.2.5 Conflicts between functional principles

The principles identified in §0.2.4 are inherently conflicting:

- Minimization of effort often conflicts with minimization of confusion. Citing Passy (1891: 224): “The consonant [r] seems difficult to acquire, and often changes into [ʀ] or [ɹ]; but this tendency can be victoriously fought by a conscious or unconscious pursuit of clarity, [r] being a particularly sonorous and distinct sound.”

- Minimization of categorization sometimes conflicts with maximization of recognition. For instance, the tendency of Dutch listeners to put English /æ/ and /ɛ/ into the same perceptual category, will slightly hamper their understanding of English utterances. On the other hand, these functional principles of perception sometimes collaborate: if a contrast between two perceptual classes is not reliable, i.e., if an acoustic feature is sometimes classified into an adjacent category, successful recognition is actually helped by not trying to use this contrast for disambiguating utterances. If the listener accepts the phonological ambiguity of an utterance, she will take recourse to alternative (semantic, pragmatic) disambiguation strategies, which might otherwise not have been invoked. Labov (1994) showed that this principle can be responsible for segment merger in cases of dialect mixture (§17.1.4).

- Maximization of information flow conflicts with both minimization of effort and minimization of categorization (§9.6).

- Conflicts also arise *within* the various principles, e.g., the minimization of the number of gestures conflicts with the minimization of energy.

Conflicts like these have been noticed in other areas of linguistics. In the realm of syntactic theory, for instance, the theory of **Functional Grammar** (Dik 1978, 1989, 1997) acknowledges the existence of potentially conflicting functional principles for constituent ordering: “– The actual constituent ordering patterns found in a language are the resultant of a number of interacting principles. – Each of these principles is in itself functionally motivated (...) – [N]o language can conform to all the ordering principles at the same time or to the same degree. (...) – Shifts in the relative force of the different principles may lead to (sometimes radical) changes in constituent ordering.” (Dik 1989: 337)

0.2.6 Formalizing functional principles

The hypothesis that languages are organized in a way that reflects the primacy of communication, has met with little enthusiasm on the part of generative linguists. This is partially due to the lack of formalizability of grammars that consist of interacting functional principles. In the realm of speech, for instance, generative phonology has often been able to reach descriptive adequacy by proposing sets of sequentially ordered formal ***rules***; by contrast, functionalist accounts like those of Passy (1891), Martinet (1955), and Boersma (1989), while explaining several facts of language from potentially conflicting functional principles, have failed to give adequate descriptions of the behaviour of actual speakers, i.e., they have failed to produce any formal ***grammars***.

Fortunately, the advent of **Optimality Theory** (OT; Prince & Smolensky 1993), though rooted in the generative tradition (its original version explicitly denied any role for function in the grammar), has allowed us to put an end to this situation. This theory proposes that phonological grammars consist of allegedly innate ***constraints***, each of which can be ***violated*** if it is crucially ***dominated*** by a stronger constraint; the interaction between the constraints is based on the principle of ***strict ranking***, i.e., a high-ranked constraint will always outweigh any number of lower-ranked constraints. This scheme of evaluating candidate surface representations is perfectly amenable to the interaction of functional principles (Jun 1995; Flemming 1995; Hayes 1995, 1996ab; Boersma 1997abcde, 1998ab; Kirchner 1998). To stay with Passy's example: a speaker will turn [r] into [ʀ] or [ɹ] if the constraint expressing the articulatory effort of [r] dominates the constraint that aims for clarity, and she will pronounce [r] faithfully if the clarity constraint is ranked higher. In chapter 6, I will present the functional version of OT and show that if we express articulatory and perceptual principles directly as constraints in the language user's production and perception grammars, the desired properties of their interactions will follow from the Optimality-theoretic notion of violability: because the principles are inherently conflicting, the corresponding constraints, if stated in their naked, most general forms, must be violable.

Chapters 7, 8, and 9 show how functional principles translate into a plethora of constraints or universal constraint rankings, ready for immediate use in autosegmental phonology. In the production grammar, the principle of minimization of articulatory effort branches into many families of articulatorily motivated constraints (ch. 7), formulated within a space of articulatory gestures indentified in ch. 1; likewise, the principle of minimization of perceptual confusion branches into many families of input-output faithfulness constraints (ch. 9), formulated within a space of perceptual features identified in ch. 1. In the perception grammar, constraints against perceptual confusion branch into families of categorization constraints (ch. 8), likewise formulated within a space of perceptual features. All these constraint families can be ranked individually in each language.

0.2.7 Interactions between the constraints

The remaining part of this book (chs. 10-19) will centre on the interactions between the constraints.

Phonological structures and processes follow from the interaction between faithfulness and articulatory constraints. As a first example, chapter 10 describes how this interaction determines the realization of vowel height in phonetic implementation, and how phonetic and pragmatic circumstances influence the result by shifting the rankings of the constraints. A transition from gradient to discrete vowel reduction follows naturally.

The ***local-ranking principle*** (ch. 11), rooted in general properties of motor behaviour and perception, determines which constraints can be ranked universally, and which must be ranked on a language-specific basis. The examples of nasal place assimilation and obstruent voicing illustrate the typological adequacy of this approach. It leads to a straightforward strategy for the phonologization of phonetic principles.

Faithfulness to specifications of “vertical” (simultaneous) and “horizontal” (sequential) perceptual connections creates the illusions of segments and autosegments in the grammar (ch. 12).

Many arguments for all-or-none (instead of gradient) ***underspecification*** vanish if we distinguish between articulatory and perceptual features, and between high- and low-ranked specifications (chs. 13, 17).

0.3 Production and perception grammars

Part III (“Grammar”, chs. 14-19) shows that functionally based constraints can settle several recalcitrant issues in phonology. With the help of the distinction between articulation and perception, we can solve problems in the study of acquisition (ch. 14), segmental inventories (ch. 16), sound change (ch. 17), and synchronic autosegmental phenomena like spreading (ch. 19) and the Obligatory Contour Principle (ch. 18). Traditionally separate devices like the segment (ch. 12), spreading (chs. 11, 19), licensing (ch. 12), underspecification (chs. 13, 17), feature geometry (chs. 1, 19), and OCP effects (chs. 12, 18), will appear to be surface phenomena created by the interaction of more fundamental principles.

Phonological features, representations, and constraints need not be ***innate***, because they can be ***learned*** (ch. 14). If constraint evaluation has a noisy component, we can even learn stochastic grammars, i.e., we can learn to reproduce our parents’ degree of variation and optionality (ch. 15). The finiteness of the number of feature values in every language is a result of general properties of motor learning and perceptual categorization, and explains the symmetries found in segment inventories (ch. 16); the gaps in these inventories are explained by universal local rankings of constraints, which we need not learn because they appeal to general capabilities of human motor behaviour and perception.

The subjects treated in part III lie in the realm of common phonological debate; by handling them successfully, the functional theory of phonology, developed on a priori grounds in part II, may become an acceptable alternative to theories that start from the data of the languages of the world, because of its capacity of generating less ad-hoc accounts of these data, which, after all, have the last word on the empirical adequacy of any theory.

Part I

REPRESENTATIONS

Chapter 1 discusses the need for a principled distinction between articulatory and perceptual features and representations in phonology. Chapters 2 and 3 introduce a computational model of how we can determine the “automatic” acoustic output from specifications of muscle lengths and tensions. This model will be used to corroborate statements about the interaction between articulation and perception in the “phonological” parts II and III. Chapter 4 treats some models of perception that we will need, and chapter 5 tests the workings of the articulation model in the simulation of vowels and consonants.

1 Representations and features

Abstract. Phonological theory would benefit from making a principled distinction between articulatory and perceptual features and representations.

In order to define functional phonological grammars, we need sets of constraints to put into those grammars. In order to identify these functional constraints, we need to know on what representations these constraints perform their evaluations. As will be apparent from the introduction, the representations that can be subjected to an evaluation of functional principles are the articulatory and perceptual representations of speech utterances. The current chapter identifies these representations. We will start with an example that will recur throughout this book.

1.1 Articulatory and perceptual representations of an utterance

Consider the English utterance *tense*. Its underlying phonological form is

|tɛns| (1.1)

I will take this to be the ***perceptual specification*** of the utterance: if the speaker produces the specified perceptual features in the specified order with the specified time alignment, the listener will recognize the utterance as |tɛns|, and a substantial part of the communication has succeeded. This basic insight should be reflected in our theory of grammar.

Several articulatory strategies can be followed to implement the utterance (1.1). In some varieties of English, a part of the ***articulatory implementation*** is (time runs from left to right; *critical* = constricted so as to produce friction):

<table>
<tr><td>tongue tip</td><td colspan="1">closed</td><td colspan="3">open</td><td colspan="3">closed</td><td colspan="1">critical</td></tr>
<tr><td>velum</td><td colspan="3">closed</td><td colspan="2">open</td><td colspan="3">closed</td></tr>
<tr><td>glottis</td><td colspan="2">wide</td><td colspan="4">narrow</td><td colspan="2">wide</td></tr>
<tr><td>lips</td><td colspan="8">spread</td></tr>
</table>

(1.2)

This will give rise to an acoustic output that we can translate into the following table of perceptual phonetic events, time-aligned with the articulatory score (1.2) (“+” = present; *tr* = transition; *side* = the resonance of a side branch, see §19.1.2; *bu* = burst; *cont* = continuant; *asp* = aspiration; *max* = maximum):

silence	+					+	
coronal		burst		tr.	side		bu. cont
voice			sonorant				
noise		asp					sibilant
F1		open mid					
F2		max					
nasal				+			

(1.3)

In a microscopic transcription (§1.3.3), this ***perceptual result*** can be written as [[thɛɛ̃n_ts]] ("_" = silence). With the help of the processes of categorization and recognition, the listener may reconstruct |tɛns|.

The theory of Functional Phonology, introduced in this book, claims that the principle of ***minimization of articulatory effort*** (§0.2.1) evaluates the articulatory implementation (1.2) and its competitors, and that the principle of ***minimization of perceptual confusion*** (§0.2.1) evaluates the differences between the perceptual specification (1.1) and the perceptual result (1.3). Together, these principles will determine which candidate articulatory implementation will actually be chosen to surface.

In the present chapter, I will defend the hypothesis that the distinction between articulation and perception is an integral part of the grammar. This involves determining the nature of the phonological spaces (§1.2) and representations (§1.3) on which the functional constraints will be defined, and will lead to a replacement of the traditional hybrid features and representations with systems based on general properties of human motor behaviour and perception.

1.2 Articulatory, perceptual, and hybrid features

A thread of this work is the idea that features of speech sounds, language-dependent though they may be, can be divided into two large classes: articulatory and perceptual features. These two groups play different roles in phonology, and an awareness of the difference between them will solve many hitherto unsettled problems in several realms of phonological debate.

The difference between the two groups of features can be traced to their different roles in speech production and perception.

1.2.1 Articulation versus perception in speech production

Figure 1.1 shows a simplified view of how the articulatory and perceptual aspects of phonology are integrated into speech production. The point labelled "start" marks the interface of the rest of the grammar to the phonological/phonetic component. In the following paragraphs, I will explain this figure. The main point that I am trying to make, is that phonology controls both the articulatory and the perceptual specifications of the utterance, i.e. both the representations that we saw in (1.1) and (1.2).

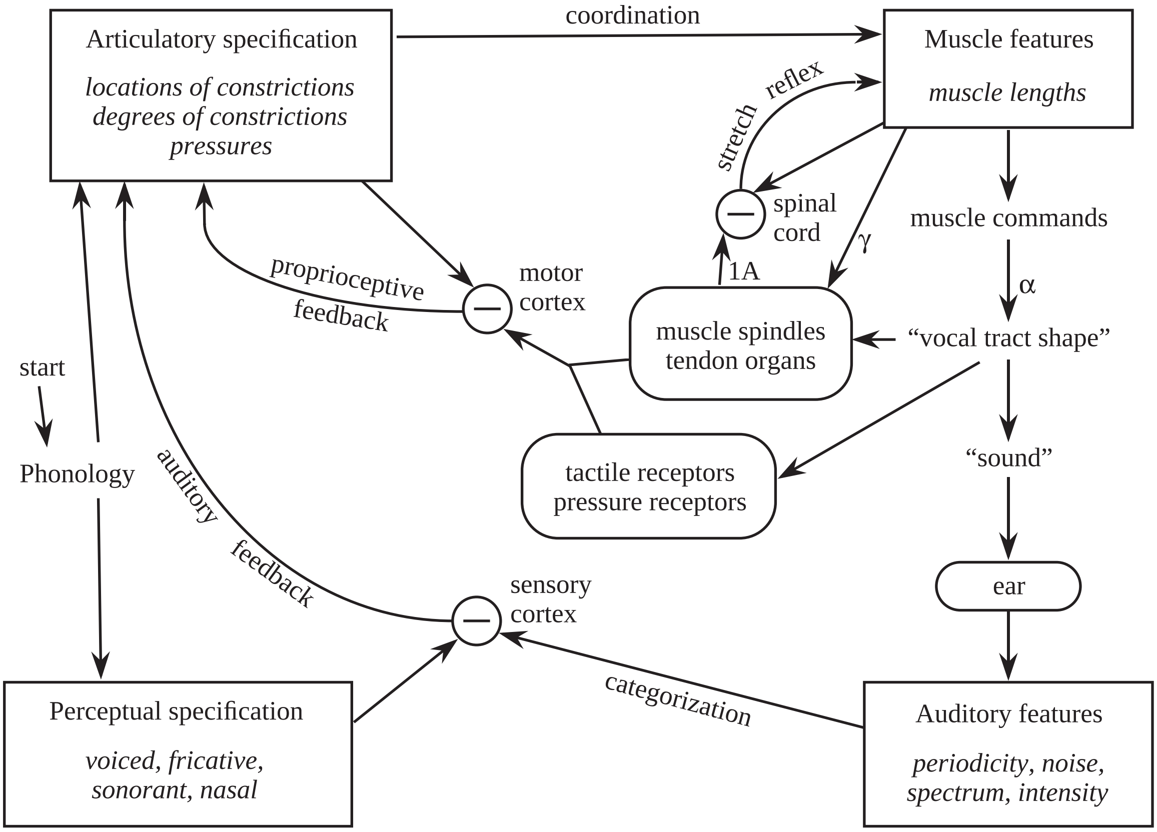


Fig. 1.1 Integration of phonology into speech production.
Rectangles = representations. Rounded rectangles = sensors.
Encircled minus signs = comparison centres. Arrows = causation.
α, γ, 1A = nerve fibers.

Top right: control of muscle length. The speaker can control the ***tension*** of a muscle. For this, a direct *muscle command* (every term set in plain italics can be found in figure 1.1) is conducted by the α neuron fibers from the spinal cord or the brain stem to the muscle fibers, whose contraction then results in a change in the shape of the human body, e.g., a change in *vocal tract shape*. The length and length change of a muscle are measured by the *muscle spindles* (and the tension by the *tendon organs*), which send this information back (through the afferent fibers marked *1A*) to the *spinal cord* or the brain stem. If the muscle is stretched by an external cause, a direct excitatory synapse of the afferent with the α motor neuron then causes the *stretch reflex*: a compensatory contraction of the muscle.

With the help of the γ efferent fibers, the muscle spindles can be actively stretched, so that the afferents fool the spinal cord into thinking that the muscle itself is stretched by an external cause. Consequently, the reflex mechanism described above will cause the muscle to contract. Thus, while direct α activity would cause an uncontrolled contraction, this γ-loop system, which does not go further up than the spinal cord, can be used to control ***muscle length*** (Hardcastle 1976; Gentil 1990). The learning of a fast, shape-oriented gesture probably involves the learning of an efficient mix of α and γ activity, innervating the muscle spindles simultaneously with the other fibres.

Conclusion: the speaker can set her muscles to a specified length. In chapter 2, I will present a computational model of the vocal apparatus that is controlled by setting the lengths (and a few tensions) of 29 muscles and muscle pairs.

Top left: control of articulator position. For most gestures, the control of muscle length is not sufficient. Rather, the motor cortex specifies the actual position of the body structures. For the vocal tract, this means that the *locations* and *degrees of constrictions* are specified. That the muscle lengths as such are not the target positions specified in speech production, can be seen from bite-block experiments (Lindblom, Lubker & Gay 1979): speakers immediately compensate for the constraints on the jaw, even before phonating, in such a way that the tongue muscles bring about approximately the same area function in the vocal tract as in normally articulated vowels, while having very different shapes.

The proprioceptive sensory system, consisting of muscle spindles, tendon organs, *tactile receptors*, and *pressure receptors*, sends the information about the realized shapes back to the motor cortex, where it is compared to the intended shapes, i.e., the *articulatory specification*, and appropriate action is taken if there are any differences. This system is called *proprioceptive feedback*.

Conclusion: the speaker can directly control muscle tensions, muscle lengths, and the locations and degrees of the constrictions in the vocal tract.

Hypothesis: the phonological component of the speaker's grammar can specify any of these articulatory variables.

Right side: generation of sound. The step from "vocal tract shape" to "sound" involves no actions of the speaker or listener: the sound is the automatic acoustic result of the muscle tensions, positions, and movements. In chapter 3, I present a comprehensive physical-mathematical model of this automatic conversion, and an algorithm for its implementation on a computer.

Bottom right: auditory perception. The human *ear* will analyse any *sound*, perhaps one arising from a speech utterance, into *auditory features* like *periodicity* (pitch and noisiness), *spectrum* (timbre), and *intensity* (loudness), all of them functions of time. I will illustrate the perceptual part of speech production with the development of phonology in young children.

The infant is born with an innate control of some of the gestures that are also used in speech: breathing, vocal-fold adduction (crying), and repetitive jaw movements (drinking). Other gestures, like the movements of the limbs, are still largely uncoordinated. After a few months, the infant learns that she can control her environment (i.e. her perceptual impressions), by contracting some muscles. Like the use of one of her deltoid muscles gives her the visually pleasing result of a swinging object (her arm), a certain combination of expiration and vocal-fold adduction gives her the auditorily pleasing result of a periodic sound (voicing). A little later, when she has a command of some agonist/antagonist pairs, she will start exploring the benefits of repetitive movements; like hitting the mills and bells that are within her reach, she will superponate opening and closure gestures of the jaw on a background of phonation, thus getting nice alternations of silence and sound (babbling).

Conclusion: speakers learn the forward relationship between articulatory coordinations (top left) and perceptual results (bottom right). I will return to this in the chapter on phonological acquisition (ch. 14).

Bottom left: speech perception. At the time she starts to imitate the speech she hears, the little language learner will have to compare her own utterance with the model (*auditory feedback*). At first, the *perceptual specification* (initially, the adult utterance, see ch. 15), is an unsegmented gestalt. The articulatory specifications, which she is now constructing for the sake of faithful imitation and the reproduction of her own speech, are not very sophisticated yet either, because the orosensory (proprioceptive) feedback mechanism is still under development.

But the child learns to group perceptual events into categories. For speech, this ultimately leads to a language-dependent ***categorization*** of perceptual features. The skilled speaker will also have highly organized articulatory specifications in terms of degrees of constrictions and air pressures, with a language-dependent degree of underspecification, determined by economical considerations, i.e., the balance between perceptual invariance and articulatory ease. She will use the auditory feedback only as a check and for maintenance.

Conclusion: the speaker can compare the realized perceptual categories with the perceptual specification of the utterance.

Hypothesis: comparing the perceptual result and the perceptual specification of the utterance is integrated in the speaker’s organization of her speech.

1.2.2 The two targets of speech production: two levels of specification

For a skilled speaker, the perceptual specifications must be the ***ultimate*** (distal) targets of speech production. They cannot be the ***immediate*** (proximal) targets, because the auditory feedback loop is much too slow for that. The immediate targets are the locations and degrees of constriction and the air pressures in the vocal tract. These proprioceptive targets can be monitored by the collective effort of tactile and pressure receptors, muscle spindles, tendon organs, and joint receptors.

The ***task-dynamic*** approach advocated by Kelso, Saltzman, & Tuller (1986) and Browman & Goldstein (1986, 1990), maintains that the input to an articulation model should consist of specifications of ***tract variables***, such as locations and degrees of constrictions, as functions of time. This approach explicitly focuses on describing the coordination of the muscles of speech production: specification of these tract variables refers to *learned* motor behaviour. Kelso et al. notice, for example, that an experimentally induced perturbation of the movement of the jaw does not prevent the completion of the bilabial closure in [aba] or the achievement of an appropriate alveolar near-closure in [aza]. Thus, if the upper and lower teeth are externally constrained to be more than 1 cm apart (by a bite block), the required alveolar closure will still be attained. Crucially, however, the smallest bilabial closure will then be much larger than in the case of an unconstrained [aza]. Apparently (Kelso et al. argue), the immediate task for producing [b] is: “make a complete closure with the lips”, and for [z] it is: “make a near closure at the alveoli”. Crucially, the task for [z] does not specify bilabial closure at all; this is why there can be a large variation in the degree of bilabial closure during [z]. Therefore, there is some underspecification in the immediate targets of speech production.

However, as will be apparent from our separation of perceptual and articulatory specifications, a part of the ultimate *perceptual* specification of /z/ (in some languages) should be in these terms: “make a periodic sound that will produce strong high-frequency noise”. Speakers will learn that the only articulatory implementation (“task”) that achieves this, is: “make a near closure at the alveoli; meanwhile, the bilabial and dorsal constrictions should be wider than this alveolar constriction, the naso-pharyngeal port should be closed, the lungs should exert pressure, and the vocal cords should be in a position that enables voicing”. We see that the perceptual specification does require a constraint on bilabial closure after all (the lips must not be completely or nearly closed), and that the articulatory specification *follows* from the perceptual specification for /z/.

That the perceptual features, not the proprioceptive features, form the distal targets of speech production, can be seen in a simple experiment that embroiders on the bite-block experiments. If you ask someone to pronounce a central (e.g. Dutch) [a] with her teeth clenched, she will make compensating tongue and lip movements; however, because [a] is not specified for horizontal lip spreading, she will not draw the corners of her mouth apart, though this would yield a much more [a]-like sound; she will only learn this trick after some practice, using auditory feedback.

Conclusion: the articulatory specifications are the proximal targets of speech production, the perceptual specifications are the distal targets.

Hypothesis: the speaker’s phonology controls both the articulatory and the perceptual specifications.

1.2.3 Perceptual specifications

According to one of the hypotheses above, the speaker’s phonology compares the perceptual specification with the perceptual result of the utterance. As we will see later, she does this in order to evaluate the extent to which the functional principle of minimization of perceptual confusion is honoured. This evaluation takes place in a space of ***perceptual features***, which include periodicity (voicing and tone), noise (frication, aspiration), silence, burst, continuancy, and frequency spectrum (place, nasality).

All these features are measured along continuous scales, but languages discretize these scales into a language-dependent number of ***categories***. An example of the perceptual specification of labial sounds for a language that has two categories along each of the voicing, friction, sonorancy, and nasality scales, can be read from the following table, where ‘+’ means ‘present’, ‘–’ is ‘absent’ (suggesting a privative feature), and ‘|’ is a perceptual contour, i.e. a temporal change in the value of a perceptual feature:

	p	f	v	b	m	w	pʰ	ʋ	hʷ	u	b̤	ũ	ṽ
voiced	–	–	+	+	+	+	–	+	–	+	+	+	+
noise	–	+	+	–	–	–	–|+	–	+	–	–|+	–	+
sonorant	–	–	–	–	+	+	–	+	–	+	–	+	–
nasal	–	–	–	–	+	–	–	–	–	–	–	+	+

(1.4)

No universal feature values. The language-dependency of perceptual feature values can be most clearly seen from the different divisions of the height continuum for languages with three and four vowel heights (ch. 8): if the lowest vowel is [a] and the highest vowel is [i], a language with three vowel heights will have an “e” whose height is approximately midway between [a] and [i], and a language with four vowel heights will have two vowels close to canonical [ɛ] and [e]; this shows that the height continuum is divided on a basis of equal perceptual distance rather than on a basis of Chomsky & Halle’s (1968) maximum use of universal binary features.

1.2.4 Articulatory specifications

The speaker’s phonology will also evaluate the extent to which the functional principle of minimization of articulatory effort is honoured. This evaluation takes place in a space of ***articulatory features***, which include the possible positions, shapes, movements, and tensions of the lips, cheeks, tongue tip, tongue body, velum, tongue root, pharynx walls, epiglottis, laryngeal structures, vocal folds, and lungs. The trajectory of the implementation of the utterance through this articulatory space is a voyage along many positions, each of which is characterized as a vector measured along scales of degree of closure or tension. Though these scales are continuous, languages discretize most of them. For instance, supralaryngeal degrees of closure can be: *complete* (usually brought about by a ballistic movement: plosives and nasals); *critical* (usually brought about by a controlled movement, which makes it precise enough to maintain friction noise or vibration: fricatives); *approximant* (strong secondary articulation, pharyngealization); *narrow* (0.3 - 1 cm^2; high vowels, glides, liquids, retracted tongue root); *open* (1 - 4 cm^2; neutral vocalic); or *wide* (4 - 15 cm^2; spread lips, advanced tongue root).

I classified these degrees of closure according to perceptual differences, i.e., every pair of successive labels is found somewhere in the world to contrast two phonemes on the same articulator. Still, there is nothing canonical, preferred, or universal about this subdivision. Besides the obvious articulatory implementation of the language-dependent subdivision of vowel height, here is an example with non-vocalic closures: Dutch contrasts a noisy voiced labiodental fricative ([viɫ] ‘fell’) and a noiseless approximant ([ʋiɫ] ‘wheel’); in between those two, as far as noisiness and, therefore, degree of constriction are concerned, are the [v]-like sounds of German ([vaen] ‘wine’), English ([vain] ‘vine’), Afrikaans ([vət] ‘white’), and French ([vil] ‘city’).

The labial, coronal and dorsal articulators can be used independently to a large extent in doubly articulated sounds (labial-velars, clicks) or even triply articulated sounds (Swedish [ɧ], Holland Dutch syllable-final <l> [ɫ^w]), but there are no sounds that use the same articulator simultaneously twice (e.g. no clicks with dorso-palatal front closure). The articulatory space is organized in tiers, with one tier for every degree of opening and tension. The independence of these tiers represents the independence of the articulators, and reflects the independence of articulatory features in phonology.

An example of the articulatory specifications of some labial sounds in a language that would faithfully implement the perceptual features of (1.4), is given in the following table, where ‘0’ = closed, ‘1’ = critical, ‘2’ = approximant, ‘3’ = narrow, ‘4’ = open, ‘5’ = wide, ‘|’ = an articulatory contour (change in time), and ‘2-5’ = any value from 2 to 5:

	p	f	v	b	m	w	$\mathrm{p^h}$	ʋ	w̃	b̥	b̤	ɓ	$\mathrm{h^w}$	u	ɔ
lip opening	0	1	1	0	0	3	0|2-5	2	3	0	0|2-5	0	3	3	4
tongue tip opening	2-5	2-5	2-5	2-5	2-5	3-5	2-5	2-5	3-5	2-5	2-5	2-5	3-5	4-5	5
tongue body opening	2-5	2-5	2-5	2-5	2-5	3	2-5	2-5	3	2-5	2-5	2-5	3	3	4
velum opening	0	0	0	0	4	0-1	0	0-1	4	0	0	0	0-1	0-1	0-2
pharynx opening	2-5	2-5	2-5	2-5	2-5	3-5	2-5	2-5	3-5	2-5	2-5	2-5	3-5	4-5	3
glottis opening	2-3	2-3	1	1	1	1	3	1	1	1	2	1	3	1	1
supralar. tension	+			–						–	–	–			

(1.5)

Articulatory underspecification. There is a lot of underspecification in (1.5). For instance, if the lips are completely or almost closed, the coronal and dorsal constrictions have a lot of freedom: they can be anywhere between the approximant closure and a wide vocalic opening without affecting the perceptual features too much. As an example, consider the articulatory and perceptual features and specifications of [b] in the utterance [aba]. During the pronunciation of [a], the tongue will be pulled down in the throat. This state will last during the whole of the utterance [aba]. The jaw will travel a long distance in going from the [a] position to the [b] position and back again. The lips will also make a closing-opening movement. If, however, the lips are less closed, as in [u], the coronal constriction should be quite wide so that it will not sound like a front vowel, and the pharyngeal constriction should also be quite wide so that the vowel does not sound more open or centralized. Thus, as already argued in §1.2.2, the articulatory specifications follow from the perceptual specifications.

Conclusion: articulatory underspecification is constrained by faithfulness to perceptual invariance.

1.2.5 Perceptual versus articulatory features

Though it is often the case that similar articulations produce similar perceptual results, as with most place features, there are several sources of asymmetry between perceptual and articulatory features. In the following, I will disentangle the hybrid features used in generative phonology.

Voicing. If we define voicing as the vibration of the vocal cords, we are talking about the perceptual feature [voice], which refers to a high degree of periodicity in the sound. There is no single articulatory gesture that can be associated with voicing: for the vocal folds to vibrate, they must be close enough and air has to flow through the glottis with a sufficient velocity. The articulatory settings needed to implement the voicing feature, vary depending on the degree of constriction above the larynx. If the air is allowed to exit freely, as in sonorants, there is spontaneous voicing if the vocal folds have been adducted; sufficient airflow is then guaranteed.

If the passage is obstructed, as in [b], active laryngeal or supralaryngeal gestures are often needed to maintain voicing, especially in non-intervocalic position: the larynx may be lowered, the width of the glottis or the tension of the vocal folds may be adjusted, the walls of the pharynx, the cheeks, or the velum may be expanded passively or actively, or the stop may be pre-nasalized. The effects of all of these tricks have been confirmed in simulations with a simple model of the vocal tract (Westbury & Keating 1986) as well as with a more comprehensive model (§5.12). Since it is not always easy to find out which trick (other than implosion or prenasalization) is used by a specific language, we can supply plain voiced obstruents with the implementationally formulated articulatory feature [obstruent voicing] (or Steriade's 1995 suggestion [pharyngeally expanded], though the term "expanding" might be more correct).

Likewise, active gestures are sometimes needed for voiceless obstruents, especially in intervocalic position: widening or constriction of the glottis, raising of the larynx, stiffening of supralaryngeal walls, or active narrowing of the supralaryngeal tract. For this, we can similarly imagine a goal-oriented articulatory feature [obstruent devoicing].

Since assimilation processes are normally associated with changes of articulatory timing, we expect that obstruents can trigger voice assimilation, and that sonorants cannot. Acceptance of the distinction between articulatory and perceptual voicing features, will lead to a rejection of the main argument for underspecification in phonological processes (ch. 13). Thus, an early decision to posit a single feature [voice] for underlying and surface representations resulted in the underspecification of sonorants for this feature: the fact that many languages do contrast voiced and voiceless obstruents but do not contrast voiced and voiceless sonorants, combined with the phonological inertness (with respect to spreading) of voicing in sonorants, was considered evidence for the analysis that sonorants were not voiced at all underlyingly; a late rule would insert the voicing feature for sonorants. A distinction between an articulatory voicing feature, which only applies to obstruents because sonorants are spontaneously voiced, and a perceptual voicing feature common to sonorants and voiced obstruents, would quite simply solve the mysteries associated with the voicing problem. However, this will not go without a struggle: the one phenomenon that seems immune to a simple functional approach, NC voicing (i.e., the phenomenon that plosives tend to be voiced after nasals), tempted Itô, Mester & Padgett (1995) into the following remarks:

> "the trouble lies not with [voice], (...) the challenge is to resolve the paradox without destroying the unity and integrity of the distinctive feature [voice]." (Itô, Mester & Padgett 1995, p. 581)

Their resolution of the paradox entails that nasals, *because* they are redundantly voiced, like to share a non-redundant voicing feature with their neighbours; no explanation is given for the absence of CN voicing. An articulatory explanation was advanced by Hayes (1995): in the case of a voiced NC, the velum goes on raising even after the moment of closure, so that the enlarging pharyngeal cavity facilitates the maintenance of voicing; the exactly reverse situation from the CN case. The question of how such details are phonologized, is answered in chapter 11.

Noise. In the phonological literature, fricatives are economically divided into *non-strident* (/ɸ/, /θ/, /x/) and *strident* (/f/, /s/, /ʃ/, /χ/). In contrast with what the label suggests, this division is based on distributional grounds: the strident fricatives are louder (make more noise) than their non-strident counterparts *on the same articulator* (Chomsky & Halle 1968, p. 327), and are, therefore, on the average more suitable for human communication in a world with distances and background noise; the non-strident fricatives, on the other hand, often alternate, or are historically related to, plosives at the same place of articulation; as so happens, plosives tend to occur at locations where perfect closures are easy to make (bilabial, corono-postdental, dorso-velar), and fricatives prefer locations with small holes (labio-dental, corono-interdental) or unstable structures (dorso-uvular). From the perceptual standpoint, however, we could divide the continuous noise scale into four levels of a combined loudness/roughness nature (which is rather arbitrary, especially for the non-peripherals):

- [aspirated]: as in [h], [pʰ], and so-called “voiceless sonorants”.
- [mellow friction]: resulting from airflow through a smooth slit ([ɸ], [x]).
- [strident friction]: airflow along sharp edges ([f], [θ]) or loose structures ([χ]).
- [sibilant]: a jet of air generated in one place (alveolar) and colliding at a rough structure at another place (teeth): [s], [ʃ]; this causes a 15 dB intensity increase with respect to the normal strident [θ].[1] According to Ladefoged (1990a), the distance between the lower and upper teeth is critical,[2] and sibilants are the only English sounds with a precise specification for jaw height (see the discussion below for vowel height).

Sonorant. Chomsky & Halle’s (1968: 302) definition of sonorants is that they are “sounds produced with a vocal tract configuration in which spontaneous voicing is possible”. This is neither an articulatory nor a perceptual definition, and, as such, not likely to play a role in phonology. Since, as Ladefoged (1971: 109) states, “the rules of languages are often based on auditory properties of sounds”, I will simply take [sonorant] to refer to a high degree of loudness and periodicity that allows us to hear a clear formant structure.[3] Thus, [sonorant] implies [voice]. Its implementation is as follows. From the openings associated with each articulator, we can derive the following abstract openings:

- Oral opening. This equals the minimum of the labial, coronal, and dorsal openings.
- Suprapharyngeal opening. The maximum of the oral opening and the nasal opening.
- Supralaryngeal opening. Minimum of suprapharyngeal and pharyngeal openings.

These derivative features can help as intermediaries in formulating the mapping from articulatory to perceptual features. For instance, the supralaryngeal articulatory setting needed for spontaneous voicing is:

$$\textit{supralaryngeal opening} \geq \text{“approximant”} \qquad (1.6)$$

This condition is not sufficient, of course. Vocal-fold adduction and lung pressure have to be added.

Fricatives versus approximants. So-called voiceless sonorants aren’t that: they are just very mellow fricatives (aspirates). The binarily categorizing language of table (1.1) shows a perceptual contrast between fricatives and approximants, but only if these are voiced ([v] and [ʋ]), not if they are voiceless ([f] and [hʷ]). This is because a voiced

[1] Which the reader may verify by saying [sθsθsθsθsθ].

[2] The reader may verify that she cannot produce a faithfully sibilant [s] with a finger between her teeth.

[3] This raises the question whether [sonorant] can be considered a primitive feature at all: it can be seen as a *value* of a loudness feature, or as a derived feature based on the presence of formant structure.

approximant will not produce friction, but a voiceless (aspirated) articulation with the same degree of closure, will. So, voiced fricatives and approximants can occur together in such a language (e.g., Dutch [v] and [ʋ]), because voiced fricatives are noisy and voiced approximants are not; their voiceless counterparts cannot occur together in such a language, because voiceless fricatives and voiceless approximants only differ in their *degree* of noisiness, which would force the listener to distinguish between the categories [aspirated] and [fricative].

Nasality. The perceptual feature [nasal] more or less coincides with the articulatory feature [lowered velum]. But not precisely. Table (1.5) shows a less restricted nasal specification for [ɔ] than for [u]. A slightly open nasopharyngeal port is allowed in lower vowels, because it can hardly be heard if the oral opening is large (Van Reenen 1981). Thus, the same small amount of velum lowering may give rise to a perception of nasality in high vowels, and of no nasality in low vowels.

Continuant. This feature has been used to distinguish plosives from fricatives, and to be able to treat nasal and “oral” stops as a natural class. As a perceptual feature for audible oral airflow, I will replace it with [oral]; thus, [f], [h], and [a] are oral, and [p] and [m] are not, while [ã] is both oral and nasal. This move reflects the articulatory symmetry between the nasal and oral pathways. However, because most speech sounds are oral but not nasal, commonness considerations (§9.5) lead us to expect that the values [–oral] and [+nasal] play more visible roles in phonological processes than their counterparts [+oral] and [–nasal].

In another respect, oral stricture works just like velar stricture: the degree of perceived oral airflow does not necessarily reflect the degree of closure. A sound made with the articulatory setting for a labial fricative will normally lose its friction when the velum is lowered: the air will follow the path of lowest resistance[4]. This is why nasalized fricatives like [ṽ][5] in table (1.4) are so rare in the languages of the world; to make one, you’ll have to come up with a very precise setting of your tongue blade, with different muscle tensions and positions from normal fricatives. Again, the perceptual specification determines the articulatory gestures.

If two articulations produce the same sound, the easier one is more likely to be used. At most places of articulation, a complete closure is easier to make than a critical closure, because it involves a ballistic instead of a controlled movement (Hardcastle 1976). For labiodentals, even a ballistic movement often results in an incomplete closure; so, labiodental plosives are very rare, but labiodental nasals quite common. Every non-labiodental nasal forms a natural class with its corresponding plosive because both are implemented with the same ballistic articulatory gesture, e.g., [complete labial closure].

[4] You can check this by pinching your nose, making a “nasal” [z], and then suddenly releasing your nose.

[5] If we take a perceptual definition for [ṽ]. The IPA is a hybrid notation system, and often ambiguous: if [i] and [u] are vowels with minimal F_1, what does the IPA symbol [y] mean? Is it a front rounded vowel with minimal F_1, or a vowel with the tongue shape of [i] and the lip shape of [u]?

Plosives. The intervocalic plosive in [ata] is perceptually marked by a sequence of formant transition [[t˺]] + silence [[_]] + release burst [[t]] + formant transition. Their has been a gigantic literature about the importance of all these cues in the perception of speech. While the formant transitions are shared with most other consonants at the same place of articulation, the silence and the burst together signal the presence of a voiceless plosive. In [[thɛɛ̃n_ts]], both release bursts are heard, but silence associated with the first [t] merges with the ambient stillness, thus giving up its identity. A cluster of plosives, like /atpa/, is pronounced with overlapping gestures in most languages (with French as a notable exception), so that the result [[at˺_ːpa]] shows the demise of the main place cue for the recognition of [coronal]. In English, this may lead to place assimilation ([[ap˺_ːpa]]), because the articulatory gain of not having to perform a blade gesture outweighs the perceptual loss of losing the remaining place cue. We will see (ch. 11, 16) that this kind of phonetic detail can be expressed directly in the grammar of spreading phenomena.

Duration. Duration could be called a *derived* perceptual feature, because the perception of duration presupposes the recognition of another feature (the presence of sound, timbre) as being constant. In the above example of place assimilation, the duration of the silence was preserved, which is a sign of the independence of the silence cue for plosives.

Vowel height. According to Kenstowicz (1994, p. 20), “we may interpret [+high] as the instruction the brain sends to the vocal apparatus to raise the tongue body above the neutral point”. However, since different tongue muscles are involved in [i] and [u], such a standpoint testifies to a view that speech is organized very differently from other motor activities: no proprioceptors for non-low tongue height are known; the correlation of vowel height with jaw height is weak, regarding the highly varying strategies that speakers adopt to implement this feature (Ladefoged 1990a). Therefore, with Ladefoged (1971, 1990a) and Lindau (1975), I will assume that vowel height inversely corresponds to the first formant (F_1), i.e., that the phonological effects of vowel height correspond to the perception of the first peak in the excitation pattern of the basilar membrane in the inner ear (the higher the vowel, the lower its F_1). Simplistically, the muscles used in implementing vowel height are roughly: genioglossus (higher front vowels), styloglossus (higher back vowels), and hyoglossus (low vowels).

Vowel height does define natural classes in inventories and rule targets (as a result of perceptual categorization, see ch. 8), but vowel harmonies and assimilations are largely confined to the more articulatorily tractable features of rounding, backness, and ***advanced tongue root***; the rule ɔ → o / _ i is relatively rare (as compared with ɔ → ø / _ i), and assimilation of vowel height is expected to occur only if all the vowels involved use the same articulator, as in ɛ → e / _ i. Apparent exceptions are treated in chapter 16.

Tensions. A direct relation between articulation and perception is found in the tension of the vocal cords, which is the main determiner of the pitch of voiced sounds. The tension of the lung walls determines the subglottal pressure, which influences the loudness (spectral slope and intensity) and pitch of the perceived sound. A rather indirect relation between articulation and perception is found with the tension of the walls of the pharynx and the cheeks, which can play a role in the voicing of obstruents.

Place. The perceptual distinction between the various places of articulation is primarily made on the basis of the associated auditory spectra. For vowels, the first formant, which is in the lower part of the spectrum and represents the degree of closure, seems to be an independent perceptual feature; it disappears in the transitions to neighbouring obstruents. Thus, place information for vowels is restricted to the upper part of the spectrum, and we can imagine that it is a multi-valued perceptual feature, encompassing [front], [back], and [round]; all these colour features assume [sonorant]. In the auditory spectrum, the front-back distinction is represented by the ***second formant*** (F_2); I will take it to specify the *strongest* spectral peak above the first formant.[6] Specifying the value “max” for F_2 means that F_2 should be at a maximum, given a fixed value of F_1; this is most faithfully rendered by producing a front vowel with lip spreading. The value “min” specifies a minimum value of F_2 given F_1; this is most faithfully implemented as a rounded back vowel. No “enhancement” of an allegedly distinctive feature [back] by an allegedly redundant feature [round], as proposed by Stevens, Keyser & Kawasaki (1986) for reasons of lexical minimality, is implied here: the two gestures just implement the same perceptual feature symmetrically.

For consonants, place cues can be found in the formant transitions from and to neighbouring sounds. Other cues must be found in noises (fricatives and release bursts). The perceptual place feature is a rather continuous path through a multidimensional space, ranging from [bilabial] to [glottal], and does not respect the discrete articulatory distinctions between the articulators: labiodental and corono-dental fricatives sound quite similar, and so do corono-postalveolars and dorso-palatals; perceptually, [glottal] must be included in the set of values of the [place] feature (adjacent to [epiglottal]), though it shows no formant transitions to surrounding vowels because these have glottal constrictions, too. For nasals, the place information contained in the oral side branch is very weak: an isolated nasal stop produced with simultaneous lip and blade closures will sound as [n] in the dark, and as [m] if the listener sees the speaker: the visual cue overrides the auditory cue. Release cues without noise occur for nasal stops and laterals.[7]

Vocalic place cues can be used with stops and fricatives to a certain extent: in many languages, lip rounding contributes to the perceptual contrast between [s] and [ʃ]. By contrast, lip rounding does not influence at all the stationary part of the sound of [n].[8]

[6] Known in the phonetic literature as $F_2{}'$, the usual definition of F_2 being: the *second* spectral peak, measured from 0 Hz upwards. This peak is commonly determined by a computer program that is forced to find five peaks between 0 and 5000 Hz. For [i], this second peak (at 2500 Hz or so) usually incurs a much weaker impression on the inner ear than the third and fourth peaks, which tend to conspire to build a very strong perceptual peak near 4000 Hz. See ch. 4 for an alternative proposal.

[7] You can hear these cues if you record [ana] or [ala], create a backward copy of this sound, and compare the two CV transitions.

[8] Try saying [n::] and superpose the lip movements of [wiwiwi]. The colour does not change. An analogous experiment with [ŋ::] and [wawawa] shows velar excitation of a closed front cavity.

1.2.6 The speech-neutral position and privative features

Some features must be considered ***privative*** (mono-valued, unary), because only a single value can be phonologically active (Anderson & Ewen 1987, Ewen & Van der Hulst 1987, Van der Hulst 1988, 1989, Avery & Rice 1989). For instance, only [+nasal] is thought to be able to spread.

Steriade (1995) provides an articulatory explanation for the existence of privative features. The presence of an articulatory gesture like [lowered velum], she argues, is qualitatively different from its absence, because it constitutes a deviation from the speech-neutral position (Chomsky & Halle 1968, p. 300).

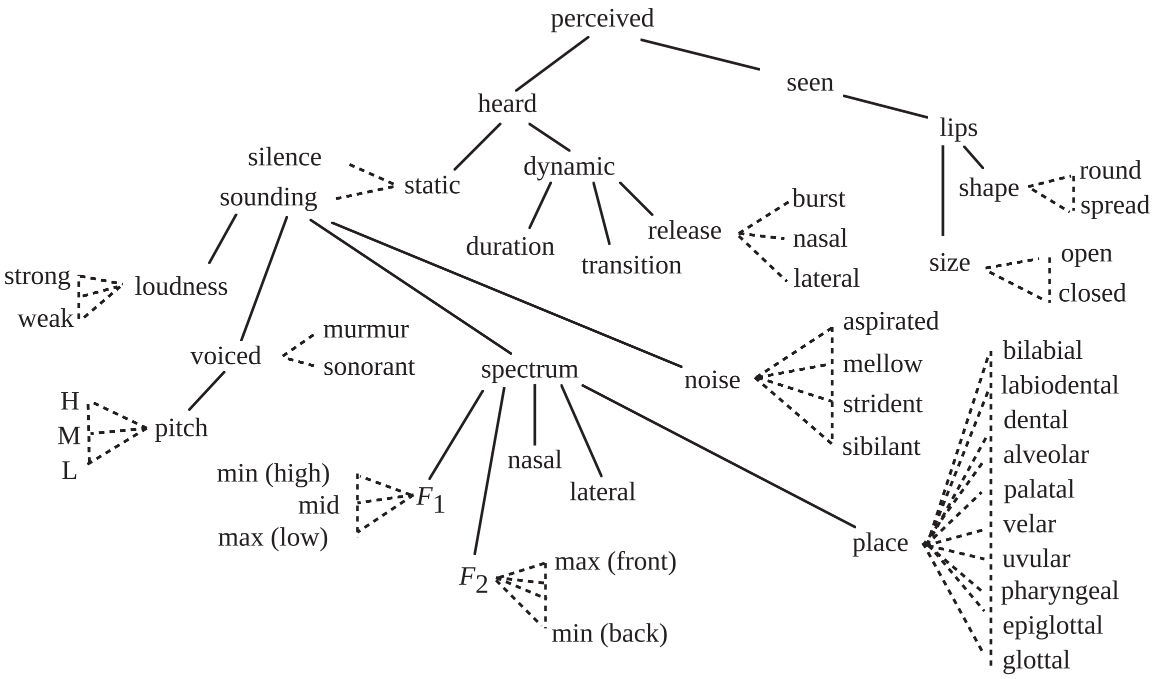
The only real neutral position is the one in which most muscles are relaxed, namely, the neutral position for breathing, which involves a wide glottis and a lowered velum. The alleged speech-neutral position would have glottal adduction and a raised velum, which involve active muscular effort (interarytenoid and levator palatini).

This speech-neutral position can only be explained with reference to requirements of perceptual contrast: we can produce better spectral contrasts for non-nasals than for nasals, and voicing allows us to produce tone contrasts, better formant structures, and louder sounds. Thus, nasal sounds will occur less often in an utterance than non-nasal sounds, and voiceless sounds will occur less often than voiced sounds. Instead of a *neutral* position, we now have the *most common* position.

So, instead of invoking a mysterious speech-neutral position, it seems more appropriate to explain privativity directly by arguments that start from the frequency of occurrence of the feature values in the average utterance: the presence of a perceptual feature like [nasal] is quantitatively different from its absence, because the latter would not signal any deviation from the more common non-nasality. In §9.5, I will show that differences in the phonological activities of various articulatory gestures can be related directly to the listener’s adaptation of recognition strategies to frequency differences in the corresponding perceptual features. I will argue there and in chapter 13 that the common values like [–nasal] are not absent, but only relatively invisible because of their weak specifications.

1.2.7 Feature geometries

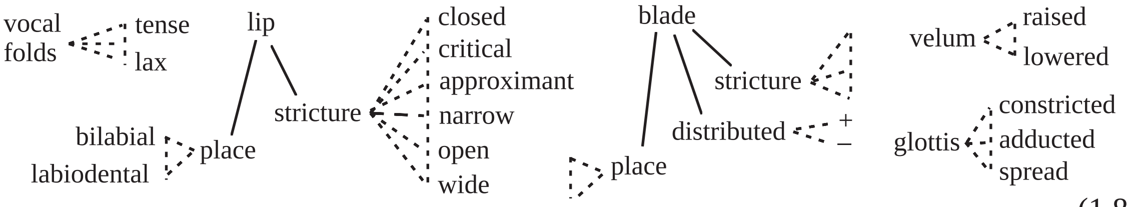
The above story gives rise to the following partial geometry of implications for the presence of perceptual features; conjunctions are shown by “vertical” (solid) branches, disjunctions by “horizontal” (stippled) branches:



(1.7)

This figure only shows perceptual dependencies, so it does not show which features cannot co-occur because of articulatory constraints; for instance, an aspirated sonorant is easy ([ɦ]), but a sibilant sonorant would be much harder to produce. Some of the implications have to be taken with a grain of salt, as it is not unthinkable that pitch is perceived on voiceless syllables (as in Japanese), etc.

The implicational geometry for articulatory gestures is extremely flat, because of the near independence of the articulators; as in (1.7), many features have values along a continuous range:



(1.8)

The picture that arises from these geometries is rather different from the hybrid feature geometries that have been proposed by Clements (1985), Sagey (1986), McCarthy (1988), and Keyser & Stevens (1994). Those geometries will be seen to result from a confusion of the roles of articulatory and perceptual features (ch. 19).

1.2.8 Conclusion

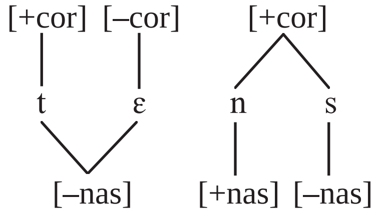
As the examples show, the relations of the traditional hybrid features with their supposed articulatory and acoustic correlates are rather vague. Every instance of asymmetry between articulatory and perceptual features causes problems to theories that do not distinguish them. Therefore, now that phonological theories have gotten rid of the early generative segmentality, binarity, representations, grammar organization, and rule ordering, the time has come to replace the content of the features with concepts rooted in general properties of human motor behaviour and perception.

1.3 Hybrid, articulatory, and perceptual representations

The purpose of linguistic proposals for phonological representations is the efficient description of phonological structures and processes. Derived from the evidence of language data, the usual phonological representation of an utterance is a hybrid of articulatory and perceptual specifications.

1.3.1 Hybrid representations

If we return to the English word *tense*, we see that linear phonology (Chomsky & Halle 1968) described it as a sequence of four bundles of binary features, called ***segments***: /t+ɛ+n+s/. The ***autosegmental*** approach (Leben 1973, Goldsmith 1976) stressed the autonomy of the various features:

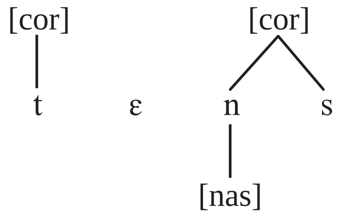


(1.9)

This would seem phonetically more satisfying, as it reflects the independence of the articulators and heeds two other principles that can be seen as consistent with articulatory phonetics: the ***Obligatory Contour Principle*** (OCP: “adjacent identical autosegments are forbidden”) ensures that the single coronal gesture of /ns/ is represented as a single feature value, and the ***No-Crossing Constraint*** (NCC: “association lines do not cross on the same plane”) ensures that the two successive coronal gestures of /t/ and /ns/ are represented as two separate feature values.

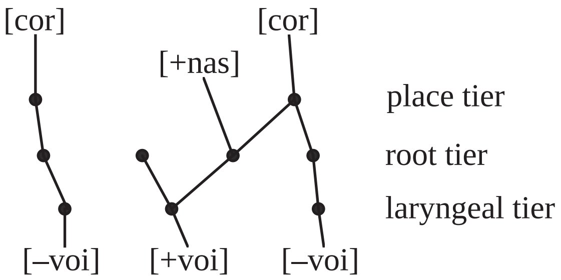
Important predictions of these representational constraints are that phonological processes cannot change two non-adjacent identical elements at a time, and that they cannot change only a single element out of a sequence of two adjacent identical elements. Thus, they allow only a limited range of primitive phonological processes, like ***delinking*** and ***spreading***. From the functional point of view, these processes are advantageous if delinking is seen as the deletion of an articulatory gesture, and spreading as the change in the timing of an articulatory gesture, often in order to compensate for the loss of another gesture; for instance, in the common process of place-assimilation of nasals (/n+b/ → [mb]), the coronal gesture is deleted, and the labial gesture is extended in such a way that the nasal still has consonantal perceptual properties. However, this interplay between articulatory and perceptual needs could not be expressed in autosegmental phonology, because articulatory features like [closed tongue blade] could not be distinguished from perceptual features like [consonantal].

The advent of theories of privative features (§1.2.6), whose presence is qualitatively different from its absence, brought phonology again somewhat closer to function. In the interpretation of Archangeli & Pulleyblank (1994), the representation of /tɛns/ is[9]



(1.10)

Theories of Feature Geometry (Clements 1985, Sagey 1986, McCarthy 1988) subsumed the features [labial], [coronal], and [dorsal] under the [place] node, the features [voiced], [spread glottis], and [constricted glottis] under the [laryngeal] node, and all features together under the ROOT NODE. For instance, a partial representation of /tɛns/ along the lines of Archangeli & Pulleyblank (1994) would be



(1.11)

Articulatory detail was put under the relevant articulator node: the [coronal] node dominates the feature [±anterior], and the [labial] node dominates [±labiodental]. The idea of this implicational interpretation of feature geometry is that if a node spreads, the dependent features also spread; for instance, place assimilation of /n+f/ can only give /ɱf/, never /mf/, because [labial] cannot spread without its dependent [labiodental].

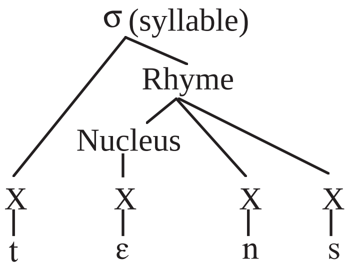
Directly under the root node are those features that we would associate with independent articulatory tiers, for instance, [nasal]. The features that do not spread, except if the whole segment spreads, can be seen as part of the root node. These ***major class features***, it will come as no surprise, are exactly the perceptual features [sonorant] and [consonantal].

The remaining traditional feature [continuant] causes the greatest problems. If it is associated with the stop/fricative distinction, it should be dependent on each articulator tier, and, indeed, we see that clicks in Nama (Beach 1938) can have separate specifications for continuancy on their coronal and dorsal articulators. A reason *not* to put the feature [continuant] there is the fact that continuancy does not necessarily spread if the articulator spreads.

In chapter 19, I will show that only implicational hierarchies as in (1.7) and (1.8) can be maintained, and that the place node and the problems with [continuant] are illusions caused by the interaction of more fundamental perceptual and articulatory phenomena.

[9] The interpretation of the NCC and OCP implicit in (1.10) is the only alternative that stays compatible with the gestural analogy. It makes it hard to describe long-distance anti-repetition phenomena as OCP effects, but this is actually an advantage, as shown in chapter 18.

Finally, theories of metrical phonology (Clements & Keyser 1983, Hyman 1985, McCarthy & Prince 1986, Hayes 1989) would propose hierarchical structures like (after Blevins 1995):

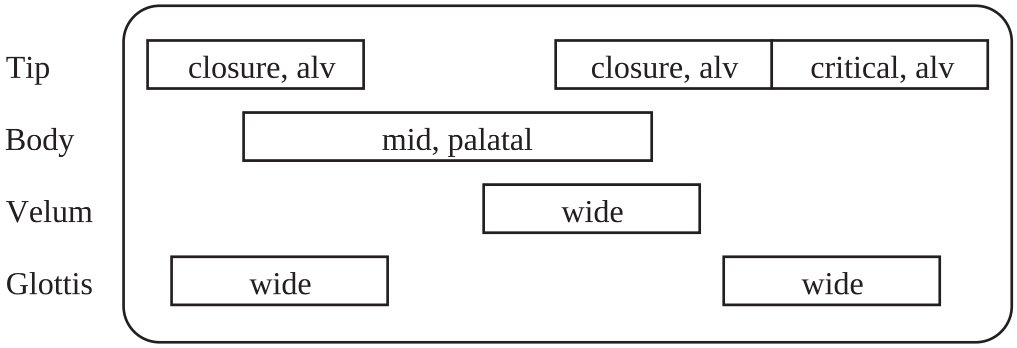


(1.12)

In this work on Functional Phonology, I will not touch metrical phenomena like accent, stress, and rhythm, because these have no obvious functional correlates in the speech-production and perception systems other than purely organizational principles: if we want to know what those principles are, we can only look at how languages handle them, and the current bottom-up approach, which starts from physiological principles, seems impossible.

1.3.2 Articulatory phonology

An interesting attempt to get at least one of the representations right, is Articulatory Phonology (Browman & Goldstein 1984, 1986, 1989, 1990a, 1990b, 1992, 1993): each articulator has its own tier, and the ***gestural score*** is a representation of the values on all relevant tiers. For instance, Bird & Klein (1990) give the following gestural score for the English word /tɛns/:



(1.13)

This representation shows the three overlaps between the four consecutive segments: the glottal widening, needed to make the stop voiceless, is continued after the release of the stop, giving the result of aspiration or a voiceless vowel; the lowering of the velum before the closing of the tongue tip causes nasalization of the preceding vowel; and the raising of the velum before the lowering of the tongue tip, which is needed to create the conditions for sibilant noise, causes an intrusive stop (silence + burst) to appear between /n/ and /s/ (Sievers 1876: 141; Fourakis & Port 1986; Clements 1987).[10]

[10] In this book, I will only talk about the variant without phonologization of the plosive, i.e. without the glottal stop that English interposes between a short vowel and a following voiceless plosive.

In Articulatory Phonology, the values on the tiers represent immediate articulatory specifications only: these are the proximal targets of speech production and implement the forward path that we saw in the top left of figure 1.1, typical of skilled motor behaviour. But the auditory system will monitor the acoustic result, and the speaker/listener will assess the faithfulness of the perceptual result to the original perceptual specification: between the stretches of gestural specification in (1.13), for instance, the articulators return to their neutral positions, but the freedom of the articulators to go anywhere depends on the local perceptual specification of this utterance.

As a theory of phonology, therefore, Articulatory Phonology neglects the organizing power of perceptual invariance and segmental linearization. The solution to this problem involves a radical discrimination between the underlying perceptual specification, candidate articulatory implementations, and perceptual surface representations.

1.3.3 The specification – articulation – perception triad

All the representations that we saw in §1.3.1 were proposed on the basis of studies of phonological structures and processes: the top-down approach. In this paper, I will use the bottom-up approach: to derive what languages could look like, starting from the capabilities of the human speech-production and perception system.

When turning a set of functional explanations into a theory of phonology, the first step is to posit the existence of ***underlying forms***. In perceptuomotor terms: the intended effects of one’s movements on the environment. In speech terms: specifications of how my utterances should sound. We can see in figure 1.1 why phonology is different from other parts of the grammar: as a control mechanism for motoric events, it contains a feedback loop, which compares the perceptual result of the utterance with its specification. My hypothesis is that all strata of our phonological system mirror this loop, although it can only actually be proven to apply to phonetic implementation. This approach allows various degrees of abstractness in underlying specifications at each stratum, and the output of each stratum will generally be different from its input.

Thus, I propose the following three representations within each stratum:

1. **Specification:**
   The underlying form (input), specified in perceptual features.
2. **Articulation:**
   A candidate implementation, expressed on articulatory tiers.
3. **Perception:**
   The surface form (output), expressed in perceptual features.

As an example, we show a fairly complete (“phonetic”) specification for /tɛns/ (the symbols /t/ etc. are nothing more than mnemonic symbols for bundles of feature specifications, reminding us of the predominant segmentality of English phonology;):

Specify:	/t/	/ɛ/	/n/	/s/
timing	C or X	V, X, or µ	C, X, or µ	C, X, or µ
coronal	burst		+	
voice		sonorant	sonorant	
noise	aspirated			sibilant
F1		open mid		
F2		max		
round				
nasal			+	

(1.14)

where ‘C’ stands for ‘consonant’, ‘V’ for ‘vowel’, ‘X’ for ‘timing slot’, and ‘µ’ for ‘mora’. This specification contains exclusively perceptual features, whose content was discussed in §1.2.5. The criterion for entering a specification in this table is the answer to the question whether the value of that feature matters for the recognition of the utterance as more or less representing the English word /tɛns/: only the values that seem to matter most, are visible in (1.14). The formalization of the verb *matter* and the adverbial phrase *more or less* will be presented in §10.1.

Besides the values of perceptual features, the table also specifies relations of simultaneity and precedence between the features. Thus: there is an “open mid” specification somewhere; the *first* segment is specified as voiceless (simultaneity relation between C and [voiceless]); there is a link between voicelessness and sibilancy; aspiration precedes voicing; a V precedes [nasal]. The specification also implicitly tells us what should *not* be there: no labial burst (because there is no labial specification), no voiced sibilancy (because these features are not simultaneous); no nasality during the vowel (because the privative feature [nasal] is not specified for the vowel).

The usual articulatory implementation of /tɛns/ in English and its perceptual result are as follows:

Articulate:

<table>
<tr><td>tip</td><td>closed</td><td colspan="3">open</td><td colspan="3">closed</td><td>critical</td></tr>
<tr><td>body</td><td colspan="8">open</td></tr>
<tr><td>velum</td><td colspan="3">closed</td><td colspan="2">open</td><td colspan="3">closed</td></tr>
<tr><td>glottis</td><td colspan="2">wide</td><td colspan="4">narrow</td><td colspan="2">wide</td></tr>
<tr><td>lips</td><td colspan="8">spread</td></tr>
</table>

Perceive:

<table>
<tr><td>silence</td><td>+</td><td colspan="4"></td><td colspan="2">+</td><td></td></tr>
<tr><td>coronal</td><td></td><td colspan="3">bu. tr.</td><td>side</td><td colspan="2"></td><td>bu. cont</td></tr>
<tr><td>voice</td><td colspan="2"></td><td colspan="3">sonorant</td><td colspan="3"></td></tr>
<tr><td>noise</td><td></td><td>asp</td><td colspan="5"></td><td>sibilant</td></tr>
<tr><td>F1</td><td></td><td colspan="3">open mid</td><td colspan="4"></td></tr>
<tr><td>F2</td><td></td><td colspan="3">max</td><td colspan="4"></td></tr>
<tr><td>rounding</td><td colspan="8"></td></tr>
<tr><td>nasal</td><td colspan="3"></td><td colspan="2">+</td><td colspan="3"></td></tr>
<tr><td></td><td>_</td><td>t h</td><td>ɛ</td><td>ɛ̃</td><td>n</td><td colspan="2">_</td><td>t s</td></tr>
</table>

(1.15)

Articulation. In the articulatory representation, time runs from left to right on each tier, and the tiers are time-aligned; thus, there are no simultaneous articulatory contours in this example. The specification on each tier is complete, for consonants as well as for vowels.

From all possible articulations that implement /tɛns/, table (1.15) shows the one that involves the fewest contours. The openness of the tongue body and the spreading of the lips are only needed for giving the correct vowel height during /ɛ/. During the other parts of the utterance, these shapes may remain the same, since they would not interfere with the perceptual invariants of /t/, /n/, and /s/; here, a less spread lip shape would give almost the same perceived utterance, though a *complete* labial closure must be forbidden. In reality, lip spreading is achieved during the closure of /t/, and undone during /n/ or /s/; this is related to the fact that the active maintenance of lip spreading costs more energy than keeping the lips in a neutral position. Thus, there is a conflict between two aspects of laziness: minimization of number of contours and minimization of energy (for the formalization of this conflict, see §7.2).

Perception. In the representation of the uncategorized (“acoustic”) perceptual result, time runs from left to right on each tier, and the tiers are time-aligned with each other and with the articulatory tiers above. If a feature has no value, no value is shown (see the noise tier); for some binary features, only positive values are shown, suggesting privativity (§9.9). In the perceptual score, many features are specific to either the consonantal or the vocalic class of sounds, in line with the implications shown in (1.7).

A complete (i.e., intervocalic) plosive is represented as a sequence of (pre-consonantal) transition (tr), silence, and release burst (bu). On the coronal tier, [side] means the acoustical correlate of the oral side branch with a coronal closure (barely distinguishable from other oral closures), and [cont] means a continuant coronal sound.

Microscopic transcription. Though the set of perceptual tiers is the ultimate surface representation of the utterance, a linear transcription would be more readable. Because all phonetic details will be involved in assessing the faithfulness relations between specification and output, such a transcription should be very narrow. Instead of a traditional narrow transcription like [tʰɛ̃nᵗs], we shall use a transcription that introduces a new symbol in the string every time that any perceptual feature changes its value. For instance, the coronal gesture in /ata/ will normally be heard as transition + silence + burst; this will give [[at˺_ta]] in a ***microscopic transcription***:

- A transition is denoted in microscopic phonetic notation as an unreleased stop: [t˺].
- Silence is denoted by an underscore: [_].
- A release burst is denoted by the symbol for the stop itself: [t].

Thus, a readable shorthand for the perceptual result is [[thɛɛ̃n_ts]]. The [h] part could equally well be transcribed as a voiceless vowel [ɛ̥].

1.4 Formalization of functional principles

We see that the specification |tɛns| (1.14) and the perceptual result [[thɛɛ̃n_ts]] (1.15) are different: there are several aspects of ***unfaithfulness*** of the perceptual result to the specification. These differences arise through properties of the speech-production system, and their interactions with properties of the speech-perception system. The properties and their interactions will be formalized in part II: functional principles can be expressed explicitly as ***gestural constraints*** that evaluate articulations, and as ***faithfulness constraints*** that evaluate specification-perception correspondences.

At this point, the phonologically oriented reader may jump to chapter 6. In chapters 2 to 5, I will describe a physical-mathematical model of the “automatic” relations between articulation and acoustics.

2 Articulation model[1]

Abstract. This chapter describes a model which represents the entire vocal apparatus as a structure of tubes with moving walls. The model is capable of simulating more features of the interaction between myo-elastical and aerodynamical properties, than any previous model.

While we are investigating the relation between articulatory and perceptual features of speech sounds, it would be advantageous to have an articulatory-acoustic model that could produce almost any speech utterance. As existing models had too few capabilities to cope with all the physical phenomena that are used in speech, I designed a comprehensive model of the speech-production apparatus, including lungs, glottis, and vocal and nasal tracts (Boersma 1991, 1993a, 1995). The coming chapters describe the construction of this ***articulatory synthesizer*** in detail. Chapter 2 shows how, starting from the activities of the main muscles involved, the model computes the target positions and tensions of the articulators. Chapter 3 shows how the realized positions and motions of these structures are computed numerically, simultaneously with the acoustic output. Chapter 5 puts the model to the test, showing that it can faithfully simulate many physical speech phenomena. In parts II and III, the articulation model will be used, together with the auditory model of chapter 4, to support explanations of cross-linguistic tendencies in phonetic implementation, sound structures, and autosegmental phonology. The reader who prefers to dive into phonological problems directly, can jump ahead to chapter 6; knowledge of chapters 2 to 5 is not needed for understanding parts II and III.

2.1 Requirements

The vocal tract can be viewed as a structure of ducts (channels that contain air). Human speech uses the following structural properties of these ducts:

- some of the ducts are open to the atmosphere at one end (lips, nostrils);
- some ducts are closed at one end (diaphragm);
- some ducts branch into two others (pharynx - mouth - nose).

Moreover, human speech takes advantage of the following physical properties of these ducts:

- the walls of the ducts yield to air pressure and are able to vibrate passively under the right circumstances (vocal folds, uvula, tongue tip);
- noise is generated wherever turbulent conditions arise (fricatives, release bursts);
- the lengths of some ducts vary in time (in lip rounding, ejectives, implosives, tongue position).

[1] This chapter is a longer version of the first halves of Boersma (1991) and Boersma (1993a).

To be acceptable as a tool for research on articulation-perception relations, the model should be able to produce almost any speech utterance. It achieves this by being:

a. **comprehensive**: all the regions of the vocal apparatus (lungs, glottis and vocal tract proper) are treated as consisting of air-filled cavities with walls that can be seen as adjustable mass-spring systems;
b. **principled**: the acoustic output is computed from basic physical laws, without some of the approximations usually found in the literature.

Several choices have to be made with regard to the specification of the input to the model, the modelling of the articulators, and the generation of the acoustic output. These will be addressed in the rest of this section.

2.1.1 Specification of the input: muscle activities or tasks?

Because the positions and shapes of the articulators are the immediate targets of speech production (§1.2.2), the ***tasks*** of the task-dynamic approach of Kelso, Saltzman, & Tuller (1986) and Browman & Goldstein (1986, 1990) would be appropriate as input to an articulatory synthesizer in a result-oriented application like a text-to-speech system. *Our* purpose, on the other hand, is to investigate the acoustic consequences of articulatory activities. Therefore, we should be able to view all the relevant muscles as independently controllable, and coordination should be *acquired*: a model that can predict anything about sound systems should be able to describe the interplay between articulatory implementations and perceptual specifications from the standpoint of the language learner who has not yet built in coordinative articulatory tasks. Rather, these tasks should follow from that model. Therefore, the input to our model should be ***muscle activities***, not tasks: the articulatory input specifications initially control the lengths and tensions of the muscles, not the positions of the articulators.

In some places, I will simplify a synergistic group of functionally related muscles and replace it by one articulatory parameter. Every articulatory parameter can thus be said to represent an articulatory degree of freedom. I will not go as far, however, as Perrier, Lœvenbruck & Payan (1996), who minimize the degrees of freedom inside the vocal tract to seven.

2.1.2 Controlling the muscles

We model the muscles (and, therefore, the walls of the ducts) as mass-spring systems. As the input to the model is formed by the activities of these muscles, we will have to decide which of the properties of the muscles are the variables controlled directly by the activities. One candidate is the ***equilibrium length*** of a muscle: the myotatic reflex loop (fig. 1.1; §1.2.1) is thought to be capable of keeping the muscle at a constant length, independent of its load, and muscle spindles are found in many places in the vocal tract (Gentil 1990), so we will control muscle length in most cases.

On the other hand, the ***stiffness*** of a muscle also changes with activity: a contracting muscle becomes harder to stretch, so stiffness control is advocated by Perrier, Abry & Keller (1989). However, even if the only controlled variable of each muscle is its rest position, the walls of the tubes still consist of muscles with fibers tangential to the walls, and if these muscles are stretched by external forces, the force (but not the stiffness) inside the muscles will increase, which causes the normal (perpendicular) tension (stiffness) of the wall to increase. Thus, even if all muscles are modelled with constant stiffness, some mass-spring systems must be modelled as stiffness-controlled. The same stiffness is therefore responsible for the velocity with which the equilibrium position is restored.

2.1.3 Smooth trajectories

A desirable property of any articulator model is the smoothness with which the articulator should approach its target position. According to Nelson (1983), the trajectory taken can be imagined to minimize duration, force, maximum velocity, energy, or jerk, while Kawato, Maeda, Uno & Suzuki (1990) propose minimization of torque change.

Coker (1968) modelled the smoothness of the trajectory in an ad-hoc way using “simple low-pass filters”. Kelso, Saltzman, & Tuller (1986) modelled the *tasks* as mass-spring systems and the *muscles* as instantaneous followers of the tasks; exactly why such a *physiological* control mechanism should show smooth behaviour, other than for the sake of a *physically* realistic outcome, remains unclear. A similar ad-hoc approach is advocated by Browman & Goldstein (1984), who use sine shapes to interpolate gestures. By contrast, my approach of modelling not the tasks, but the walls of the cavities, as mass-spring systems, has the advantage that target positions may change instantaneously; no ad-hoc constraints on the smoothness of tasks or stiffnesses is necessary. Because of the inertia of the walls, a smooth trajectory will still result; with a critically damped spring the wall will typically approach its target position exponentially.

Finally, Perrier, Abry & Keller (1989) model the stiffness as a sinusoidally varying function of time, because a natural trajectory (there and back again) resembles a sinusoid. However, that is the trajectory characteristic of an *undamped* mass-spring system. Under the paradigm of stiffness control, Perrier, Abry & Keller (1988) maintain that modelling a muscle as a single (“lumped”) damped mass-spring system gives poor results (with overshoot and time-direction asymmetry), and prefer an approach with distributed springs, in which they control articulator position by varying the stiffnesses of the muscles that pull the relevant articulator; i.e., the ***equilibrium point*** of the articulator is determined by the relative stiffnesses of the muscles involved, and the speed with which this position is arrived at is determined by the average stiffness of these muscles. This model was later defended in Perrier, Ostry & Laboissière (1996) and Perrier, Lœvenbruck & Payan (1996), implemented for the tongue by Perkell (e.g., 1996), and is used in current vowel research (Payan & Perrier 1996). However, Perrier et al.’s criticism would not hold for length-controlled systems; e.g., if the mass-spring system is at least critically damped, there will be no overshoot, and the system is invariant under time reversal.

2.1.4 Aerodynamic-myoelastic interaction

Modelling the muscles (and not the tasks) as mass-spring systems, allows as to take into account the influence of the air pressure on the walls in a natural way.

2.1.5 Timing

Fowler (1980) argues that the timing of articulatory gestures is not extrinsically controlled by things like syllable boundaries or incompatible articulatory specifications, but instead is an integral part of the mental specification of the motor plan for each segment. The present articulation model is indifferent to the distinction between the two positions, because timing strategies must reside on a higher level of abstraction than implemented here: we just tell each muscle when to contract. Intrinsic timing, if it exists, must be a property of coordination, i.e., it would reside at the articulatory task level, which we do not model. The theory of functional phonology, however, is *not* indifferent to timing models: in §19.1.8, we will see that the dichotomy between models of "feature spreading" (an example of extrinsic timing) and "coproduction" (intrinsic timing), both of which are supported by the data, actually corresponds to the dominance of different functional constraints: linguistically, it is the compatibility with *perceptual*, not articulatory, specifications that can extrinsically influence the timing of articulatory gestures (§1.2.2).

2.1.6 Generating acoustic output

Most speech synthesizers rely on direct *acoustic* synthesis. These synthesizers are designed to produce understandable and natural-sounding output in a text-to-speech system, and are not suited for investigating articulatory-perceptual relationships. Many *articulatory* synthesizers separate the vocal tract into a source and a filter part, that function relatively independently of each other. With these synthesizers, we could reliably model the interaction between articulation and some spectral properties of vowels. However, many vocal-tract properties that are used in languages for contrasts between consonants, cannot be modelled.

The algorithm most widely found is the reflection-type line-analog synthesizer of Kelly & Lochbaum (1962); it was reimplemented by Rubin, Baer, & Mermelstein (1981), Allwood & Scully (1981), Liljencrants (1985), Kröger (1990), and Scully, Castelli, Brearley, & Shirt (1992), and is currently used in an articulatory synthesizer (Rubin, Saltzman, Goldstein, McGowan, Tiede, & Browman 1996) which is used in research on the relation between articulation and perception (McGowan 1994). Though Liljencrants (1985) adds a large number of physical phenomena as perturbations to the original line analog, tube lengths that vary in space and time can still not be modelled. Another algorithm is by Maeda (1982, 1988, 1990); this is used in current research on "speech mapping" (Abry & Badin 1996), vowel systems (Boë, Schwartz, Laboissière, & Vallée 1996), iterative acoustics-to-articulation inversion (Laboissière & Galvan 1995; Perrier, Lœvenbruck & Payan 1996), and a learning model (Bailly 1997). Though Maeda's model does use varying tube lengths, it still does not model walls that yield to the aerodynamics, other than for purposes of computing a source of damping. This means that both Maeda (1982) and Liljencrants (1985) had to leave out the glottis and vocal folds from their model; instead, a voice source is posited separately. It will be clear that these methods have trouble simulating myoelastic-aerodynamic interactions such as those between the vibrating vocal folds and the resonating vocal tract, and this is why the 'current research' mentioned above focuses on the articulatory-acoustic relationships of *vowels*.

The drawbacks of the simple methods mentioned above were well-known to the people who designed them, but the rationale was:

> "Although there are synthesizers which have more sophisticated and realistic models of the acoustic sources and of the area function to sound transformation (e.g., Flanagan *et al*., 1975; Flanagan *et al*., 1980), these systems are too computationally inefficient to serve as interactive research tools on equipment which is generally available to most laboratories." (Rubin, Baer & Mermelstein,1981)

Similar considerations led to Sondhi & Schroeter's (1987) hybrid time-frequency-domain articulatory synthesizer. With the advent of faster computers, however, it is now time to take advantage of more sophisticated methods, like the one described in this book.

A model explicitly constructed with the purpose of investigating the interaction between the voice source and the vocal tract is the two-mass model of the vocal folds by Ishizaka & Flanagan (1972). However, they modelled the vocal tract very differently from the vocal folds, which undergo a rather special treatment, and though their myoelastic equations (for the vocal folds) contain an air-pressure term, their aerodynamic equations do not reflect wall movement correctly, not even when they later included a dynamic vocal tract (Flanagan, Ishizaka, & Shipley, 1975, 1980).

For our purposes, we need to combine the advantageous properties of all these models:

1. We should like to extend the two-mass model to include the entire wall of the vocal apparatus, without neglecting the 'pumping' brought about by moving walls (which Flanagan & Ishizaka (1977) stated to be negligible for the vocal folds, but which is surely a major phenomenon in obstruent consonants).
2. We require that the lengths of the tube sections are allowed to vary in space and time.
3. We require that all the walls are allowed to yield to changing air pressures.

We will see in chapter 3 that the mathematical problem can be solved numerically. Table 2.1 shows the availability of some properties desirable for simulating consonants, in several existing articulatory synthesizers. The model described in this book extends

Table 2.1 The inclusion of several desirable physical features in various synthesizers.

	Liljencrants (1985)	Maeda (1982)	Flanagan et al. (1975)	This book
space-varying tube lengths	no	yes	no	yes
time-varying tube lengths	no	yes	no	yes
glottis-tract interaction	no	no	yes	yes
pumping and sucking	no	no	no	yes
volume control (lungs)	no	no	no	yes
monopole noise source	yes	no	yes	yes
dipole noise source	no	no	no	no
boundary layer viscosity	yes			yes
Hagen-Poiseuille viscosity	yes			yes
air/wall heat conduction	apx	no	no	no
heat conduction in air	no	no	no	no

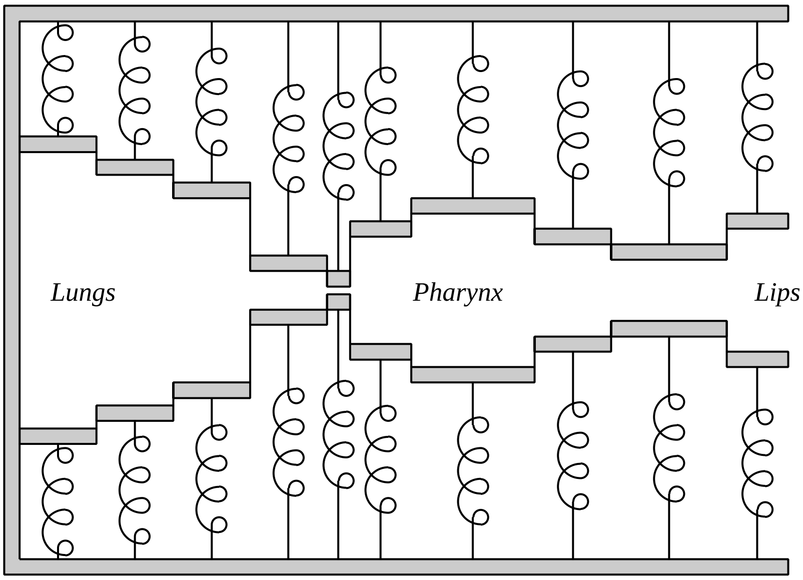


Fig. 2.1 Simplified mid-sagittal view of our model of the speech apparatus (not drawn to scale). The model features a sequence of 89 straight tubes with walls consisting of masses and springs. The leftmost of these tubes is closed at the diaphragm, the rightmost tubes form the openings between the lips (and between the nostrils, which are not shown) and are open to the atmosphere, where fluctuations in the airflow are radiated as sound. The glottis is represented by two tubes (shown as one here), which are treated exactly the same way as all other tubes. The speech muscles can alter the rest positions and the tensions of the springs. Some of the masses are connected with springs to their nearest neighbours. Not shown are: the coupling springs that connect masses to their neighbours; the springs and masses in the *z*-direction (perpendicular to the paper); the nasal tract.

Flanagan’s two-mass model of the vocal folds to include the entire speech apparatus, while treating the aerodynamics in a more principled, less ad-hoc, and more consistent way (the “apx” and “no” in the row labelled “air/wall heat conduction” are explained in §3.3).

2.2 Overview of the articulation model

Figure 2.1 shows a simplified picture of our model. As a model of the human vocal apparatus, it is a straightened approximation to the curved shapes of the *entire* speech apparatus: the pharyngeal, oral, and nasal tracts, the glottis, and the lungs. It consists of a sequence of straight tubes that contain air. Air is forced to flow into and out of these tubes as a result of its mass inertance and its elasticity. One source of acoustic output is derived from the ***airflow*** at tube boundaries that are open to the atmosphere (like the right boundary of the rightmost tube in figure 2.1): it is the sound radiated from the lips and nostrils into the atmosphere.

The walls of the tubes yield to ***pressure changes***. At the same time, the equilibrium positions of the walls can be adjusted by the articulatory muscles. The walls are, therefore, modelled as mass-spring systems. The tensions of some of these springs can be adjusted, too. This reflects the ability of the vocal folds to produce tone differences, and the ability of the supralaryngeal musculature to distinguish fortis and lenis obstruents. The second source of acoustic output is the sound radiated from the ***moving masses***.

The main source of energy in the tract is the variation of lung pressure. In some models, the lungs are modelled as an ideal pressure source. In our model, lung pressure is, more realistically, brought about by decreasing the lung volume, i.e., reducing the neutral width and/or length of the three leftmost tubes in figure 2.1 (as we will see in §2.6, the real model has more parts than that). The modelling of the respiratory mechanism as ***lung-volume control*** rather than as an ideal pressure source, expresses the fact that the lungs have a finite capacity. Furthermore, the subglottal formants will appear naturally in our model.

The walls of a tube, which are nearly parallel, can ***oscillate*** if they are close enough together and there is sufficient airflow along them. This follows automatically from the aerodynamic and myoelastic equations. Thus, the vocal folds can easily vibrate in this model. Nothing withholds other articulators, though, from vibrating as well; tongue tip, uvula and lips are likely candidates for producing trills.

If the particle velocity exceeds a certain threshold, ***noise*** is generated immediately downstream from the constriction; the portion of the kinetic energy that is converted into turbulence depends on the relative widths of both tubes involved.

The lengths of the tubes do not have to be equal. The upper part of the glottis, for instance, may be 1 mm thick, whereas in other regions, the tubes can be as long as 10 mm. More important, though, is the advantage of allowing the lengths of tubes to *vary with time*. This permits us to model appropriately the lengthening and shortening of certain tubes that is caused by lip rounding, dorsal constriction, or up and down movements of the larynx.

The articulatory synthesis is divided into two parts:

1. From muscle activities to tract parameters (chapter 2).
2. From tract parameters to sound (chapter 3).

For every moment in time, we compute as the output of step 1, the following tube parameters, which form the input to step 2, for each tube section:

- equilibrium (target) position (width) Δy_{eq}
- target length Δx_{eq}
- target “depth” (the third dimension) Δz_{eq}
- mass m
- linear and cubic spring constants $k^{(1)}$ and $k^{(3)}$
- relative damping factor $B_{open,rel}$ for the moving masses
- linear and cubic tissue stiffness constants $s^{(1)}$ and $s^{(3)}$ (during collision)
- relative damping factor $B_{closed,rel}$ (during collision)

An extra parameter of each tube is its number of ***parallel subdivisions*** (§2.3.7); this does not change over time.

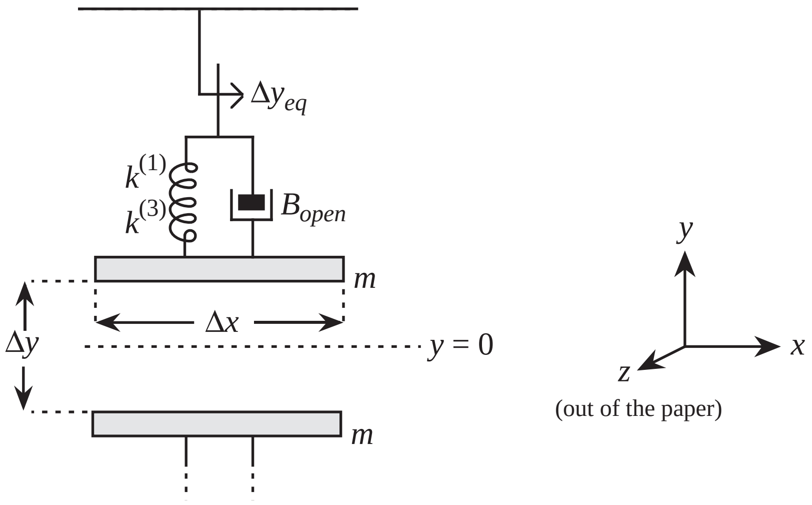


Fig. 2.2 Mid-sagittal view of one tube, showing one of its springs and both movable masses. The articulatory muscles can directly adjust the rest position y_{eq}, the linear spring constant $k^{(1)}$, and the tube length Δx, and may also indirectly vary the mass m, the damping B_{open}, and the cubic spring constant $k^{(3)}$. All these parameters, plus the air pressure inside the tube, determine the development of the state of the tube wall, which is represented by its displacement y and its velocity dy/dt.

2.3 The springs and the masses

Every tube is supposed to be enclosed along the y-axis in figure 2.1, by two opposing walls that consist of one mass and one spring each (see figure 2.2). For each tube, both masses and springs have identical properties.

2.3.1 Equation of motion

The acceleration of one wall in the y-direction is derived from the following equation, which gives the total force on this wall:

$$m\frac{d^2y}{dt^2} = \textit{tension force} + \textit{collision force} + \textit{damping force} + \textit{air pressure force} \qquad (2.1)$$

where m is the mass of either wall (in kg), and y is the displacement of the wall from the horizontal midline in figure 2.1 (in metres); thus, in our case of two opposing walls with equal properties, the distance between these walls (the width of the opening between the two masses) is $\Delta y = 2y$. The mass m need not be constant, because it is the part of the wall that actually moves; it could slowly vary in time as a function of the tension in the wall.

2.3.2 The tension force

The *tension force* (restoring spring force) is the force in the spring that tries to bring the wall to its neutral (equilibrium, rest) position. It is due to the tension of the muscles inside

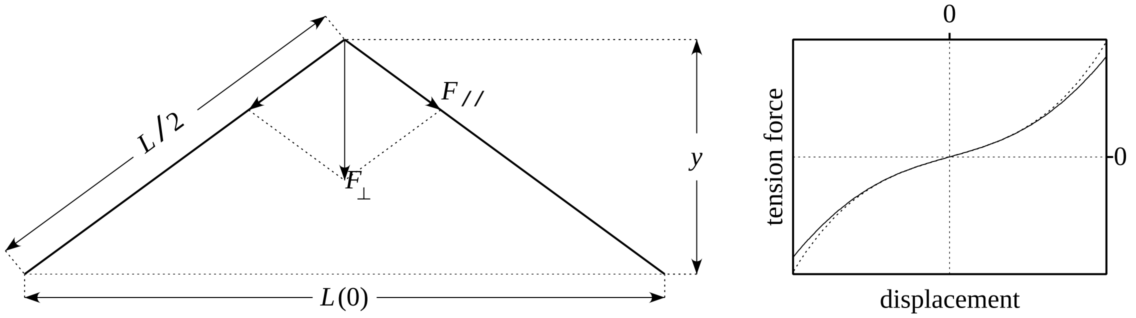


Fig. 2.3 The plucked string. On the left: the forces. On the right: the force as a function of displacement, for L_{eq} = 0.9 L_0 and a maximum displacement of ±0.3 L_0 (eq. 2.4). The dotted line represents its two-term approximation (also eq. 2.4).

the wall (e.g. vocalis muscle, pharyngeal constrictor muscles) and to the tension of the muscles that pull the edges of the wall (e.g. cricothyroid muscle). Figure 2.3 shows a model of a string with a constant stiffness k (ratio of force and longitudinal extension beyond the equilibrium length), which is plucked at the centre in the transverse direction. If the displacement is y, the length of the string becomes

$$L(y) = 2\sqrt{y^2 + \left(\tfrac{1}{2}L(0)\right)^2} \tag{2.2}$$

where $L(0)$ is the length of the unplucked string. The force inside the string depends on the difference between its current length $L(y)$ and its equilibrium length L_{eq}, which is generally smaller than $L(0)$:

$$F_{//} = 2k\cdot\left(\tfrac{1}{2}L(y) - \tfrac{1}{2}L_{eq}\right) \tag{2.3}$$

The restoring force at the centre thus becomes

$$F_\perp = 2F_{//}\frac{y}{\frac{1}{2}L(y)} = 4ky\left(1-\frac{L_{eq}}{\sqrt{L(0)^2+4y^2}}\right) \approx 4ky\left(1-\frac{L_{eq}}{L(0)}\right)+\frac{8ky^3L_{eq}}{L(0)^3} \tag{2.4}$$

where the last step is an approximation for small displacements. For very small displacements, the force is proportional to the displacement, unless the length $L(0)$ of the unplucked string equals its rest length L_{eq}, i.e., if the unplucked string was not stretched; in that case, the force is proportional to the third power of the displacement. For our model, eq. (2.4) tells us that if the walls are not in contact with each other, the (signed) tension force can be modelled as a “hard” force:

$$\textit{tension force} = k^{(1)}\left(y_{eq}-y\right) + k^{(3)}\left(y_{eq}-y\right)^3 \tag{2.5}$$

where $k^{(1)}$ is the linear spring “constant” (in N/m) of the spring, which may be a function of muscle activity, y_{eq} is the equilibrium position of the wall, which can also be adjusted by the articulatory muscles (e.g., posterior crico-arytenoid activity causes an increase of $\Delta y_{eq} = 2y_{eq}$ in the glottis, risorius does the same for the lips, and expiration is equivalent to reducing Δy_{eq} in the lungs), and $k^{(3)}$ is the cubic spring constant (in N/m^3) of the

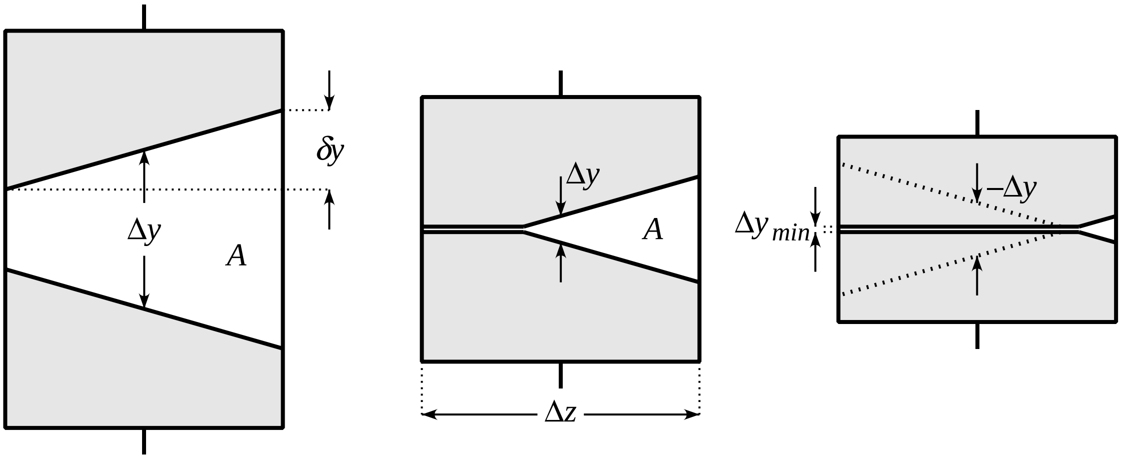


Fig. 2.4 Three consecutive cross-sectional views of a closing tube, showing that the walls make contact like a zipper (time runs from left to right). The cross section A is always positive. The distance Δy between the walls can be negative, as is seen in the last figure. There remains a small leakage Δy_{min} between the walls, even if they are completely closed.

spring. For more circular-shaped elastic walls (say, the alveoli in the lungs), which have a linear displacement-stretch dependence, the cubic spring constant may be 0.

2.3.3 The collision force

When the two masses approach one another, they collide and fold into each other. The masses are not exactly parallel, so the collision is not simultaneous for all points along the z-axis. Figure 2.4 shows a series of cross-sectional views of our stylization of this process; the walls smoothly close upon one another, like a zipper. The cross-sectional area A of a tube equals $\Delta y \cdot \Delta z$ if the distance Δy between the walls is larger than $+\delta y$. For smaller distances, the cross section is determined from figure 2.4, where we see that it can even become negative. The area A, however, cannot be negative, or our physical laws would not work any longer. What's more, our choice of modelling the walls as flat surfaces means that the area A cannot even be a very small positive number: if we allowed very small values of A (which would appear when Δy comes near $-\delta y$), the aerodynamics would show unrealistic behaviour. This is because the existence of very small values of the volume of air in a tube with a constant cross-sectional shape along its length would cause very high positive or negative pressures to arise immediately before or after the moment of contact. We shall circumvent this by allowing a very small leakage Δy_{min} through every tube, so that we can write the width of the opening as a function of the z-coordinate, which runs from 0 to Δz:

$$\Delta y(z) = \Delta y_{min} + \max\left(0, \Delta y - \delta y + \frac{2z\,\delta y}{\Delta z}\right) \tag{2.6}$$

giving for the average width $\langle \Delta y \rangle_{av}$ a smooth function of the distance Δy (i.e., differentiable in $-\delta y$ and in δy), which is always positive:

$$\langle \Delta y \rangle_{av} = \frac{1}{\Delta z}\int_0^{\Delta z} dz\, \Delta y(z) = \begin{cases} \Delta y + \Delta y_{min} & \text{for } \Delta y \geq \delta y \\ \dfrac{(\delta y + \Delta y)^2}{4\delta y} + \Delta y_{min} & \text{for } -\delta y \leq \Delta y \leq \delta y \\ \Delta y_{min} & \text{for } \Delta y \leq -\delta y \end{cases} \tag{2.7}$$

and for the cross section *A*

$$A = \langle \Delta y \rangle_{av} \Delta z \tag{2.8}$$

A good value for Δy_{min} is 0.01 mm: in this case, the relative changes in *A* during a sampling period are not too large (if $\delta y \geq \Delta y_{min}$), while the amount of air that leaks through the orifice is negligible due to the large viscous resistance.

The collision gives rise to a *collision force* which represents the reaction of the tissue against being pressed together. Every part of the wall is compressed to a depth that is a function of *z* (cf. eq. 2.6), the ***compression depth***:

$$d_c(z) = -\tfrac{1}{2}\min\left(0, \Delta y - \delta y + \frac{2z\delta y}{\Delta z}\right) = \max\left(0, \frac{\delta y - \Delta y}{2} - \frac{z}{\Delta z}\delta y\right) \tag{2.9}$$

The average compression depth along the *z* direction is

$$\langle d_c \rangle_{av} = \int_0^{\Delta z} \frac{dz}{\Delta z} d_c(z) = \begin{cases} 0 & \text{for } \Delta y \geq \delta y \\ \dfrac{(\delta y - \Delta y)^2}{8\delta y} & \text{for } -\delta y \leq \Delta y \leq \delta y \\ -\tfrac{1}{2}\Delta y & \text{for } \Delta y \leq -\delta y \end{cases} \tag{2.10}$$

The average cubed compression depth along the *z* direction is

$$\langle d_c^3 \rangle_{av} = \int_0^{\Delta z} \frac{dz}{\Delta z} d_c^3(z) = \begin{cases} 0 & \text{for } \Delta y \geq \delta y \\ \dfrac{(\delta y - \Delta y)^4}{64\delta y} & \text{for } -\delta y \leq \Delta y \leq \delta y \\ -\tfrac{1}{8}\Delta y\left(\Delta y^2 + \delta y^2\right) & \text{for } \Delta y \leq -\delta y \end{cases} \tag{2.11}$$

The force due to the linear part of the stiffness can be computed from the average linear compression depth, and the cubic part is computed from the average cubed compression depth:

$$\textit{collision force} = \begin{cases} \dfrac{s^{(1)}(\delta y - \Delta y)^2}{8\delta y} + \dfrac{s^{(3)}(\delta y - \Delta y)^4}{64\delta y} & \text{for } -\delta y \leq \Delta y \leq \delta y \\ -\tfrac{1}{2}s^{(1)}\Delta y - \tfrac{1}{8}s^{(3)}\Delta y\left(\Delta y^2 + \delta y^2\right) & \text{for } \Delta y \leq -\delta y \end{cases} \tag{2.12}$$

where $s^{(1)}$ and $s^{(3)}$ are the linear and cubic stiffnesses of a wall, respectively. This force is a smooth function of Δy (it is differentiable in $-\delta y$ and in δy).

Our modelling the walls like zippers should not be mistaken for an attempt to simulate actual asymmetric wall behaviour; rather, it is a numerical trick with the objective of ensuring smooth area functions.

2.3.4 The coupling force

If the tissue of the walls of tube *m* is elastically connected to the walls of the adjacent tubes *m*–1 and *m*+1 (not shown in figs. 2.1 and 2.2), the walls of tube *m* may experience a force in the direction of the other walls. The *y* component of this force is expressed as

$$\begin{aligned} &k^{(1)}_{m,m-1}\Big(\big(y_{m-1}-y_{eq,m-1}\big)-\big(y_m-y_{eq,m}\big)\Big)+k^{(3)}_{m,m-1}\Big(\big(y_{m-1}-y_{eq,m-1}\big)-\big(y_m-y_{eq,m}\big)\Big)^3+\\ &+k^{(1)}_{m+1,m}\Big(\big(y_{m+1}-y_{eq,m+1}\big)-\big(y_m-y_{eq,m}\big)\Big)+k^{(3)}_{m+1,m}\Big(\big(y_{m+1}-y_{eq,m+1}\big)-\big(y_m-y_{eq,m}\big)\Big)^3 \end{aligned} \tag{2.13}$$

where the k are the linear and cubic coupling constants. These forces play a role in determining the motions of the upper and lower parts of the vocal folds with respect to each other.

2.3.5 The damping force

The *damping force* is due to internal friction in the tissue. It tries to bring the velocity of the moving wall to zero. It is proportional to this velocity:

$$damping\ force=-\big(B_{open}+B_{closed}\big)\frac{dy}{dt} \tag{2.14}$$

where B_{open} is the damping (in kg/s) of the spring, which depends on the properties of the tissue and dynamically also on $k^{(1)}$, $k^{(3)}$, and m, and B_{closed} is the damping inside the compressed tissue, if the walls are in contact. These dampings are expressed relative to the critical dampings as

$$B_{open}(t)=B_{open,rel}2\sqrt{k_{eff}(t)m}\quad ;\quad B_{closed}(t)=B_{closed,rel}2\sqrt{s_{eff}(t)m_{eff}(t)} \tag{2.15}$$

(Critical damping is the damping that allows a spring to reach equilibrium as quickly as possible without oscillations.) We prefer to have damping that is constant relative to the true critical damping, which involves the cubic spring constants. Otherwise, the relaxation times of the oscillations would be longer in the cubic-force region than in the linear-force region, instead of the other way around. Therefore, we write the dynamic spring “constant” as

$$k_{eff}=-\frac{\partial(tension\ force)}{\partial y}=k^{(1)}+3k^{(3)}\big(y_{eq}-y\big)^2 \tag{2.16}$$

The effective mass for the collision is the mass of the part of the wall that touches the opposing wall:

$$m_{eff}=\begin{cases}0 & \text{for } \Delta y\geq\delta y\\ m\dfrac{\delta y-\Delta y}{2\delta y} & \text{for } -\delta y\leq\Delta y\leq\delta y\\ m & \text{for } \Delta y\leq-\delta y\end{cases} \tag{2.17}$$

The effective stiffness is (remember that $\Delta y=2y$):

$$s_{eff} = -\frac{d(collision\ force)}{dy} = \begin{cases} 0 & \text{for } \Delta y \geq \delta y \\ \frac{\delta y - \Delta y}{2\delta y}\left(s^{(1)} + \frac{1}{4}s^{(3)}(\delta y - \Delta y)^2\right) & \text{for } -\delta y \leq \Delta y \leq \delta y \\ s^{(1)} + \frac{1}{4}s^{(3)}\left(3\Delta y^2 + \delta y^2\right) & \text{for } \Delta y \leq -\delta y \end{cases} \quad (2.18)$$

2.3.6 The air pressure force

If the air pressure inside the tube is greater than the atmospheric pressure, the ***air pressure force*** will try to push the walls apart; if the pressure is less than the atmospheric pressure, the force will try to pull the walls together:

$$air\ pressure\ force = P\,\Delta x\,\Delta z \quad (2.19)$$

where P is the mean air pressure inside the tube, Δx is the length of the tube, and Δz the third dimension ("depth") of the tube, making $\Delta x \Delta z$ the area of the wall. This term expresses one side of the coupling between the myo-elastics and aerodynamics of the vocal apparatus and is responsible for many consonantal features in the languages of the world and for the periodic vibration of the vocal folds.

2.3.7 Parallel subdivision

In some regions of the vocal apparatus, the tubes are subdivided into a number of parallel branches: inside the nasal cavity, there are two to eight parallel branches, and we model the inferior part of the lungs as having many parallel equal branches (fig. 2.6). This branching has an influence on the viscous resistance that the air particles experience when moving along the walls.

2.3.8 The z direction

In the *z* direction ("depth") of the tubes, which is perpendicular to the longitudinal (*x*, "length") direction and also perpendicular to the direction in which the walls can collide (the *y* direction, "width"), the walls are also modelled as mass-spring systems, though they cannot collide and are modelled as exactly parallel. Thus, every tube has at least four walls. At the remaining two ends, the tube has a boundary, through which it is usually connected to a neighbouring tube, as described in the next section.

2.4 From muscles to tract shape

To derive the tract shape at every moment in time, we need as parameters both constant speaker characteristics and time-varying muscle activities. The structure of the vocal apparatus does not change in time: the number of tube sections does not change, nor do

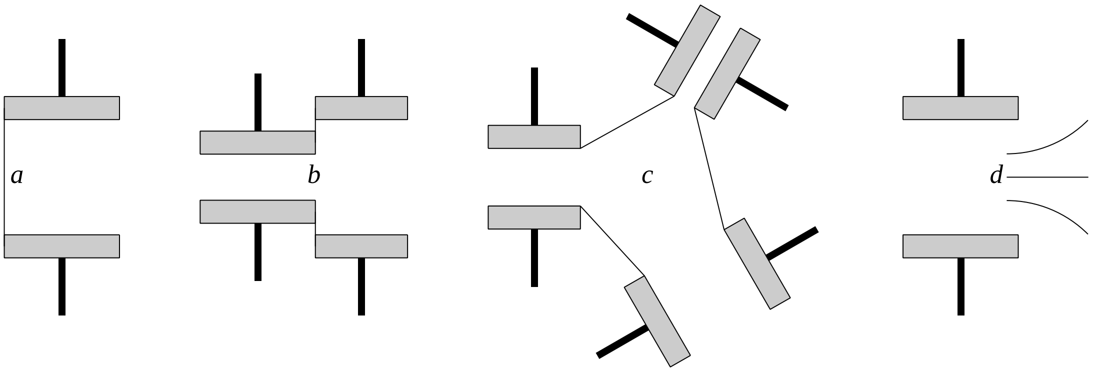


Fig. 2.5 The four types of tube boundaries in our model. (*a*) a closed boundary; (*b*) an interface between two tubes with different lengths; (*c*) an interface between three tubes with equal lengths; (*d*) an open boundary, i.e. a sound-radiating interface with the atmosphere.

their connections to their neighbours. Figure 2.5 shows the four kinds of connections a tube can have to its neighbours:

(*a*) The boundary is closed. In the vocal tract, this happens at the diaphragm.
(*b*) The boundary is an interface to one other tube. This is the most common case. It means that adjacent tube sections form an unbranching duct.
(*c*) The boundary is an interface to two other tubes. This represents a branching, like the velopharyngeal port. The three tubes involved are all treated in the same manner, i.e., we could say that the pharynx branches into the oral and nasal cavities, but we could equivalently say that the nasal cavity branches into the pharyngeal and oral cavities. For numerical reasons, the lengths of the three tubes are forced to be equal.
(*d*) The boundary is open to the air. Variations in airflow are radiated into the environment as sound.

Sections 2.5 to 2.12 describe the structure of the vocal apparatus in terms of these four boundary types, together with the speaker-dependent parameters of the tubes and their walls.

2.5 Speaker properties

Every tube in our model must be specified with its own rest length, width, depth, tension, damping, et cetera. All these properties are speaker-dependent. In order to reduce the number of independent parameters, many default values are determined beforehand (§2.5.2), and these will be the values of the tube parameters unless stated otherwise in the following sections. In the implementation of our model, we can freely change every speaker property; however, we predefined three model speakers (§2.5.1) as starting points.

2.5.1 Three sizes of speakers

Three specimens of the human species make their appearance in our model. The first is the "average speaker", an adult female. The second is the sturdy adult male, characterized mainly by being dimensioned larger by a factor of 1.1; his volumes and masses are therefore larger than hers by a factor of $(1.1)^3$, which approximately equals four thirds. So, if *she* weighs 60 kg and is 170 cm tall, *he* weighs 80 kg and is 187 cm tall. Our third speaker is a child who is characterized by being smaller than the female by a factor of 0.7; for volumes and masses, this factor is $(0.7)^3$, approximately one third. Thus, this child weighs 20 kg and is 119 cm tall.

In our model, nearly all of the speaker characteristics relevant to the vocal apparatus, such as vocal-tract length and lung volume, have these same proportions. A notable exception is the disproportionately large larynx of the male.

Unless stated otherwise, the numbers that appear in the rest of this chapter are for the female speaker. In formulas, the ***size factor*** f appears explicitly: it is 1.0, 1.1, and 0.7 for the female, male, and young speakers, respectively.

2.5.2 Default values

The following values are valid throughout our vocal-tract model, unless specified otherwise in §2.6-9.

The default thickness of the moving walls is taken as $10f$ millimetres. With a tissue density of approximately 1000 kg/m^3, the default surface mass density of the wall is $10f$ kg/m^2, and the default wall mass is

$$m = (10\,f\ \text{kg/m}^2) \cdot \Delta x \cdot \Delta z \tag{2.20}$$

The default surface stiffness density of each wall is taken as 10 mbar/mm, i.e., a pressure of 10 mbar (10 cm H_2O) will push the two walls 2 mm apart. Hence, the default linear wall stiffness is

$$k^{(1)} = (10^6\ \text{N/m}^3) \cdot \Delta x \cdot \Delta z \tag{2.21}$$

The default cubic wall stiffness $k^{(3)}$ is zero.

The zipperiness δy (fig. 2.3) and the minimum width Δy_{min} are taken to be the minimum needed for smooth contact, which is 0.01 mm in both cases.

The linear tissue stiffness is proportional to the area of the wall:

$$s^{(1)}(t) = (5.10^6\ \text{N/m}^3)\ \Delta x(t)\ \Delta z(t) \tag{2.22}$$

and the cubic stiffness constant is chosen to be

$$s^{(3)}(t) = \frac{s^{(1)}(t)}{(0.45\ \text{mm})^2} \tag{2.23}$$

This relation between the linear and cubic stiffnesses is equivalent to the one used by Ishizaka & Flanagan (1972) for the vocal folds.

We assume that the tension of the tissue is isotropic, so that we can approximate the linear coupling-spring constants between the walls of the *m*th and (*m*+1)st tubes by

$$k_{m,m+1}^{(1)}(t) = \tfrac{1}{2}\left(\frac{\Delta z_{eq,m}(t)}{4\Delta x_{eq,m}(t)} k_m^{(1)}(t) + \frac{\Delta z_{eq,m+1}(t)}{4\Delta x_{eq,m+1}(t)} k_{m+1}^{(1)}(t) \right) \tag{2.24}$$

and their cubic counterparts by

$$k_{m,m+1}^{(3)}(t) = \tfrac{1}{2}\left(\tfrac{1}{2}\left(\frac{\Delta z_{eq,m}(t)}{2\Delta x_{eq,m}(t)} \right)^3 k_m^{(3)}(t) + \tfrac{1}{2}\left(\frac{\Delta z_{eq,m+1}(t)}{2\Delta x_{eq,m+1}(t)} \right)^3 k_{m+1}^{(3)}(t) \right) \tag{2.25}$$

The damping factor of an open wall is 0.9, which means that it is slightly undercritically damped (eq. 2.15). In this way, an articulator will usually reach its target value quickly with a slight overshoot (Perrier, Lœvenbruck & Pahan (1996) use a factor of 0.945 for their tongue model). The damping factor of the extra stiffness of closed walls is 1 (critical).

2.6 Sublaryngeal system

Seen from the larynx down, the trachea branches into two main bronchi, these branch into five lobar bronchi, these into 20 segmental bronchi, and so on, until the respiratory bronchioli make contact with 300 million alveoli, whose diameter is 0.2 millimetres or less (table 2.2). We model this by a simple unbranching sequence of 29 tube sections, with constant and fixed lengths (Δx) of 10*f* mm (figure 2.6). The parallel branches are

Table 2.2 The modelling of the lower respiratory system. Some of the widths Δy_0 can be changed by the speaker. The size factor *f* is discussed in §2.5.1.

Approximate anatomy	Number of tubes	Δx	Δy_0	Δz	parallel subdivision
trachea	10	10*f*	11*f*	14*f*	1
main bronchi	2	10*f*	18*f*	9*f*	2
lobar bronchi	1	10*f*	12*f*	12*f*	3
segmental bronchi	1	10*f*	12*f*	12*f*	5
	1	10*f*	18*f*	18*f*	10
	1	10*f*	35*f*	35*f*	20
	1	10*f*	70*f*	70*f*	40
	1	10*f*	120*f*	140*f*	80
bronchioli	1	10*f*	120*f*	240*f*	160
terminal bronchioli	1	10*f*	120*f*	240*f*	320
respiratory bronchioli	1	10*f*	120*f*	240*f*	640
alveoli	1	10*f*	120*f*	240*f*	1250
	1	10*f*	120*f*	240*f*	2500
	6	10*f*	120*f*	240*f*	5000

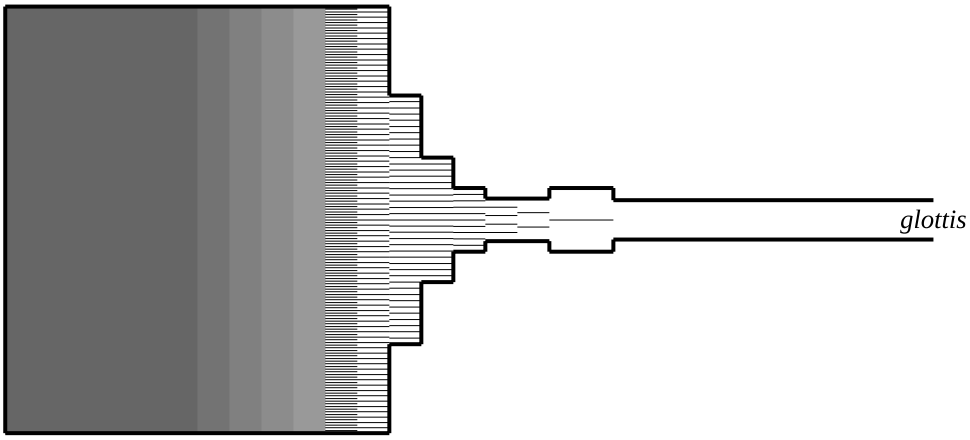


Fig. 2.6 Model of the widths and subdivisions of the subglottal system (to scale).

divided among the x, y and z directions. The fact that there is a large variation in the distance from the alveoli to the larynx, is only partly modelled by the smearing over the deepest tube sections. Our simple modelling has, e.g., all bronchioli acting synchronously with respect to myoelastics and aerodynamics. If this will appear too gross a simplification, we could model the lungs with explicit branches of various lengths.

The numbers in this section were chosen so as to yield realistic values for some macroscopic observables. For instance, the neutral lung volume, which can be computed from table 2.2, comes out as 3.4 litres for the female, 4.5 for the male, and 1.2 litres for the young speaker.

The respiratory muscles can change the equilibrium width Δy_{eq} of the deepest 12 tube sections (those with value '120*f*' in table 2.2) according to

$$\Delta y_{eq}(t) = \Delta y_0 \cdot (1 + lungs(t)) \tag{2.26}$$

where the articulatory parameter *lungs*(t) can be specified to attain values between –0.5 and +1.5: the value –0.5 represents the maximum amount of air that the speaker can exhale by force (the expiratory reserve), and the value +1.5 represents the maximum amount of air that she can inhale (tidal volume plus inspiratory reserve). There is only one *lungs* parameter, because the simplicity of the lung model does not allow us to separate the actions of the diaphragm and the abdominal muscles (vertical extension and compression of the thoracic cage) from those of the muscles that elevate or depress the ribs (horizontal extension and compression of the thorax).

Each tube has for its walls in the y direction two opposing equal masses. The thickness of the walls is taken as 30 mm (cf. eq. 2.20). The linear displacement stiffness density in the elastic part of the lungs (the deepest 12 tubes) is only 10^5 Pa/m (cf. eq. 2.21); in the cartilagenous part, we have $3{\cdot}10^6$ Pa/m. This means, e.g., that if the speaker inhales a speechlike amount of air (*lungs* = +0.2), closes her glottis firmly, and releases the inspiratory muscles, then the air pressure in the lungs will eventually settle down at about 10^5 Pa/m · 120 mm · 0.2 (= activity) / 2 (= two walls) / 2 (= pressure/tension equilibrium) = 600 Pa (6 cm H_2O); this pressure of 600 Pa with a volume change of 10% (= 0.2 / 2) of the vital capacity is a realistic value, according to measurements of the pulmonic relaxation curve (Hixon 1987).

2.7 Larynx

2.7.1 Conus elasticus

We model the conus elasticus with nine tubes, decreasing in area from the trachea to the glottis. The walls are optionally coupled to their neighbours, including the lower part of the vocal folds. This allows us to extend the usual two-mass model of the vocal folds to a system of 11 coupled springs.

2.7.2 Intermembranous glottis

The part of the glottis between the vocal folds is represented by two tubes, much like the two-mass model of Ishizaka & Flanagan (1972), from which we took the constants for the male speaker, except the damping.

For our female speaker, the lower part of the vocal folds has a thickness (tube length Δx) of 1.4 mm, and the upper part 0.7 mm. For our male speaker, these are 2 mm and 1 mm. For the child, they are 0.7 and 0.3 mm. The lengths of the vocal folds (Δz_0) are 10, 18 and 6 mm, respectively.

There are two equal opposing walls in both tubes. In the lower part, their masses are 0.02, 0.1, and 0.003 grams, in the upper part 0.01, 0.05, and 0.002 grams. The neutral tensions are 10, 12, and 6 N/m for the lower part, and 4, 4, and 2 N/m for the upper part. The relative coupling between the two parts is 1. The relative damping is 0.2; this is different from Ishizaka & Flanagan (1972), see §5.5.1 for a discussion.

The cricothyroid and vocalis muscles influence the tension and length of the vocal folds. Assume that their muscle activities α_{ct} and α_{voc} influence both the equilibrium lengths L_{ct}^{α} and L_{voc}^{α} and the stiffnesses k_{ct}^{α} and k_{voc}^{α} of these muscles. Some examples are

$$k_{voc}^{\alpha} = k_{voc}^{min} + \alpha \cdot \left(k_{voc}^{max} - k_{voc}^{min}\right) \quad ; \quad L_{voc}^{\alpha} = L_{voc}^{max} - \alpha \cdot \left(L_{voc}^{max} - L_{voc}^{min}\right) \tag{2.27}$$

but any monotonic functions of α will do. In (2.27), k_{voc}^{min} represents the stiffness in the presence of average spontaneous activity, and L_{voc}^{max} is the length that the muscle would have in the absence of external forces. By definition, the maximum value of α is 1. Its minimum value is somewhat less than 0; the exact value depends on the degree of inhibition the muscle can be subjected to.

The actual length of the vocal folds will then be determined by an equilibrium between the torques of the two muscles around the cricothyroid joint:

$$r_{voc,ctj} k_{voc}^{\alpha}\left(L_{voc} - L_{voc}^{\alpha}\right) = r_{ct,ctj} k_{ct}^{\alpha}\left(L_{ct} - L_{ct}^{\alpha}\right) \tag{2.28}$$

where $r_{voc,ctj}$ and $r_{ct,ctj}$ are the ***arms*** of the two muscle torques, i.e., the perpendicular distances between their lines of force and the joint. The actual length of the vocal folds can be expressed as a perturbation on its length L^{rest} in the absence of extra activity, which, again, can be seen as a perturbation on the length in the absence of other muscles:

$$L_{voc} = L_{voc}^{rest} + \Delta L_{voc} = L_{voc}^{min} + \Delta L_{voc}^{rest} + \Delta L_{voc} = L_{voc}^{min} + \frac{F_{voc}^{rest}}{k_{voc}^{min}} + \Delta L_{voc} \tag{2.29}$$

where F_{voc}^{rest} is the tension in the (more or less) relaxed connected vocal folds. The cricothyroid muscle is shortened by the same amount by which the vocalis is lengthened (taking into account the different arms):

$$L_{ct} = L_{ct}^{min} + \frac{F_{ct}^{rest}}{k_{ct}^{min}} - \frac{r_{voc,cjt}}{r_{ct,cjt}} \Delta L_{voc} = L_{ct}^{min} + \frac{r_{voc,cjt}}{r_{ct,cjt}} \left(\frac{F_{voc}^{rest}}{k_{ct}^{min}} - \Delta L_{voc} \right) \tag{2.30}$$

where the second step makes use of the fact that the sum of the moments about the joint was zero for the relaxed connected system. We write the equilibrium lengths as perturbations on the relaxed equilibrium length:

$$L_{voc}^{\alpha} = L_{voc}^{0} + \Delta L_{voc}^{\alpha} \tag{2.31}$$

Substituting (2.29), (2.30), and (2.31) into (2.28) allows us to compute the lengthening as

$$\Delta L_{voc} = \frac{k_{ct}^{\alpha} \left(\frac{F_{voc}^{rest}}{k_{ct}^{0}} - \frac{r_{ct,ctj}}{r_{voc,ctj}} \Delta L_{ct}^{\alpha} \right) - k_{voc}^{\alpha} \left(\frac{F_{voc}^{rest}}{k_{voc}^{0}} - \Delta L_{voc}^{\alpha} \right)}{k_{voc}^{\alpha} + k_{ct}^{\alpha}} \tag{2.32}$$

In this equation, we see the effects of length control and stiffness control separately. For instance, if stiffness is the only controlled parameter, (2.32) reduces to

$$\Delta L_{voc} = \frac{F_{voc}^{rest} \left(\frac{k_{ct}^{\alpha}}{k_{ct}^{0}} - \frac{k_{voc}^{\alpha}}{k_{voc}^{0}} \right)}{k_{voc}^{\alpha} + k_{ct}^{\alpha}} \tag{2.33}$$

In this case, the results of the actions of both muscles are additive if the sum of their stiffnesses is constant, i.e., if a stimulation of one of the muscles is accompanied by an appropriate inhibition of the antagonist. Such a mechanism is very common in the human body.

If, on the other hand, length is the only controlled parameter, (2.32) reduces to

$$\Delta L_{voc} = \frac{k_{voc} \Delta L_{voc}^{\alpha} - k_{ct} \frac{r_{ct,ctj}}{r_{voc,ctj}} \Delta L_{ct}^{\alpha}}{k_{voc} + k_{ct}} \tag{2.34}$$

In this case, the actions of both muscles are unconditionally additive; with respect to the intended result, of course, which is a length change, the inhibition of the antagonist is still favourable.

We will assume only length control. The length of the vocal folds is thus something like

$$\Delta z_{eq}(t) = \Delta z_0 \cdot \left(1 + cricothyroid(t) - 0.3 \cdot vocalis(t) - 0.2 \cdot externalThyroarytenoid(t)\right) \tag{2.35}$$

where the *externalThyroarytenoid* parameter is shorthand for the synergistic aryepiglottic sphincter muscles, which also include at least the aryepiglottic folds, the oblique arytenoid muscles, and the thyroepiglottic muscles (Lindqvist 1969, 1972).

We guess the equilibrium lengths of the vocalis muscles as

$$L_{voc}^{\alpha}(t) = \frac{L_{voc}^{0}(t)}{1 + 2 \cdot vocalis(t)} \tag{2.36}$$

Note that this is much shorter than the actual length that can be reached by (2.35), because the restoring forces of the cricothyroid muscle and several other structures must be overcome. If we also guess that the unconnected relaxed length L_{voc}^{0} is 90% of the connected relaxed length Δz_0, we can write eqs. (2.4) and (2.5) for the glottis as

$$k^{(1)} = 8k\left(1 - \frac{0.9 \cdot \Delta z_0}{\Delta z \cdot \left(1 + 2 \cdot vocalis(t)\right)}\right) \quad ; \quad k^{(3)} = \frac{32k \cdot 0.9 \cdot \Delta z_0}{\Delta z^3 \cdot \left(1 + 2 \cdot vocalis(t)\right)} \tag{2.37}$$

where we substituted $2y$ for y in (2.4) because we need the *average*, not the *maximum*, displacement here. With a stiffness of k = 12.5, 15, and 7.5 N/m for the lower parts of our three speakers' folds, we find (from 2.37, with $\Delta z = \Delta z_0$ and $vocalis(t) = 0$) the linear neutral tensions $k^{(1)}$ as 10, 12 (used by Ishizaka & Flanagan), and 6 N/m. We find the neutral cubic stiffness constants $k^{(3)}$ as 360, 432, and 216, divided by Δz^2.

We see that we can expect the following phenomena:

- Cricothyroid increases Δz, which causes an increase in $k^{(1)}$, and a decrease in $k^{(3)}$ (which unphysically beats the increase in $k^{(1)}$ if y is greater than $L_0 \cdot \frac{1}{2}\sqrt{2}$). This gives a rising vibration frequency F_0.
- The external thyroarytenoid fibers cause a lower F_0.
- Vocalis has an ambiguous effect: with our choice of parameters, the tightening effect (2.36) usually wins over the relaxing effect (2.35), but if the other sphincter muscles strongly cooperate, the effect on the fundamental frequency is reversed. While vocalis activity is generally found to correlate with pitch raising (Hirano, Vennard & Ohala 1970; Hirose & Gay 1972), the possible ambiguity is noted by Hardcastle (1976: 80).

We ignore the influence of the hyoid depressor muscles, which can lower F_0 by decreasing the vertical tension of the vocal folds (Ohala 1972).

The interarytenoid muscles and the posterior and lateral cricoarytenoid muscles influence the equilibrium width of the glottis as (in units of millimetres)

$$\begin{aligned} \Delta y_{eq}(t) &= 5f - 10f \cdot interarytenoid(t) + \\ &+ 3f \cdot posteriorCricoarytenoid(t) - 3f \cdot lateralCricoarytenoid(t) \end{aligned} \tag{2.38}$$

Thus, an interarytenoid activity of 0.5 brings the vocal folds together into a position suitable for voicing (equilibrium width around 0), and an interarytenoid activity of 1 brings about an *effort closure* of the glottis. During speech, the glottis-opening activity of the posterior cricoarytenoid can be superposed, as happens during aspiration.

Instead of as a sequence of two tubes, the glottis can also be modelled as a single tube by adding the thicknesses, masses, and tensions of the vocal folds.

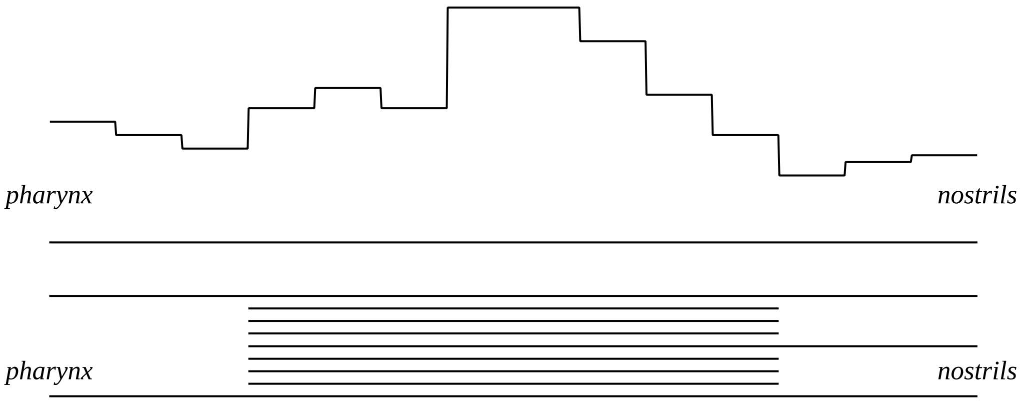


Fig. 2.7 Midsagittal view (above) and transverse view (below) of the nose in our model.

2.7.3 Intercartilagenous glottis

The part of the glottis between the arytenoid cartilages (optional in our model) is implemented separately with the help of branching tubes. This allows a more natural model of phonation than if it were not modelled (Cranen 1987). With the activity of the lateral cricoarytenoid muscles, we can simulate some aspects of breathiness and whispering. We take the width of the space between the arytenoids as

$$\begin{aligned}&\Delta y_{eq}(t) = 5f - 10f \cdot interarytenoid(t) + \\ &- f \cdot posteriorCricoarytenoid(t) + 3f \cdot lateralCricoarytenoid(t)\end{aligned} \tag{2.39}$$

2.8 Nasal cavities

The nasal cavity proper (figure 2.7) consists of 14 tubes, all with a length (Δ*x*) of 7*f* and a width (Δz) of 14*f*. The neutral widths of these tubes, counted from the velopharyngeal port to the nostrils, are 18*f*, 16*f*, 14*f*, 20*f*, 23*f*, 20*f*, 35*f*, 35*f*, 30*f*, 22*f*, 16*f*, 10*f*, 12*f*, and 13*f* mm (Fant 1960). The parallel subdivision in the last three tubes is two, and in the middle eight tubes it is eight, because there are three nasal conchae on each side.

The nasal cavity branches from the 5th and 6th points on the outer contour (fig. 2.8), or, more precisely, from the 13th and 14th tube sections of the pharyngeal and oral cavities (fig. 2.10). Levator palatini lifts the velum and closes the velopharyngeal port. Therefore, the rest width of the first tube section is

$$\Delta y_{eq}(t) = 18f - levatorPalatini(t) \cdot 25f \tag{2.40}$$

According to Maeda (1982), the modelling of a nasal sound will be more natural if it includes a representation of the paranasal sinuses. Therefore, we could (as an example of a possible extension to our model) make the fourth tube section from the nostrils branch to the maxillary sinus cavities, which could then terminate after three tube sections with a closed boundary.

2.9 Pharyngeal and oral cavities

The shape of the oral and pharyngeal cavities is based on the model by Mermelstein (1973), which was also used by McGowan (1994), and, with some undocumented improvements, by Rubin, Saltzman, Goldstein, McGowan, Tiede, & Browman (1996). Another model (Maeda 1982, 1989, 1990), which uses articulatory parameters like tongue-body height and tongue-tip closure, is used by Perrier, Lœvenbruck & Payan (1996) and Vallée (1996). I chose Mermelstein's model because of its explicitness: most of the actual numbers in this section were directly copied from his paper, or measured from one of his figures.

The outline of our model of the (non-nasal) vocal tract, with its basic parameters, is shown in figure 2.8. The outline is computed as 11 points on the outer contour, and 14 points on the inner contour. The outer contour has a relatively fixed position given by the points ($x_{ext,i}$, $y_{ext,i}$) (i = 1...11), and is formed by the rear pharyngeal wall, the velum, the palate, the upper teeth, and the upper lips. The inner contour has a more variable position given by the points ($x_{int,i}$, $y_{int,i}$) (i = 1...14), and is formed by the hyoid bone, the tongue root, the tongue body, the tongue tip, the lower teeth, and the lower lips.

The input to the model is formed by the activities of all the muscle parameters mentioned in this chapter. The workings of some of the muscles are shown in figure 2.9 (where the values of the relevant muscle parameters are 1).

2.9.1 Upper part of the larynx

The position of the hyoid bone is given by the 4th point on the inner contour. It is determined by the speaker's neutral hyoid position and (ignoring some other muscles) the activities of the stylohyoid muscle, which pulls the hyoid bone up by at most 20 mm, the sternohyoid muscle, which pulls it down by at most 20 mm, and the middle pharyngeal constrictor muscle, which pulls it backwards by at most 5 mm:

$$\begin{aligned} \delta x_{hyoid}(t) &= -middleConstrictor(t) \cdot 5f \\ \delta y_{hyoid}(t) &= stylohyoid(t) \cdot 20f - sternohyoid(t) \cdot 20f \end{aligned} \tag{2.41}$$

The larynx moves up and down with the hyoid bone. The anterior larynx does not follow completely the horizontal movements:

$$x_{int,1}(t) = -14f + \tfrac{1}{2}\delta x_{hyoid}(t) \qquad y_{int,1}(t) = -53f + \delta y_{hyoid}(t) \tag{2.42}$$

The top of the larynx:

$$x_{int,2}(t) = -20f + \delta x_{hyoid}(t) \qquad y_{int,2}(t) = -33f + \delta y_{hyoid}(t) \tag{2.43}$$

The epiglottis:

$$x_{int,3}(t) = -20f + \delta x_{hyoid}(t) \qquad y_{int,3}(t) = -26f + \delta y_{hyoid}(t) \tag{2.44}$$

The hyoid bone:

$$x_{int,4}(t) = -16f + \delta x_{hyoid}(t) \qquad y_{int,4}(t) = -26f + \delta y_{hyoid}(t) \tag{2.45}$$

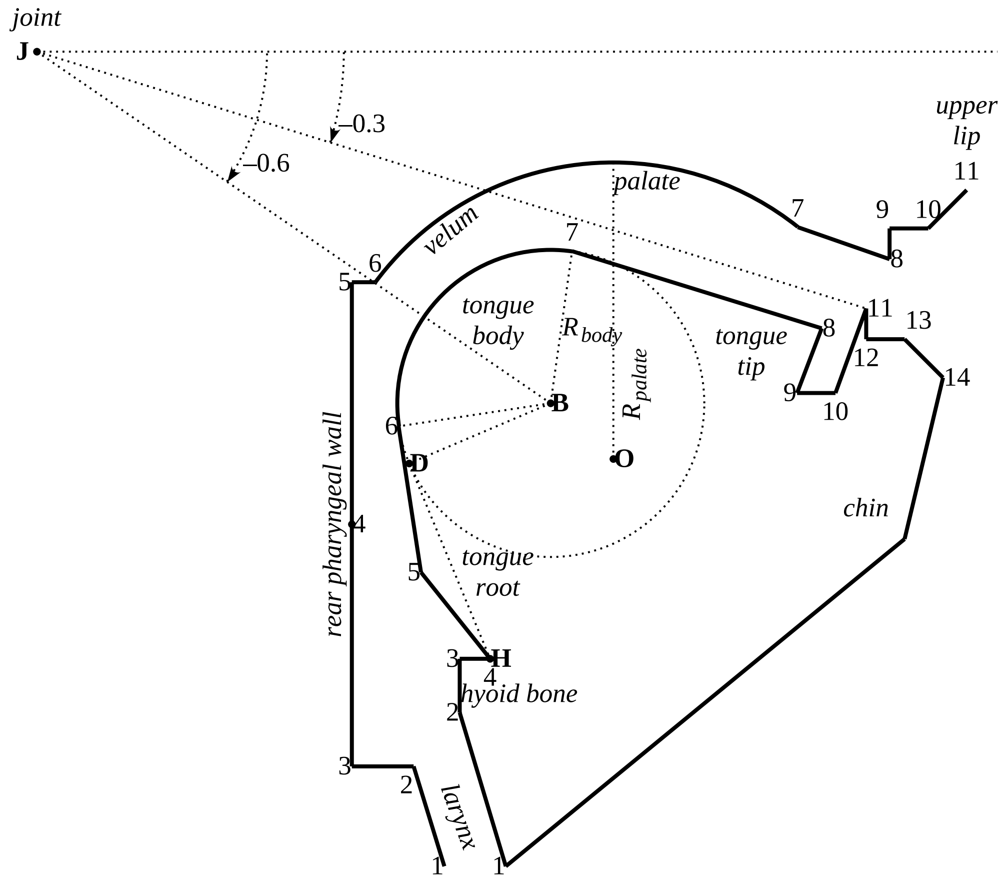


Fig. 2.8 Geometry of the pharyngeal and oral cavities, after Mermelstein (1973).

The posterior larynx:

$$x_{ext,1}(t) = -22f + \delta x_{hyoid}(t) \qquad y_{ext,1}(t) = -53f + \delta y_{hyoid}(t) \tag{2.46}$$

The oesophagus:

$$x_{ext,2}(t) = -26f + \delta x_{hyoid}(t) \qquad y_{ext,2}(t) = -40f + \delta y_{hyoid}(t) \tag{2.47}$$

The lower pharynx moves up and down with the hyoid bone. The neutral horizontal rest position of the back pharyngeal wall is a characteristic of the speaker. The lower constrictor muscle pulls the rear pharyngeal wall forwards:

$$x_{ext,3}(t) = -34f + lowerConstrictor(t) \cdot 5f \qquad y_{ext,3}(t) = y_{ext,2}(t) \tag{2.48}$$

2.9.2 Jaw and tongue body

The angle of the jaw is influenced by the muscles that raise and lower the jaw:

$$\delta\alpha_{jaw}(t) = masseter(t) \cdot 0.15 - mylohyoid(t) \cdot 0.20 \tag{2.49}$$

The location of the centre of the tongue body is determined by the jaw position and by the extrinsic tongue muscles. The styloglossus muscles pull the tongue back up to the styloid process of the temporal bone, the hyoglossus muscles pull the tongue down to the hyoid bone, and the genioglossus muscles pull the tongue forwards to the frontal part of the mandible (**B** and **J** refer to points representing the tongue body and mandibular joint in figure 2.8):

$$\begin{aligned} B_x(t) &= J_x + 81f \cdot \cos\left(-0.60 + \delta\alpha_{jaw}(t)\right) + \\ &\quad - styloglossus(t) \cdot 10f + genioglossus(t) \cdot 10f \\ B_y(t) &= J_y + 81f \cdot \sin\left(-0.60 + \delta\alpha_{jaw}(t)\right) + \\ &\quad - hyoglossus(t) \cdot 10f + styloglossus(t) \cdot 5f \end{aligned} \tag{2.50}$$

where the location of the mandibular joint is given by:

$$J_x = -75f + lateralPterygoid(t) \cdot 20f \qquad ; \qquad J_y = 53f \tag{2.50a}$$

2.9.3 Tongue root

The shape of the tongue root (fig. 2.8) is computed from the position **H** of the hyoid and the position **B** of the tongue body and its radius R_{body} (which is 20*f* mm), with the help of the point **D** where a line through **H** is tangent to the circular mass of the tongue body. The actual tongue-root contour deviates from the flat one given by the line piece *HD*, by having the midpoint of this linepiece replaced in a direction perpendicular to *HD*:

$$\begin{pmatrix} x_{int,5}(t) \\ y_{int,5}(t) \end{pmatrix} = \begin{pmatrix} x_{int,4}(t) \\ y_{int,4}(t) \end{pmatrix} + \begin{pmatrix} \cos\angle HD & -\sin\angle HD \\ \sin\angle HD & \cos\angle HD \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2}HD \\ 0.57(34.8f - HD) \end{pmatrix} \tag{2.51}$$

where the distance *HD* is given by

$$HD = \sqrt{HB'^2 - R_{body}^2} \quad \text{where} \quad HB' = \max\left(HB, R_{body}\right) \tag{2.52}$$

and the angle $\angle HD$ (counterclockwise from the rightward horizontal half-line) is given by:

$$\angle HD = \angle HB + \angle BHD = \operatorname{arctan2}\left(B_y - H_y, B_x - H_x\right) + \arcsin\frac{R_{body}}{HB'} \tag{2.53}$$

If the factor 0.57 appearing in equation (2.51) had been 0, the tongue root would always have been flat. Now, the tongue root moves forward as the tongue body rises.

The sixth point $\mathbf{X}_6$ on the innner contour is the point where a line through the fifth point is tangent to the circular mass of the tongue body:

$$\begin{pmatrix} x_{int,6}(t) \\ y_{int,6}(t) \end{pmatrix} = \begin{pmatrix} x_{int,5}(t) \\ y_{int,5}(t) \end{pmatrix} + X_5X_6 \cdot \begin{pmatrix} \cos\angle X_5X_6 \\ \sin\angle X_5X_6 \end{pmatrix} \tag{2.54}$$

where the distance X_5X_6 is given by

$$X_5X_6 = \sqrt{X_5B^2 - R_{body}^2} \tag{2.55}$$

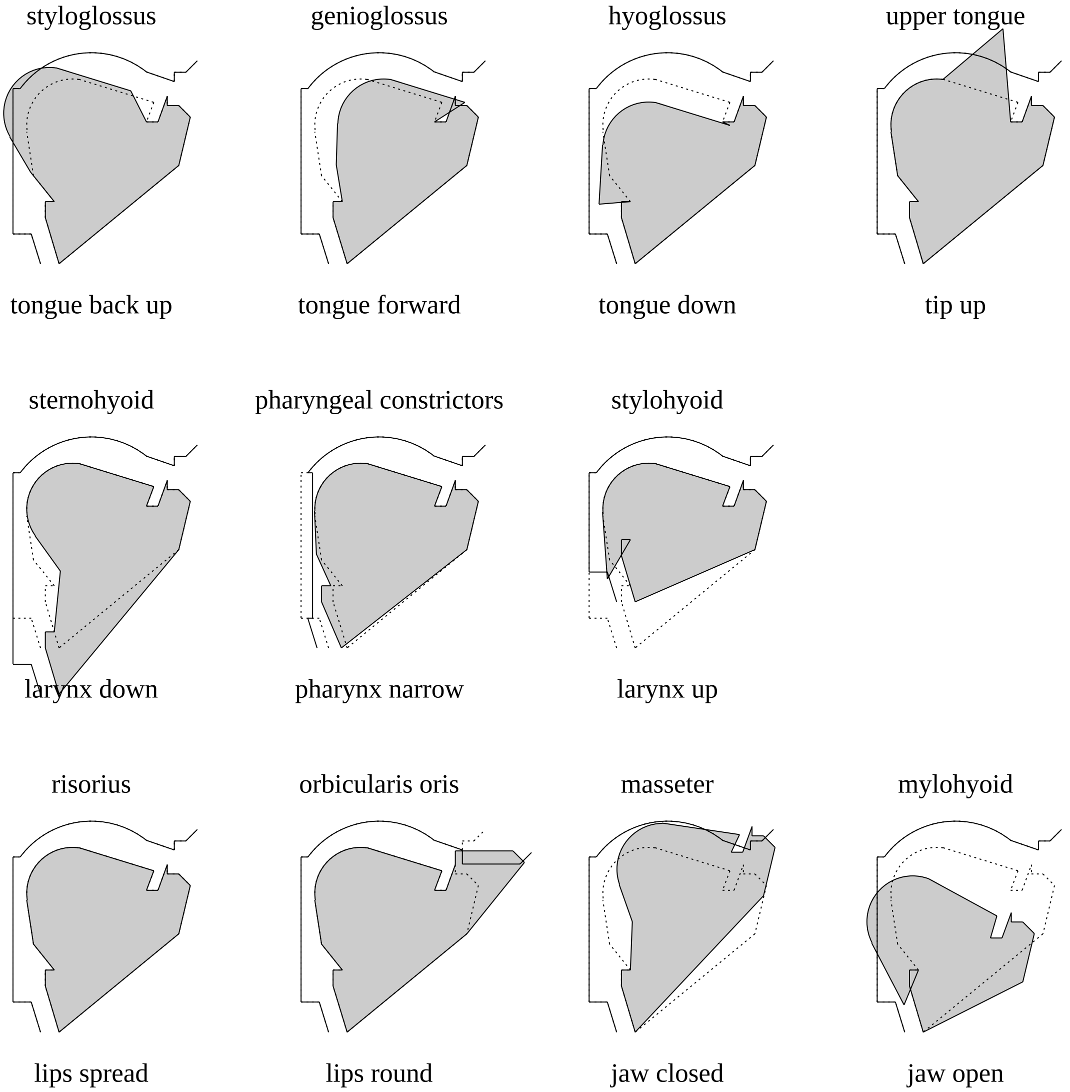


Fig. 2.9 The workings of some of the muscles that determine the shape of the pharyngeal and oral cavities. Some muscles are seen to close off the vocal tract. 'Pharyngeal constrictors' stands for the combined actions of the three pharyngeal constrictor muscles. In each muscle shown, its activity, as defined in the formulas in this chapter, equals 1. The risorius and orbicularis oris muscles also change the tract shape in the *z* direction (not shown).

and the angle $\angle X_5X_6$ is given by (for arctan2, see above eq. 2.70):

$$\angle X_5X_6 = \angle X_5B + \angle BX_5X_6 = \arctan2\left(B_y - y_{int,5}, B_x - x_{int,5}\right) + \arcsin\frac{R_{body}}{X_5B} \qquad (2.56)$$

The procedure described in (2.51) to (2.56) is only a crude approximation of the tongue, which is an only slightly compressible, but highly deformable mass. More sophisticated approaches could now be found in a multi-mass-and-spring representation of the tongue (Perkell 1996), or in a finite-element description (Wilhelms-Tricarico 1995, 1996; Honda 1996). However, because we will focus on the interaction between articulator shapes and aerodynamics, these models should be extended with a method of accounting for the influences of air pressure on articulator shape, which would require an investigation outside the scope of this book, so we do with a simpler approach for now.

The shape of the rear pharyngeal wall is given by (2.48) and by

$$\begin{aligned} x_{ext,5}(t) &= -34f + upperConstrictor(t)\cdot 5f \qquad y_{ext,5}(t) = 23f \\ x_{ext,4}(t) &= -34f + middleConstrictor(t)\cdot 5f \quad y_{ext,4}(t) = \frac{y_{ext,3}(t) + y_{ext,5}(t)}{2} \end{aligned} \tag{2.57}$$

2.9.4 Velum and palate

The position of the back of the velum is given by

$$x_{ext,6}(t) = -31f \qquad y_{ext,6}(t) = 23f \tag{2.58}$$

The palate is a quarter of a circle around the origin O of the coordinate system (see figure 2.8), with a radius of $r_{palate} = f\cdot\sqrt{31^2 + 23^2}$. Therefore, the front end of this arc, which represents the alveolar ridge, is found at

$$x_{ext,7}(t) = y_{ext,6}(t) = 23f \qquad y_{ext,7}(t) = -x_{ext,6}(t) = 31f \tag{2.59}$$

2.9.5 Tongue tip

The posterior position of the tongue blade moves with the jaw and with the tongue body:

$$\begin{pmatrix} x_{int,7}(t) \\ y_{int,7}(t) \end{pmatrix} = \begin{pmatrix} B_x(t) \\ B_y(t) \end{pmatrix} + R_{body}\cdot\begin{pmatrix} \cos(1.43 + \delta\alpha_{jaw}(t)) \\ \sin(1.43 + \delta\alpha_{jaw}(t)) \end{pmatrix} \tag{2.60}$$

The intrinsic tongue-tip elevation angle is a function of the superior and inferior longitudinal tongue muscles, which curl the tongue tip (apex) up and down, respectively:

$$\delta\alpha_{tip}(t) = 1.00\cdot upperTongue(t) - 1.00\cdot lowerTongue(t) \tag{2.61}$$

The angle of the tongue blade relative to the horizontal plane is determined by the angle of the jaw, by the intrinsic tongue-tip elevation, and by the degree to which the tongue body is pressed against the mandible, i.e., the excess distance from mandibular joint to tongue body:

$$\alpha_{blade}(t) = -0.32 + \delta\alpha_{jaw}(t) + \delta\alpha_{tip}(t) + 0.004\cdot(JB(t) - 81f) \tag{2.62}$$

If the tongue blade has a constant length of 34*f*, the position of the tongue tip is

$$\begin{pmatrix} x_{int,8}(t) \\ y_{int,8}(t) \end{pmatrix} = \begin{pmatrix} x_{int,7}(t) \\ y_{int,7}(t) \end{pmatrix} + 34f\cdot\begin{pmatrix} \cos\,\alpha_{blade} \\ \sin\,\alpha_{blade} \end{pmatrix} \tag{2.63}$$

The transverse intrinsic tongue muscle pulls the tongue into a cylindrical shape, like in [l]; the vertical intrinsic tongue muscle flattens the tongue, which causes the tongue tip to be less damped, as in the trill [r] and the tap [ɾ].

2.9.6 Teeth

The distance from the mandibular joint to the cutting edges of the lower teeth is 113*f*, and the angular position of these teeth is directly determined by the position of the mandible:

$$\begin{pmatrix} x_{int,11}(t) \\ y_{int,11}(t) \end{pmatrix} = \begin{pmatrix} J_x \\ J_y \end{pmatrix} + 113f \cdot \begin{pmatrix} \cos\left(-0.30 + \delta\alpha_{jaw}(t)\right) \\ \sin\left(-0.30 + \delta\alpha_{jaw}(t)\right) \end{pmatrix} \tag{2.64}$$

The same formula holds for the deeper (9th and 10th) points on the inner contour, which have distances to the joint of 110*f* and 114*f*, and neutral angles to the horizontal of –0.43 and –0.41 radians, respectively.

The upper teeth are fixed on the outer contour at

$$x_{ext,8}(t) = 36f \qquad y_{ext,8}(t) = 26f \tag{2.65}$$

2.9.7 Lips

The lips can be brought together or drawn apart without spreading or rounding, because the lower lip moves with the jaw. *Spreading* is achieved by the risorius muscles, which spread the lips and pull them back back against the teeth; *rounding* is achieved if the orbicularis oris muscles pull the upper and lower lips together while protruding them, thus changing all three dimensions of the tubes in the lip region:

$$\begin{gathered} \delta x_{lip}(t) = orbicularisOris \cdot 20f \qquad \delta y_{lip}(t) = orbicularisOris \cdot 10f \\ x_{int,12}(t) = x_{int,11}(t) \qquad x_{int,13}(t) = x_{int,12}(t) + 5f + \delta x_{lip}(t) \\ x_{ext,9}(t) = x_{ext,8}(t) \qquad x_{ext,10}(t) = x_{ext,9}(t) + 5f + \delta x_{lip}(t) \\ y_{int,12}(t) = y_{int,13}(t) = y_{int,11}(t) - 4f + \delta y_{lip}(t) \\ y_{ext,9}(t) = y_{ext,10}(t) = y_{ext,8}(t) + 4f - \delta y_{lip}(t) \\ \Delta z_{eq}(t) = 30f - orbicularisOris(t) \cdot 25f + risorius(t) \cdot 25f \end{gathered} \tag{2.66}$$

Finally, the interface to the free air is formed by two line segments making an angle of 45 degrees relative to the horizontal, ending in the points:

$$\begin{gathered} x_{int,14}(t) = x_{int,13}(t) + 5f \qquad y_{int,14}(t) = y_{int,13}(t) - 5f \\ x_{ext,11}(t) = x_{ext,10}(t) + 5f \qquad y_{ext,11}(t) = y_{ext,10}(t) + 5f \end{gathered} \tag{2.67}$$

For the purposes of drawing, a point on the chin is added to close the inner contour, as can be seen in figure 2.8.

2.10 Meshing of the vocal tract

The pharyngeal and oral cavities are represented by 27 tube sections (this number is a trade-off between accuracy and computation time). The 14 points on the inner contour and the 11 points on the outer contour determine the equilibrium lengths and widths of these tube sections.

2.10.1 Mesh points on the outer contour

The 27 mesh points on the outer contour, shown as the outer endpoints of 27 line segments in figure 2.10, are in fixed positions relative to the 11 outer points gotten with the procedure described in 2.9.1 through 2.9.6. The first three points lie on the posterior wall of the upper part of the larynx (the time dependence is suppressed from the following formulas):

$$\vec{x}_1 = \vec{x}_{ext,1} \quad \vec{x}_2 = 0.8 \cdot \vec{x}_{ext,1} + 0.2 \cdot \vec{x}_{ext,2} \quad \vec{x}_3 = 0.4 \cdot \vec{x}_{ext,1} + 0.6 \cdot \vec{x}_{ext,2} \tag{2.68}$$

The next ten points lie at equal intervals on the rear pharyngeal wall:

$$\begin{aligned} i = 4\ldots 8: &\quad \vec{x}_i = 0.2 \cdot \left((8.5 - i)\vec{x}_{ext,3} + (i - 3.5)\vec{x}_{ext,4}\right) \\ i = 9\ldots 13: &\quad \vec{x}_i = 0.2 \cdot \left((13.5 - i)\vec{x}_{ext,4} + (i - 8.5)\vec{x}_{ext,5}\right) \end{aligned} \tag{2.69}$$

The 19th mesh point is on the palatal roof right above the origin, and the 14th through 18th mesh points are on the posterior palatal arc, in such a way that the angles from the origin between consecutive mesh points are equally spaced between the 13th and 19th points (the function arctan2 (y, x) is defined as the angle from the positive x axis to the line connecting the origin and the point (x, y), and lies in the range ($-\pi$, $+\pi$]):

$$\delta\alpha = \frac{\arctan2(y_{13}, x_{13}) - \frac{1}{2}\pi}{6}; \quad i = 14\ldots 19: \quad \vec{x}_i = r_{palate} \begin{pmatrix} \cos(19 - i)\delta\alpha \\ \sin(19 - i)\delta\alpha \end{pmatrix} \tag{2.70}$$

There are three mesh points equally spaced along the horizontal axis between the roof and the 7th point on the outer contour, one mesh point on the 7th contour point, one mesh point halfway between the 7th and 8th contour points, and one mesh point on the 8th contour point (upper teeth). Finally, there is a mesh point in the middle of each of the two line segments that make up the upper lip contour:

$$\begin{aligned} &i = 19\ldots 23: \quad x_i = 0.25(i - 19)x_{ext,7} \quad y_i = \sqrt{r_{palate}^2 - x_i^2} \\ &\vec{x}_{23} = \vec{x}_{ext,7} \quad \vec{x}_{24} = 0.5 \cdot \left(\vec{x}_{ext,7} + \vec{x}_{ext,8}\right) \quad \vec{x}_{25} = \vec{x}_{ext,8} \\ &\vec{x}_{26} = 0.5 \cdot \left(\vec{x}_{ext,9} + \vec{x}_{ext,10}\right) \quad \vec{x}_{27} = 0.5 \cdot \left(\vec{x}_{ext,10} + \vec{x}_{ext,11}\right) \end{aligned} \tag{2.71}$$

2.10.2 The midlines of the tube sections

From each mesh point on the outer contour (except the three in the larynx), a straight line segment is drawn with the following properties (see figure 2.10):

- Its **direction** is independent of the location of the inner contour: for the mesh points below the origin O (the centre of curvature of the palate), the mesh line is horizontal, for the mesh points anterior to the origin it is vertical, and for the mesh points superior

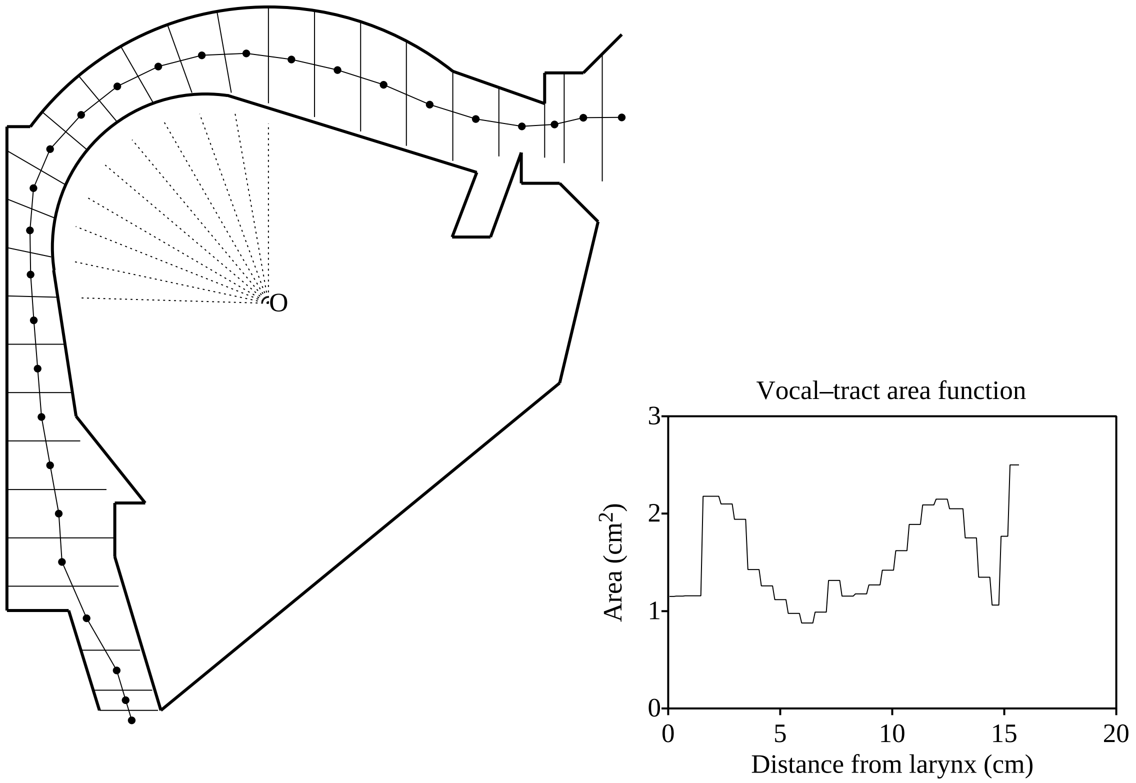


Fig. 2.10 Meshing of the pharyngeal and oral cavities in a neutral position. Each of the 27 mesh lines has a horizontal, radial, or vertical direction and a length that is equal to the distance from the mesh point on the outer contour to the nearest point on the inner contour. The lengths of the mesh lines represent the equilibrium widths of the 27 tube sections. The equilibrium lengths of the 27 tube sections are given by the lengths of the lines that connect the mid-midpoints. The mid-midpoints are shown as dots and are positioned in the middle of the (invisible) lines that connect the midpoints of the mesh lines. The equilibrium area function, which results from multiplying the widths by Δz (which is equal everywhere, except between the lips, see §2.11), is shown at the right.

and posterior to the origin it is radially directed to the origin. If, instead, the direction of the mesh line were chosen as the direction to the nearest point on the inner contour, we would have the unwelcome situation that this direction would not be a continuous function of time, because it would suddenly change as the closest moving structure recedes and another approaches.

- The **length** of the line segment equals the distance from the mesh point on the outer contour to the nearest location on the inner contour (this causes some mesh lines in figure 2.10 to end in the air). If, instead, the length of the mesh line were chosen as equal to the distance from the mesh point to the inner contour in the direction of the mesh line, then this length would not be a continuous function of time, because it would suddenly change as a moving structure touches the mesh line from the side.

Thus, our choice for relatively fixed directions and smallest distances ensures continuity in time of the directions and lengths of the mesh lines, even in situations of wild movements of the articulators; this guarantees that the equilibrium lengths and widths of the tube sections will also be continuous in time. This continuity is a prerequisite for a faithful numerical articulation-to-acoustics transformation (chapter 5).

2.10.3 The lengths of the mesh lines

The length of the mesh line from a certain mesh point on the outer contour, is the minimum of the distances to each of the line and curve segments that constitute the inner contour.

The distance to a line segment. The distance *d* of a mesh point (*x*, *y*) to the line segment that connects the *i*th and (*i*+1)st points on the inner contour, is computed as follows: if the inner product of the vector from the *i*th point to the mesh point and the vector from the *i*th point to the (*i*+1)st point is negative, then the *i*th point is the nearest point of the inner line segment:

$$\begin{aligned}&\text{if}\quad \left(x-x_{int,i}\right)\cdot\left(x_{int,i+1}-x_{int,i}\right)+\left(y-y_{int,i}\right)\cdot\left(y_{int,i+1}-y_{int,i}\right)<0\\&\text{then}\quad d=\sqrt{\left(x-x_{int,i}\right)^2+\left(y-y_{int,i}\right)^2}\end{aligned}\tag{2.72}$$

The same formula goes for the (*i*+1)st point, with the subscripts *i* and *i*+1 reversed. If both the inner products are positive, *d* is the distance from the mesh point to the line through the *i*th and (*i*+1)st points on the inner contour, which equals the absolute value of the outer product of the vector from the mesh point to the *i*th point and the vector from the *i*th point to the (*i*+1)st point, divided by the length of the line segment:

$$d=\frac{\left|\left(x_{int,i}-x\right)\cdot\left(y_{int,i+1}-y_{int,i}\right)-\left(y_{int,i}-y\right)\cdot\left(x_{int,i+1}-x_{int,i}\right)\right|}{\sqrt{\left(x_{int,i+1}-x_{int,i}\right)^2+\left(y_{int,i+1}-y_{int,i}\right)^2}}\tag{2.73}$$

The distance to a curve segment. The distance *d* of a mesh point (*x*, *y*) to the curve that represents the tongue body between the 6th and 7th points on the inner contour (as in fig. 2.8), is computed as follows (angles seen from the centre of the tongue body): if the angle traversed counterclockwise from the 7th point on the inner contour to (*x*, *y*) is smaller than the angle traversed counterclockwise from the 7th to the 6th point on the inner contour, the mesh point is within the pie slice defined by O and the 6th and 7th points, and *d* is the distance to the arc:

$$d=\left|\sqrt{\left(B_x-x\right)^2+\left(B_y-y\right)^2}-R_{body}\right|\tag{2.74}$$

otherwise, if the angle traversed counterclockwise from the midpoint between the 6th and 7th points on the inner contour to (*x*, *y*) is less than π, *d* is the distance from the mesh point to the 6th point; if it is greater than π, it is the distance to the 7th point.

2.10.4 Equilibrium widths of pharyngeal and oral tube sections

The absolute values of the equilibrium widths Δy_{eq} of the 27 tube sections are equal to the lengths of the 27 line segments. However, we must still determine whether Δy_{eq} is positive (open) or negative (closed). The formulas (2.72) to (2.73) are valid in either case.

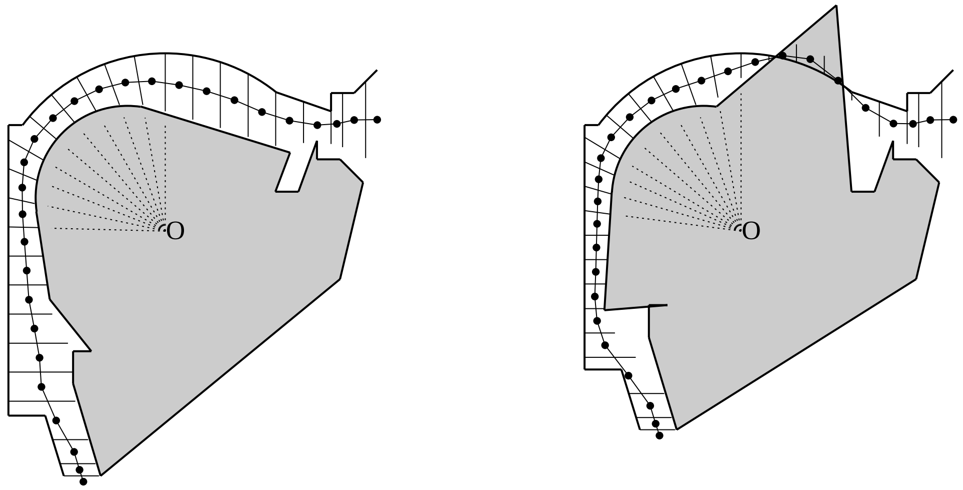


Fig. 2.11 Meshing of the vocal tract in a neutral position (left) and during the closure of the ejective stop [t’] (right). Two images from a film of the simulated utterance [ət’ə]. Each shape shown is the ‘rest’ (target) shape, not the actually realized shape, which depends on inertia, elasticity, and aerodynamics.

Wherever the inner contour crosses the outer contour, the equilibrium width of the tract becomes negative (the walls are pressed together), i.e., the equilibrium width equals minus the length of the line segment if the mesh point is inside the closed inner contour (as at the tongue tip in figure 2.11). The mesh point is inside the inner contour if either its distance to the centre of the tongue body is less than the radius of the tongue body (the sign of the expression between the bars in eq. (2.74) is negative), or it is inside the closed polygon defined by the 14 points that build the inner contour.

To determine whether a mesh point is inside a polygon, we draw an imaginary horizontal line through the point. We then follow the polygon and if two consecutive points of the polygon are on different sides of this line, i.e., if there is a zero crossing, we determine the point of intersection. If this lies to the left of the mesh point, we mark whether the zero crossing was in the upward or downward direction; we ignore zero crossings to the right of the mesh point. If the number of upward zero crossings is different from the number of downward zero crossings, the mesh point is inside the polygon.

2.10.5 Equilibrium lengths of pharyngeal and oral tube sections

The equilibrium length Δx_{eq} of a tube section is the distance between two points that are each halfway between the midpoints of two adjacent mesh lines.

If the ith mesh line runs from the mesh point $\vec{x}_i$ to the inner end $\vec{x}_i'$, the midpoint of the mesh line is at

$$\vec{x}_{m,i} = \tfrac{1}{2}\left(\vec{x}_i + \vec{x}_i'\right) \qquad (2.75)$$

The boundary between two adjacent tube sections is thought to run through a point $\vec{x}_{mm,j}$ that is midway between two adjacent midpoints:

$$\vec{x}_{mm,j} = \tfrac{1}{2}\left(\vec{x}_{m,j-1} + \vec{x}_{m,j}\right) \qquad (j = 2\ldots27) \qquad (2.76)$$

The first and 28th of these points are found by linear extrapolation:

$$\vec{x}_{mm,1} = 2 \cdot \vec{x}_{m,1} - \vec{x}_{mm,2} \qquad \vec{x}_{mm,28} = 2 \cdot \vec{x}_{m,27} - \vec{x}_{mm,27} \tag{2.77}$$

The equilibrium length of the ith tube section in the pharyngeal or oral cavity is now the distance between two consecutive mid-midpoints:

$$\Delta x_{eq,i}(t) = \sqrt{\left(x_{mm,i} - x_{mm,i+1}\right)^2 + \left(y_{mm,i} - y_{mm,i+1}\right)^2} \tag{2.78}$$

2.11 Other oral and pharyngeal properties

Section 2.10 treated how the equilibrium lengths Δx_{eq} and widths Δy_{eq} of the pharyngeal and oral tube sections were to be found. As Baer, Gore, Gracco & Nye (1991) showed, constrictions in the vocal tract have an almost circular cross section. Therefore, we take the third dimension Δz_{eq} to approximate the width everywhere, without actually becoming zero:

$$\Delta z_{eq}(t) = \Delta y_{eq}(t) + 2f\,\text{mm} \tag{2.79}$$

Between the lips, however, we follow (2.66).

The cubic spring constant is chosen as

$$k^{(3)} = k^{(1)} \left(\frac{10}{\Delta z} \right)^2 \tag{2.80}$$

which means that the distance where the third-power force equals the linear force, is $\Delta z/10$.

2.12 Time

The activities of the articulating muscles are slowly varying functions of time. In our implementation, their values are interpolated linearly between the nearest target values specified. For instance, if the spring k is specified as k_1 at a time t_1 and as k_2 at a time t_2, and there are no specifications for k at times between t_1 and t_2, the spring at every time t between t_1 and t_2 is expressed as

$$k(t) = k_1 + \frac{t - t_1}{t_2 - t_1}\left(k_2 - k_1\right) \tag{2.81}$$

This linear behaviour of a control parameter is also used by Perrier, Lœvenbruck & Payan (1996).

At least two targets have to be specified for each articulatory dimension:

1. The starting points at $t = 0$. The starting values of the equilibrium dimensions are the starting values of the dimensions themselves as well.

2. The end points at $t = T$, which is the time at which the simulation stops.

Instead of having a linear interpolation between targets that are separated in time, we could also have immediate changes in target positions. This may be one of the differences between ballistic and controlled movements; the model accepts these immediate changes as a special case of linear interpolation.

In chapter 5, we will see many examples of articulatory parameters as functions of time, and the shapes that they realize.

2.13 Conclusion

The novelties in our articulation model are:

- The entire speech apparatus is modelled in the same way.
- Tube lengths can vary as functions of time, so that we can model faithfully speech sounds that crucially depend on longitudinal movements.
- The meshing algorithm is resistant to the wildest movements, as shown in §2.10.2.
- The zipperiness allows smooth closing and opening phases with only one parameter.

Room for improvement is found at the following points:

- If we allowed more structures and muscles in our model, we would model the dependencies between the actions of different muscles more faithfully. For instance, the amount by which mylohyoid can lower the jaw or front the hyoid depends on the activities of masseter, sternohyoid, and middle constrictor, which are capable of fixing the position of one or the other bone. In reality, therefore, the results of the actions of the muscles are not as linearly additive as they are modelled here. We could model this by replacing our linear array with a general system of connected masses and springs.
- If we used a finite-element simulation of the tongue mass, we would honour the constancy of the tongue volume more correctly. However, the current models that incorporate this (Wilhelms-Tricarico 1995, 1996; Honda 1996) do not yet allow predictable control strategies or an interaction with air pressure.

For now, we must be satisfied with the novelties of our model. In chapter 5, we will see that we can faithfully simulate several real-life phenomena that were never simulated before.

3 Acoustical simulation[1]

Abstract. This chapter derives the aerodynamic equations needed in our articulation model, translates the myoelastic and aerodynamic equations into difference equations for numerical simulation, and presents the actual computer algorithm.

The procedure described in chapter 2 gives us a number of properties of a network of tubes. To compute the state of this network for every moment of time, we need equations that describe the physical behaviour of the walls of these tubes and equations that describe the evolution of the movements and pressures of the air in the network. These ***myoelastic*** and ***aerodynamic*** equations are coupled. Chapter 2 described the myoelastic equations; in this chapter, we will derive the aerodynamic equations.

3.1 The equation of continuity of mass flow

The principle of the conservation of mass is expressed as follows: “The increase during a certain amount of time of the mass contained in a volume is equal to the mass that flows into that volume during that time minus the mass that leaves that volume during that time”. We will derive an integral equation directly from the wording of this conservation law (§3.1.1). This approach is different from those found in the speech literature so far (§3.1.3). We will show why our approach is the only correct one (§3.1.2).

3.1.1 The integral equation of continuity

Consider a channel (duct, tube) extending along the x direction, with a time- and position-dependent cross section (area) $A(x,t)$, expressed in m^2 (fig. 3.1). If this channel contains a fluid (e.g., air) with a mass density of $\rho(x,t)$, expressed in kg/m^3, that is constant across its cross section, the mass contained in the channel between two arbitrary positions x_1 and x_2 (expressed in metres) is

$$\int_{x_1}^{x_2} \rho(x,t)\, A(x,t)\, dx \tag{3.1}$$

The mass flow (in units of kg/s) in the positive x direction at x_1 at any time t is

$$\rho(x_1,t)\, v(x_1,t)\, A(x_1,t) \tag{3.2}$$

where $v(x,t)$ is the particle velocity (expressed in m/s) along the channel, averaged over all y and z positions across the cross section of the channel; a ***particle*** is considered a homogeneous piece of the fluid: it is infinitesimally small but contains infinitely many

[1] This chapter elaborates on the second halves of Boersma (1991) and Boersma (1993a).

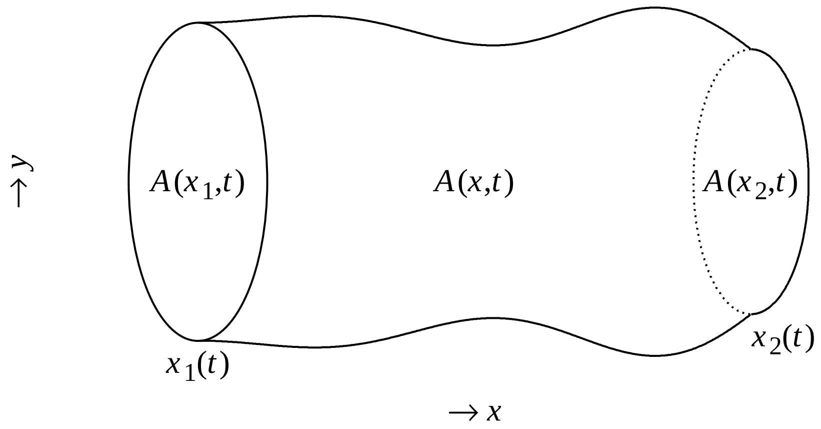


Fig. 3.1 A part with moving boundaries, of a channel with moving walls.

molecules. The exact integral equation describing the mass gain between the times t_1 and t_2 is thus

$$\begin{gathered}\int_{x_1(t_2)}^{x_2(t_2)} \rho(x,t_2)A(x,t_2)dx - \int_{x_1(t_1)}^{x_2(t_1)} \rho(x,t_1)A(x,t_1)dx = \\ \int_{t_1}^{t_2} \rho(x_1,t)v(x_1,t)A(x_1,t)dt - \int_{t_1}^{t_2} \rho(x_2,t)v(x_2,t)A(x_2,t)dt\end{gathered} \tag{3.3}$$

This ***continuity equation*** is still correct if x_1 and x_2 depend on time; the velocities at x_1 and x_2 should then be taken relative to the velocities by which the positions x_1 and x_2 are moving. The closest we can get (3.3) in the direction of a differential equation is therefore

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} \rho A\,dx = (\rho v A)_{x_1} - (\rho v A)_{x_2} \tag{3.4}$$

3.1.2 Pumping and sucking

Though our difference equations will be derived directly from the integral equation (3.3), it is instructive to take the place derivative of (3.4), assuming that x_1 and x_2 are constant, and rewrite the integral equation as an exact differential equation for the continuity of mass flow in a channel:

$$\frac{\partial(\rho A)}{\partial t} + \frac{\partial(\rho v A)}{\partial x} = 0 \tag{3.5}$$

Now, we could have tried to derive (3.5) from the continuity equation of hydrodynamics (e.g. Landau & Lifshitz 1953), which must hold everywhere inside our channel:

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho\mathbf{v}) = 0 \tag{3.6}$$

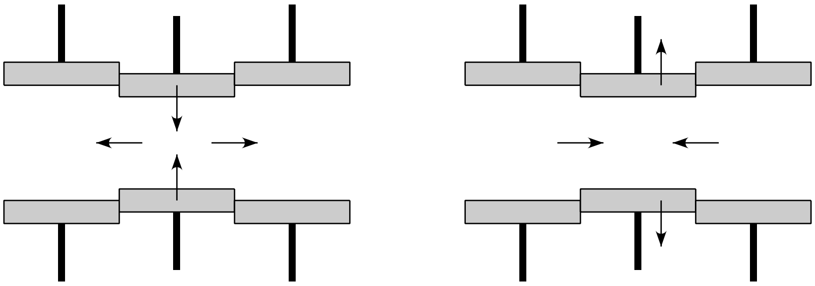


Fig. 3.2 Pumping (left) and sucking (right).

where the ***divergence*** operator is defined as

$$\operatorname{div}(\rho\mathbf{v}) \equiv \frac{\partial(\rho v_x)}{\partial x} + \frac{\partial(\rho v_y)}{\partial y} + \frac{\partial(\rho v_z)}{\partial z}$$

where v_x, v_y, and v_z are the local particle velocities in the three independent directions. Thus, the one-dimensional continuity equation, which describes the case where flow is constrained along one direction, reads

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial x} = 0 \tag{3.7}$$

This equation is correct for a straight tube with rigid walls. It is incorrect, however, for the case of a channel with moving walls or with a non-uniform cross section, because in these cases the transverse velocities v_y and v_z are typically *not* zero. For this reason, we call the correct differential equation (3.5) a ***pseudo-one-dimensional*** equation.

An illustration of the correct position of A in (3.5) and its incorrect position in (3.7) is the case of an incompressible fluid, where the density ρ is independent of time and position. The differential equation (3.5) then reduces to

$$\frac{\partial A}{\partial t} + \frac{\partial(vA)}{\partial x} = 0 \tag{3.8}$$

where vA is the ***volume flow*** along the channel, expressed in m^3/s. Equation (3.8) states correctly that if the cross section of a certain region in the channel shrinks, the fluid will flow out of that region, and that if it widens, fluid will be drawn into the region. We shall call these processes ***pumping*** and ***sucking***, respectively (fig 3.2).

Flanagan and Ishizaka (1977) state that the pumping effects of the vibrating vocal cords on the aerodynamics of the glottis are negligible. However, as chapter 5 will show, we cannot ignore these effects when we model changes in lung volume or variations in the tension of the supralaryngeal musculature like those that are partly responsible for voicing contrasts in obstruent consonants.

3.1.3 Others' choices for the continuity equation

Most of the representations of the continuity equation found in the literature differ from our equations (3.3) to (3.5). This does not mean that they produce incorrect results for the kinds of utterances they were designed for, like sustained vowels, but it does mean that we cannot use them for simulating many of the physical phenomena utilized by consonants.

The bare reflection-type synthesizer of Kelly & Lochbaum (1962) and others, the vocal tract of Ishizaka & Flanagan (1972), and the integration along characteristics of Sondhi & Resnick (1983) freely ignore the time-dependence of the cross section *A* in (3.5), dragging *A* outside and inside the parentheses as comes in handy (this does allow of much faster numerical algorithms than the one used in this book).

The synthesizer by Maeda (1982) does contain an equation reminiscent of (3.5), with ρA replaced by an expression involving the pressure (as we will do in §3.3), and ρ taken outside the time-derivative (which is a sensible simplification that will serve *our* solution, too). However, the yielding of the walls to the air pressure is treated as a small perturbation of the wall position, which makes the connection with the equation of motion inaccurate; nevertheless, as the walls are extrinsically moved anyway, it seems that Maeda could as easily have taken the more exact approach here.

For the other major aerodynamic equation, which is the equation of motion described in the next section, the differences between our integral equation and most of those found in the literature are even greater than in the case of the equation of continuity.

3.2 The equation of motion

The forces on a particle within a fluid can be described as follows: “The particle experiences a force in the down-hill direction of the pressure gradient. At the same time, its velocity relative to other particles is impeded by viscous friction”.

3.2.1 Pressure gradient

Just like we did with the continuity equation, we shall derive this part of the equation of motion from first principles, in this case from Newton’s law. Let us assume that the following approximations hold:

1. All motion is parallel to the *x* axis: $v_y = v_z = 0$, so that we can define $v \equiv v_x$;
2. The velocities are constant along the *y* and *z* axes: $\partial v / \partial y = \partial v / \partial z = 0$;
3. The pressure is constant in the *yz* plane: $P = P(x,t)$.

The force on a particle of air (fig. 3.3) with mass *m* and dimensions *dx*(*t*), *dy*(*t*) and *dz*(*t*), travelling through the positions *x*(*t*), is then

$$\begin{aligned} m\frac{dv(t)}{dt} &= F_{pressure,left}(t) - F_{pressure,right}(t) = \\ &= P\Big(x(t) - \tfrac{1}{2}dx(t),t\Big)dy(t)\,dz(t) - P\Big(x(t) + \tfrac{1}{2}dx(t),t\Big)dy(t)\,dz(t) \\ &= \frac{\partial P(x(t),t)}{\partial x}\,dx(t)\;dy(t)\;dz(t) \end{aligned} \tag{3.9}$$

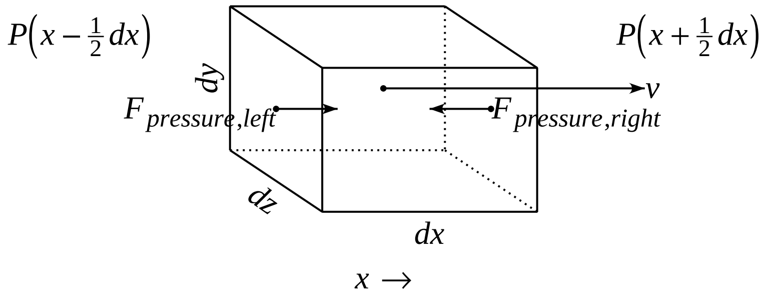


Fig. 3.3 An air "particle" moving along the *x* direction. The pressure gradient is also along the *x* direction.

where m is the particle's mass, $v = v(x,t)$ is the signed particle velocity along the x direction, and $P(x,y,t)$ is pressure expressed in Pascal (Pa, Newtons per square metre). If we substitute for the mass

$$m = \rho(t)\, dx(t)\, dy(t)\, dz(t) \tag{3.10}$$

and divide by the volume $dx(t)\, dy(t)\, dz(t)$, the equation of motion becomes

$$\rho(t)\frac{dv(t)}{dt} = -\frac{\partial P(x(t),t)}{\partial x} \tag{3.11}$$

3.2.2 Bernoulli effect

If we now replace the material derivative by a local derivative according to

$$\frac{df(t)}{dt} = \frac{\partial f(x(t),t)}{\partial t} + \frac{\partial f(x(t),t)}{\partial x}\frac{dx(t)}{dt} = \frac{\partial f(x(t),t)}{\partial t} + \frac{\partial f(x(t),t)}{\partial x}v(x(t),t) \tag{3.12}$$

we can translate the frame of reference from the particle to a coordinate system fixed in space, so that we get

$$\rho(x,t)\frac{\partial v(x,y,z,t)}{\partial t} = -\frac{\partial P(x,t)}{\partial x} - \rho(x,t)v(x,y,z,t)\frac{\partial v(x,y,z,t)}{\partial x} \tag{3.13}$$

For an incompressible fluid, this would reduce to

$$\rho\frac{\partial v}{\partial t} = -\frac{\partial P}{\partial x} - \frac{\partial\left(\frac{1}{2}\rho v^2\right)}{\partial x} \tag{3.14}$$

Thus, for a stationary flow ($\partial v/\partial t = 0$) in an incompressible fluid, an increase in the particle velocity caused by a narrowing along the tube, is accompanied by a pressure drop along this narrowing. Though the mathematics of (3.14) does not distinguish cause and consequence, our intuitive idea of the direction of causality must be reversed here: while we normally think that a pressure gradient causes a change in velocity, we should now see that a change in velocity causes a pressure gradient. This ***Bernoulli effect*** is responsible for lifting airplanes (partly) and for sucking together the walls of a duct at a constriction. As such, it is the indispensable physical phenomenon that delivers energy to the vibrating vocal folds; this ***aerodynamic-myoelastic theory*** of vocal-fold vibration is due to Van den Berg, Zantema & Doornenbal (1957).

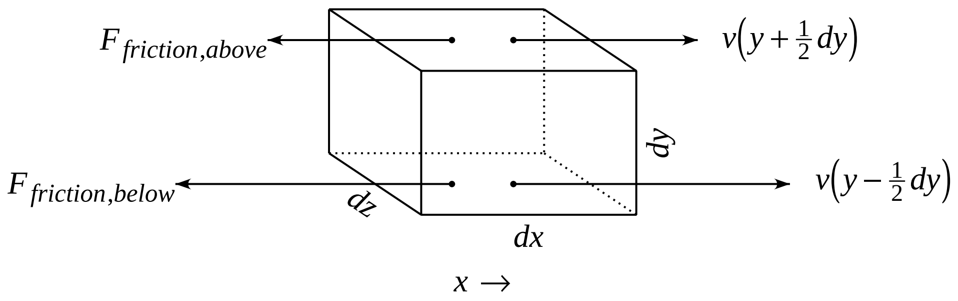


Fig. 3.4 The forces of friction in the case of a velocity gradient in the *y* direction.

3.2.3 Friction

We now relieve the constraint on the velocity v in the x direction: it depends on the y position, i.e, on the distance to the walls, so that $\partial v/\partial y$ does not vanish (fig. 3.4). As $\partial v/\partial z$ is still zero, this situation is representative of a straight tube whose extent in the z direction is much larger than in the y direction. The part of the equation of motion that describes friction reads

$$m\frac{dv(t)}{dt} = F_{friction,above}(t) + F_{friction,below}(t) =$$
$$= \mu\frac{\partial v\left(x(t),y(t)+\frac{1}{2}dy(t),t\right)}{\partial y}dx(t)\,dz(t) - \mu\frac{\partial v\left(x(t),y(t)-\frac{1}{2}dy(t),t\right)}{\partial y}dx(t)\,dz(t) =$$
$$= \mu\frac{\partial^2 v(x(t),y(t),t)}{\partial y^2}\,dx(t)\,dy(t)\,dz(t) \tag{3.15}$$

where μ is the coefficient of shear (laminar) viscosity, which is $1.86\cdot 10^{-5}$ Ns/m^2 for air. Dividing again by the volume, and adding the pressure term, yields the one-dimensional Navier-Stokes equation:

$$\rho(x,t)\frac{\partial v(x,y,z,t)}{\partial t} = -\frac{\partial P(x,t)}{\partial x} - \rho(x,t)v(x,y,z,t)\frac{\partial v(x,y,z,t)}{\partial x} + \mu\frac{\partial^2 v(x,y,z,t)}{\partial y^2} \tag{3.16}$$

As our model consists of a one-dimensional array of tube sections, we would like to suppress the y and z dependence of this equation by averaging it over all values of y and z between the walls. In this procedure, we can replace the velocity in the left-hand side by the average velocity, and the first term on the right hand does not change. We are left with the task of finding a constant $c_{Bernoulli}$ and a function $R(x,t)$ so that

$$\rho(x,t)\frac{\partial v(x,t)}{\partial t} = -\frac{\partial P(x,t)}{\partial x} - c_{Bernoulli}\,\rho(x,t)v(x,t)\frac{\partial v(x,t)}{\partial x} - R(x,t)v(x,t) \tag{3.17}$$

where v is now the velocity averaged over all y and z coordinates inside the tube, and R is called the acoustic ***viscous resistance*** per unit length, which is expressed in Nsm^{-4}. The viscous resistance of a tube can be solved from the boundary condition that $v = 0$ at the

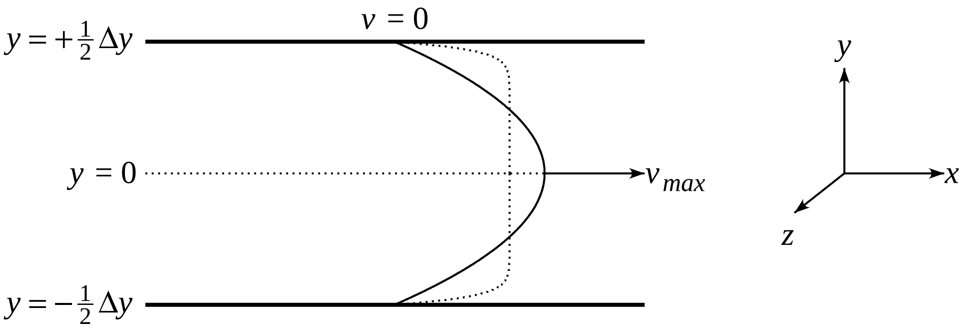


Fig. 3.5 Parabolic velocity profile for Hagen-Poiseuille flow (solid curve), and velocity profile for boundary-layer flow (dotted curve). In both cases, the velocity is zero at the walls, and maximal in the middle.

walls. In the static case, $\partial^2 v / \partial y^2$ is constant as a function of y and z (Hagen-Poiseuille flow, figure 3.5).

Now consider our zipper-shaped tube (figure 2.4). If y is measured from a plane parallel to and at equal distances from the two masses, then

$$v(y,z) = \alpha\left(\left(\tfrac{1}{2}\Delta y(z)\right)^2 - y^2\right) \tag{3.18}$$

where $\Delta y(z)$ is given by equation (2.6), and α does not depend on y or z (figure 3.5). The mean velocity between the plates is

$$v = \frac{\int_0^{\Delta z} dz \int_{-\frac{1}{2}\Delta y(z)}^{+\frac{1}{2}\Delta y(z)} dy\, v(y,z)}{\int_0^{\Delta z} dz \int_{-\frac{1}{2}\Delta y(z)}^{+\frac{1}{2}\Delta y(z)} dy} = \frac{\int_0^{\Delta z} dz \frac{\alpha}{6}\left(\Delta y(z)\right)^3}{\int_0^{\Delta z} dz\, \Delta y(z)} = \frac{\alpha}{6}\frac{\left\langle \Delta y^3 \right\rangle_{av}}{\left\langle \Delta y \right\rangle_{av}} \tag{3.19}$$

The integral for $\langle \Delta y \rangle_{av}$ is evaluated as equation (2.7), and

$$\left\langle \Delta y^3 \right\rangle_{av} = \begin{cases} \left(\left(\Delta y + \Delta y_{min}\right)^2 + \delta y^2\right) \cdot \left(\Delta y + \Delta y_{min}\right) & \text{for } \Delta y \geq \delta y \\ \dfrac{\frac{1}{4}\left(\left(\Delta y + \delta y + \Delta y_{min}\right)^4 - \Delta y_{min}^4\right) + \Delta y_{min}^3\left(\delta y - \Delta y\right)}{2\delta y} & \text{for } -\delta y \leq \Delta y \leq \delta y \\ \Delta y_{min}^3 & \text{for } \Delta y \leq -\delta y \end{cases} \tag{3.20}$$

where δy is the zipperiness defined in §2.3.3. Because

$$R(x,t)v(x,t) \equiv -\mu \frac{\partial^2 v(x,y,z,t)}{\partial y^2} = 2\mu\alpha \tag{3.21}$$

we get the following expression for the resistance, which is a continuously differentiable function of Δy:

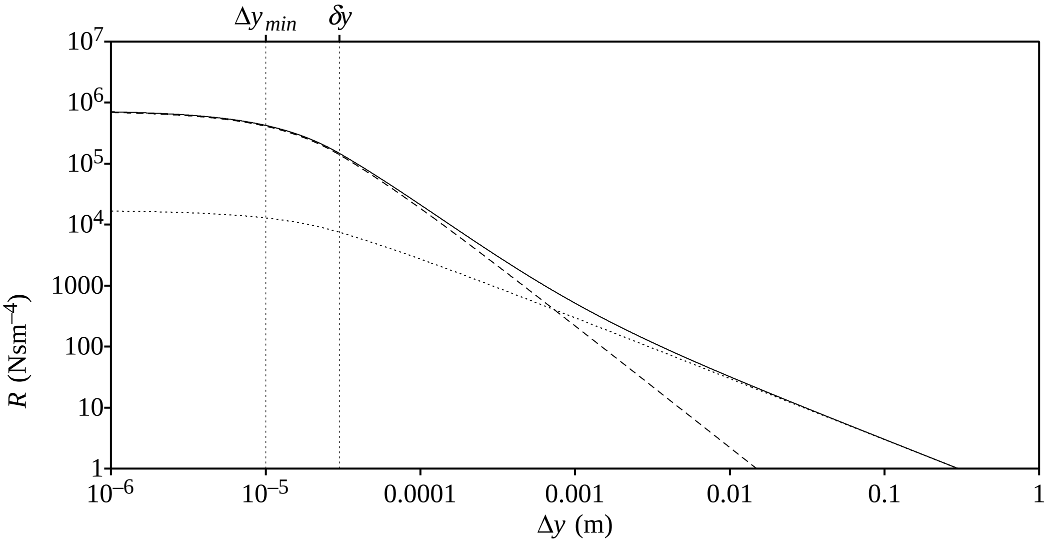


Fig. 3.6 The resistance as a function of the distance Δy between the walls (solid curve). For small distances, it approaches the Hagen-Poiseuille resistance (dotted curve), and for large distances, it approaches the boundary-layer resistance (dashed curve). The zipperiness δy was taken as 0.03 mm, the minimum width Δy_{min} as 0.01 mm.

$$R = \frac{2\mu\alpha}{v} = \frac{12\mu\langle\Delta y\rangle_{av}}{\langle\Delta y^3\rangle_{av}} \tag{3.22}$$

This formula for the fully developed laminar flow (also found in Flanagan 1965) is only valid for small tube widths. For large tube widths, the viscous resistance is only found in the boundary-layer, where it is frequency dependent (Morse & Ingard 1968). This resistance is inversely proportional to the tube width, so that the total viscous resistance can be written as the *sum* of the Hagen-Poiseuille and boundary layer resistances:

$$R = \frac{0.3\ \mathrm{Ns/m^3}}{\langle\Delta y\rangle_{av}} + \frac{12\mu\langle\Delta y\rangle_{av}}{\langle\Delta y^3\rangle_{av}} \tag{3.23}$$

Figure 3.6 shows an example of the viscous resistance as a function of Δy.

In the case of Hagen-Poiseuille flow, the second term on the right hand in (3.17) would receive a factor $c_{Bernoulli} = 6/5$, as a tedious computation would show, but I will take it to be 1, which would be appropriate for large widths.

For tubes that are subdivided into N “parallel” branches, the resistance of each branch is given by (3.23) with the widths divided by N. The total resistance of all these resistances in parallel is equal to the resistance of each branch (the parallel-resistance formula known from electric-circuit theory would be valid instead if (3.17) referred to volume velocities, not particle velocities). We further simplify the resistance R as

$$R = \frac{0.3\ \mathrm{Ns/m^3}}{\langle\Delta y\rangle_{av}} N + \frac{12\mu}{\langle\Delta y\rangle_{av}^2} N^2 \tag{3.24}$$

3.2.4 Complete equation of motion

The shortest form of the complete equation of motion is

$$\rho \frac{\partial v}{\partial t} = -\frac{\partial P}{\partial x} - \rho v \frac{\partial v}{\partial x} - Rv \qquad (3.25)$$

3.2.5 Others' choices for the equation of motion

Volume velocity. The quantity from which the time derivative is taken, is the particle velocity v. However, some authors use the volume velocity $U \equiv vA$ here.

Ishizaka & Flanagan (1972), for instance, decided to use $L\,\partial U/\partial t$, where $L \equiv \rho/A$, on the left-hand side of (3.24). They remark in a footnote that they should have used $\partial(LU)/\partial t$, but that this makes no difference in the glottis. My early simulations showed, however, that the particle velocities became much too high (80 m/s versus 40 m/s normally) and that the frequency of the vibrating vocal cords was 50% higher than with the correct formula.

Bernoulli effect. The Bernoulli effect is ignored by all authors, except those who want to model vocal-cord vibration (Van den Berg, Zantema & Doornenbal 1957; Ishizaka & Flanagan, 1972), but Ishizaka & Flanagan's implementation of this effect with positive and negative resistances is less principled than the approach in this book (Boersma 1991).

Resistance. For the viscous resistance, most other authors use either the Hagen-Poiseuille formula (Van den Berg, Zantema & Doornenbal 1957; Maeda 1982) or the boundary-layer formula (Fant 1960; Ishizaka & Flanagan 1972). Liljencrants (1985) uses both.

3.3 The equation of state

The third aerodynamic equation is the equation of state, which relates the air pressure P to the mass density ρ of the air and thus couples the equations of continuity and motion.

If the processes that we are interested in, are so swift that we can neglect heat conduction in the fluid, the temperatures vary with the material pressure, and no air particles exchange any heat. In this case, the relation between pressure and mass density is given by the ***adiabatic*** pressure law

$$\frac{P}{P_0} = \left(\frac{\rho}{\rho_0}\right)^{\gamma} \qquad (3.26)$$

where P_0 and ρ_0 are a reference pressure and density, and γ is a constant of the fluid, equal to approximately 1.4 for a diatomic gas like air. If we choose for P_0 and ρ_0 the average pressure and density in vivo (no flow, no temperature gradient), a differential pressure change is given by

$$dP = P_0 \gamma \frac{d\rho}{\rho_0} \equiv c^2 d\rho \qquad (3.27)$$

where the constant c depends only on temperature and mean pressure and has the dimensions of a velocity. The value of this constant can be computed as 353 m/s, for P_0 = $1.013{\cdot}10^5$ N/m^2 and ρ_0 = 1.14 kg/m^3. If we combine (3.27) with the one-dimensional continuity equation (3.7) and the equation of motion (3.11), we can see that c is the velocity of the propagation of a sound wave.

For slower processes, there is time for the particles to exchange heat with each other and with the walls. In this case, the relation between the pressure and the density is given by Boyle's law, i.e., eqs. (3.26) and (3.27) with $\gamma = 1$. These ***isothermic*** processes do occur in speech: they are involved in building up the lung pressure under the glottis and the pressure behind a constriction in the vocal tract. All these pressures come out 40% too high. To correct this, we should take the temperature as our third aerodynamic variable (besides flow and pressure) and add the equations of heat convection and conduction to those of continuity and motion. A further benefit of this procedure would be the automatic inclusion of damping due to heat conduction. Nevertheless, we will refrain from complicating the physical model in this way, as long as our knowledge of the muscle tensions that should produce the isothermic pressures, is not accurate within 40%. This explains the two "no" entries in the last column of table 2.1; the "apx" in that table for the model by Liljencrants' refers to his approximation of the air/wall heat conduction loss as a constant factor of 45% of the viscous loss in the boundary layer (following Fant 1960), which we could simply include in our model by changing the factor 0.3 in (3.23) to 0.435.

The excess pressure ΔP is taken as the difference of the real pressure and the atmospheric pressure, so that for small pressures the equation of state is approximated by

$$\Delta P = P - \rho_0 c^2 = (\rho - \rho_0) c^2 \tag{3.28}$$

3.4 Turbulence

Chaotic air movements arise at interfaces where the air flows from a narrower into a wider tube, if the particle velocity in the narrower tube is high enough. This turbulence causes a loss of energy of the motion in the x direction, but at the same time it generates a noisy sound.

3.4.1 Energy loss

The energy loss causes a pressure drop in the direction of flow. If tube 1 is the narrower tube, and tube 2 the wider (figure 3.7), this pressure drop for velocities greater than a critical velocity v_{crit},, is

$$P_{turb} = \tfrac{1}{2} \rho_0 v_{1r} \left(|v_{1r}| - v_{crit} \right) \left(1 - \frac{A_1}{A_2} \right)^2 \tag{3.29}$$

This pressure drop can be looked at as a failure to recover completely from the Bernoulli (kinetic) pressure drop $\tfrac{1}{2} \rho_0 v_{1r}^2$. This equation is in accordance with equation (6) of Ishizaka & Flanagan (1972), which describes the pressure recovery if $v_{crit} = 0$. From

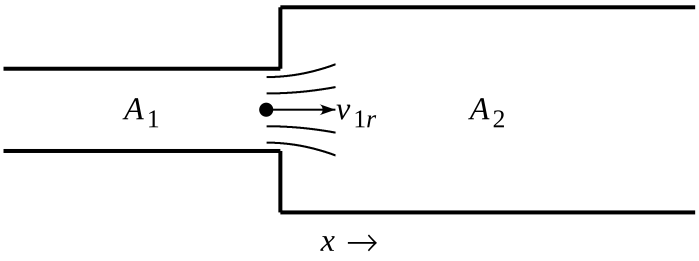


Fig. 3.7 Turbulent conditions exist when the velocity at the outlet of the narrower tube is high enough. The figure shows the idea of the related flow separation: the streamlines form a jet instead of bending around the corner.

Van den Berg, Zantema & Doornenbal (1957), the critical velocity can be computed as 10 m/s (critical volume velocity 200 cm^3/s, area 1.07 x 20 mm; for a larger opening, they found a greater critical volume velocity, which suggests a constant critical velocity).

The energy loss reveals itself as a discontinuity in the “continuous” pressure Q at the interface between the tubes.

Another way of describing the energy loss is representing it as an extra resistance term in the equation of motion inside the narrower tube section:

$$R_{turb} v = \frac{P_{turb}}{\Delta x_1} \tag{3.30}$$

where Δx_1 is the length of the tube.

Ishizaka & Flanagan (1972) use a similar resistance not only at the exit of the glottis, but also at the entrance of the glottis. The pressure drop there, however, is just the Bernoulli pressure for a ***vena contracta***, i.e., the stream is contracted and the area of the entrance is smaller than would be expected from the distance of the walls (though Ishizaka & Flanagan acknowledge this, they do not use this smaller area in their subsequent computations). Therefore, if the flow at the inlet is laminar, the “resistance” represents no energy loss, and the pressure loss is recovered somewhere in the glottis. In this case, the effect can be neglected, and (3.29) would be approximately right in predicting, for $A_1 = 0.1 \cdot A_2$, a turbulence loss of 0.81 relative to the Bernoulli pressure, comparing favourably with Van den Berg’s measured value of 0.875, as opposed to the value of 1.19 predicted by Ishizaka & Flanagan.

3.4.2 Turbulence noise

Thus, turbulence causes a pressure drop due to the loss of kinetic energy in the x-direction. This energy is converted into chaotic particle movements. Most of the energy is ultimately lost as heat, but some of it is radiated as sound. Meyer-Eppler (1953) finds that the sound source power is proportional to the square of the Reynolds number. The Reynolds number is $\rho v d / \mu$ (see e.g. Sommerfeld 1964), where d is a characteristic dimension. For a tube with circular cross section, d is the radius; as the source power is proportional to $(vd)^2$, it is also proportional to $v^2 A$, which is equivalent to the formula U^2/A (U is the volume velocity vA), which was used by Flanagan & Cherry (1969) and by Flanagan, Ishizaka & Shipley (1975) for friction noise, although the cross sections in their models were not thought of as being circular. Stevens (1971) also provides the source with a v^2 dependence. Badin (1989) and Scully, Castelli, Brearley & Shirt (1992) use the formula $(\Delta P)^{3/2} A^{1/2}$ for the source pressure (which is $v^6 A$ for the source power); they attribute this formula to Stevens (1971), though Stevens uses it as an expression for the power *radiated* through the lips, which has an additional v^2 dependence, because the radiated pressure is approximately the time derivative of the flow (and the pressure) inside the mouth (see eq. (3.50)). So a v^2 dependence of the noise pressure seems most realistic.

We implement the noise source by multiplying (3.29) by the factor

$$1 + 0.1 \cdot \mathbf{N}(t) \tag{3.31}$$

where the stochastic process $\mathbf{N}(t)$ is Gaussian white noise whose power is unity. This gives a v^2 dependence of the noise pressure, and a v^4 dependence of the noise power. In our implementation, a frequency cut-off is automatically caused as a side-effect of our method of integration (§3.11.2): if the longest tubes are approximately 1 cm long, the cut-off frequency of the integration is just above 6 kHz.

3.5 Boundary conditions

There are boundaries in space and time. The one boundary condition in time is that at the time point zero, all velocities are zero and the pressures equal the atmospheric pressure. The four types of boundary conditions in space refer to the four boundary types that were shown in figure 2.4 and are discussed in the following four subsections.

3.5.1 At a closed boundary

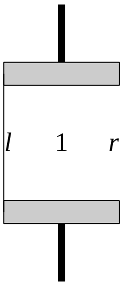


Fig. 3.8 A tube with a closed boundary on the left side.

The left boundary of the tube depicted in figure 3.8 is impenetrable for air. Therefore, the particle velocity in the x direction vanishes:

$$v_{1l} = 0 \tag{3.32}$$

Here and in the following, the subscripts 1, 2, and 3 refer to tube sections, and the subscripts l and r refer to the limit values found when approaching from the centre of a tube its left and right boundaries, respectively. The positive x direction always points from l to r.

3.5.2 At a boundary open to the atmosphere

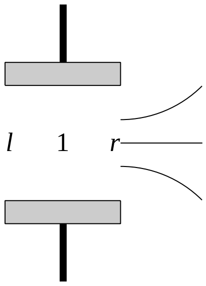


Fig. 3.9 A tube with a boundary that radiates sound into the surrounding atmosphere.

At the right boundaries of some tubes (lips, nostrils), sound radiates into the atmosphere (figure 3.9). Morse & Ingard (1968: eq. 7.4.31) derive an equation for the relation between the particle velocity and the pressure at such an interface: if the orifice is circular with a radius a and the wave is a harmonic oscillation with a radial frequency of $\omega < c/a$, this relation is

$$\frac{P}{v} = \rho c \left(\frac{1}{2} \left(\frac{\omega a}{c} \right)^2 - i \frac{8}{3\pi} \frac{\omega a}{c} \right) \tag{3.33}$$

which can be further approximated in the differential equation

$$\frac{\partial P}{\partial t} - \frac{128}{9\pi^2} \frac{\partial(\rho c v)}{\partial t} + \frac{16}{3\pi} \frac{Pc}{a} = 0 \tag{3.34}$$

In our case, we take a lip opening of at least 1 centimetre, in order not to get unrealistically low damping values for small lip openings.

3.5.3 At a boundary between two tube sections

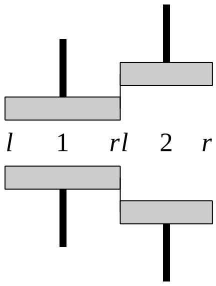


Fig. 3.10 A right and a left boundary forming an interface between two tube sections.

On the boundary between two tube sections with different cross sections (areas), many quantities are discontinuous. For example, the particle velocity suddenly increases when air flows from a wider tube into a narrower tube (e.g., from the right tube in fig. 3.10 into the left). This increase in velocity must be accompanied by a negative gradient (Bernoulli effect), so the pressure suddenly drops, and, as a consequence, the air density drops as well. However, there are two quantitities that are continuous at the interface between two tubes:

1. Mass flow. The conservation of mass ensures that the amount of air that leaves tube 1 in figure 3.10 at its right boundary during a certain time interval must enter the adjacent tube 2 at its left boundary. Thus, the *mass flow J*, which is a vector defined as

$$J \equiv \rho v A \tag{3.35}$$

and which is expressed in kilograms per second, is continuous at the interface between two adjacent tube sections:

$$\left(\rho v A\right)_{1r} = \left(\rho v A\right)_{2l} \tag{3.36}$$

2. Continuous pressure. The Bernoulli effect causes a pressure increase if air flows from a narrower tube into a wider tube. If the fluid flow stays laminar in crossing the boundary, the particle keeps moving in the x-direction, and we can define the scalar ***continuous pressure*** Q, expressed in Newtons per square meter, by

$$Q \equiv P + \tfrac{1}{2}\rho v^2 \tag{3.37}$$

This definition should include higher-order terms in v, but the fourth-order term is already less than 1% of the second-order term if $v < 0.2 \cdot c$, e.g., for velocities under 70 metres per second (which is in the order of the maximum particle velocity found in speech). The pressure Q is continuous at the interface between two adjacent tube sections:

$$\left(P + \tfrac{1}{2}\rho v^2\right)_{1r} = \left(P + \tfrac{1}{2}\rho v^2\right)_{2l} \tag{3.38}$$

If the flow is not laminar, but there is a pressure drop P_{turb} due to turbulence, the continuity relation becomes

$$Q_{1r} = Q_{2l} + P_{turb} \tag{3.39}$$

Conservation of momentum? Momentum is a conserved quantity in the hydrodynamics of unbounded fluids. However, it is not conserved at the interface between two adjacent tubes: the flowing air transfers momentum to the vertical walls, and vice versa, except in the case of complete flow separation (eq. 3.29, with v_{crit} = 0).

3.5.4 At a three-way boundary

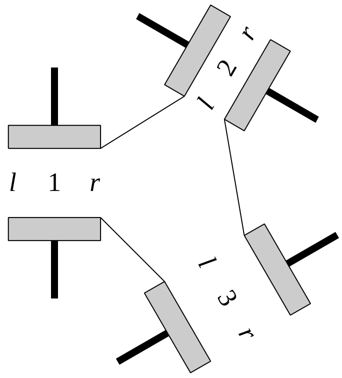


Fig. 3.11 An interface between three tube sections. The right boundary of section 1 is thought to be in immediate contact with the left boundaries of sections 2 and 3.

The extension from a two-tube interface to the three-tube interface shown in figure 3.11 is straightforward:

$$\begin{gathered}(\rho v A)_{1r} = (\rho v A)_{2l} + (\rho v A)_{3l} \\ \left(P + \tfrac{1}{2}\rho v^2\right)_{1r} = \left(P + \tfrac{1}{2}\rho v^2\right)_{2l} = \left(P + \tfrac{1}{2}\rho v^2\right)_{3l}\end{gathered} \tag{3.40}$$

In this case, the formulas for flow and pressure are not analogous. This difference reflects the vector character of the flow, and the scalar character of the pressure. These formulas are reminiscent of the Kirchhoff relations for electrical current and voltage.

3.6 Simplifying the aerodynamic equations

3.6.1 The aerodynamic equations in terms of continuous quantities

It is no coincidence that the mass flow J and the continuous pressure Q, which are continuous across the interfaces between tube sections, are exactly the quantities that appear in the divergence part of the continuity equation (3.4) and in the gradient part of the equation of motion (3.24). In terms of these continuous quantities, the aerodynamic equations read

$$\begin{gathered}\frac{\partial}{\partial t}\int \rho A dx = -\Delta_x J \\ \rho \frac{\partial v}{\partial t} = -\frac{\partial Q}{\partial x} - Rv \\ dP = c^2 d\rho\end{gathered} \tag{3.41}$$

3.6.2 Eliminating the equation of state

We can simplify the aerodynamic equations by eliminating the air mass density ρ from them. We write the integrand of the continuity equation as the ***line mass density*** e, expressed in kg/m, as a function of J and Q:

$$e \equiv \rho A = \frac{PA}{c^2} = \frac{QA}{c^2} - \frac{J^2}{2\rho_0 c^2 A} \tag{3.42}$$

The equation of motion must be rewritten so that ρ is brought inside the time derivative. This is easy if we assume that the air is incompressible. In that case, we can define the ***momentum density*** p, expressed in kg/m^2/s, and write it as a function of J and Q:

$$p \equiv \rho v = \frac{J}{A} \tag{3.43}$$

We can see that e and p are not continuous at the interface between two tubes, because they are the products of a continuous quantity (Q or J) and the discontinuous area A. Now we can write the *two* aerodynamic equations as

$$\begin{aligned} &\frac{\partial}{\partial t}\int e\,dx = -\Delta_x J \\ &\frac{\partial p}{\partial t} = -\frac{\partial Q}{\partial x} - \frac{R}{\rho_0}p \end{aligned} \tag{3.44}$$

The continuous quantities J and Q can be expressed in the non-continuous quantities e and p according to

$$\begin{aligned} &J = pA \\ &Q = \frac{ec^2}{A} + \frac{p^2}{2\rho_0} \end{aligned} \tag{3.45}$$

The boundary conditions in terms of the continuous quantities are simply

$$J = 0 \tag{3.46}$$

for the closed boundary, and the approximate

$$\frac{\partial Q}{\partial t} - \frac{128}{9\pi^2}\frac{\partial\left(\frac{cJ}{A}\right)}{\partial t} + \frac{16}{3\pi}\frac{Qc}{a} = 0 \tag{3.47}$$

for the open boundary.

3.6.3 A paradoxical factor of one half

Note that the pressure part of the equation of motion effectively reads

$$\frac{\partial(\rho v)}{\partial t} = -\frac{\partial\left(P + \frac{1}{2}\rho v^2\right)}{\partial x} \tag{3.48}$$

whereas one of the standard equations of hydrodynamics is derived from eqs. (3.13) and (3.7) as

$$\frac{\partial(\rho v)}{\partial t} = \rho\frac{\partial v}{\partial t} + v\frac{\partial \rho}{\partial t} = -\frac{\partial P}{\partial x} - (\rho v)\frac{\partial v}{\partial x} + v\cdot\left(-\frac{\partial(\rho v)}{\partial x}\right) = -\frac{\partial\left(P + \rho v^2\right)}{\partial x} \tag{3.49}$$

The refutation of the apparent contradiction involves the argument below equation (3.7).

3.7 Acoustic output

The aerodynamic and myoelastic differential equations are integrated by a finite-differencing method (described in detail in the following sections), where time is spliced into pieces with a fixed duration Δt., the ***sample period***.

The *state* of the system at a time $n\Delta t$ is defined by the dimensions of the tube sections, the velocities of the walls, the air flow, and the air pressure. The changes in this state between the times $n\Delta t$ and $(n{+}1)\Delta t$ are computed from the aerodynamic equations (§3.1–§3.6), the myoelastic equations (§2.3), and the articulation data (§2.5–12). The resulting acoustic sound pressure at a certain distance d from the lips and nostrils is computed from the time derivative of the flow J_r through the open boundaries (Morse & Ingard 1968: ch. 7; Flanagan 1972: eq. 3.40), and from the velocities of the moving walls, as follows:

$$sound(d) = \frac{4\pi}{d}\left(\sum_r \frac{dJ_r}{dt} + \sum_m 1000\rho_0 \Delta x_m \Delta z_m \frac{d\Delta y_m}{dt}\right) \qquad (3.50)$$

where the first term is a summation over all radiating tubes (r), and the second term is a summation over all tubes (m).

3.8 Digital simulation

In the rest of this chapter, we will show how both the myoelastic and aerodynamic equations are converted to difference equations suitable for computer simulation.

In our digital simulation of the speech apparatus, time is spliced into pieces with a fixed duration Δt. This *sampling period* must be smaller than the time needed for sound to travel the length of the shortest tube section, because that is the largest time that will guarantee a stable integration. In our model speakers, the shortest tube section is the one that represents the upper part of the vocal cords. If this has a length of 0.7 mm, and the velocity of sound in the vocal tract is 350 m/s, the minimum sampling frequency is 500 kHz. So, to simulate one second of speech, we have to compute 500,000 tract shapes, and if the model speaker has 80 tube sections, we have to compute 40 million times the air streams and pressures inside a tube and on the interfaces with its neighbours.

For most of the simulations presented here, the sampling frequency of the resulting sound was 22,050 Hz, which is one of the standard sampling frequencies of our Silicon Graphics Indigo computers, and half of the sampling frequency of a Compact Disk. The entire vocal-tract configuration (equilibrium values) was computed 22,050 times for one second of output, by the method that was described in §2.5 to §2.12. For every computation of one output sample (and one vocal-tract configuration), the aerodynamics and myoelastics were computed 25 times by the methods of §2.3 and §3.1 to §3.6, resulting in an internal sampling frequency of 551,250 Hz. On an Indigo built in 1991, this requires a thousand seconds of computer time for one second of speech.

The slowness of the simulation makes our method unsuitable for use in text-to-speech systems, but our method can cope with a lot of speech phenomena that pose insuperable problems to the faster algorithms. In §3.9 to §3.11, we will tackle the numerical integration of the three mathematical types of differential equations that appear in our model: dissipative, harmonic, and hyperbolic equations. In §3.12, we present the complete algorithm for the integration of the aerodynamic and myoelastic differential equations.

3.9 The dissipative part of the equations

The aerodynamic equation of motion (3.44) and the myoelastic equation of motion (2.1 in combination with 2.10) contain a part of the type

$$\frac{dv(t)}{dt} = -Rv(t) \tag{3.51}$$

where v is a flow or a velocity and R is a positive number representing resistance or damping. For a constant resistance, the solution of this differential equation is

$$v(t) = v(0)e^{-Rt} \tag{3.52}$$

which is an exponentially damped motion with a decay time (the time during which the velocity decreases by a factor e) of $1/R$ seconds. In general, however, R is an unpredictable function of time, so that (3.51) can only be solved numerically.

The evolution of v between the times $n\Delta t$ and $(n+1)\Delta t$ can be approximated in a number of ways with varying degrees of accuracy, stability, and computing effort.

3.9.1 The exponential method

From (3.52), we can derive the solution (we use superscripts as time indices)

$$v^{n+1} = v^n\, e^{-R\Delta t} \tag{3.53}$$

which is exact for constant R. Otherwise, R in (3.53) is evaluated at the old time $n\Delta t$, or at the mid-time $\left(n+\frac{1}{2}\right)\Delta t$ if R is known at that time. Formula (3.53) has three desirable properties that are true to the underlying physics of damping:

(1) The absolute value of the multiplication factor $e^{-R\Delta t}$ is always less than 1, so that the absolute value of the velocity diminishes over time. The fact that the absolute velocity does not blow up is called *stability* in the sense of Von Neumann (Press, Flannery, Teukolsky & Vetterling 1989).

(2) The higher the resistance, the greater the fractional decrease in the velocity during a sampling period.

(3) The multiplication factor is always positive, so that the velocity does not change sign.

Moreover, the decay time t_{decay} associated with the exponential method is equal to the true decay time:

$$\frac{1}{e} = \left(e^{-R\Delta t}\right)^{\frac{t_{decay}}{\Delta t}} \qquad \rightarrow \qquad t_{decay} = \frac{1}{R} \tag{3.54}$$

The one disadvantage of the exponential method is the relatively large cost of computing an exponential function (ten multiplications or so). However, for small and moderate values of R, the multiplication factor $e^{-R\Delta t}$ is very near to 1, so that it can be replaced by any of the following three approximations, the first two of which are first-order accurate, and the last of which is second-order accurate:

$$e^{-R\Delta t} \approx 1 - R\Delta t \approx \frac{1}{1+R\Delta t} \approx \frac{1-\frac{1}{2}R\Delta t}{1+\frac{1}{2}R\Delta t} \tag{3.55}$$

These approximations represent the three differencing schemes that follow from the three most natural ways of directly rewriting the differential equation (3.51) as a difference equation, without knowledge of the special-case solution (3.52). Though we could deal with the integration of the simple equation (3.51) in a few words, we will look at these three methods in detail, because we will encounter some of the same problems and terminology when tackling the harmonic and hyperbolic equations.

3.9.2 The first-order explicit method

The differential equation (3.51) can be rewritten as the difference equation

$$\frac{v^{n+1} - v^n}{\Delta t} = -Rv^n \tag{3.56}$$

This differencing scheme is called ***first-order explicit*** or ***forward Euler*** (Press, Flannery, Teukolsky & Vetterling 1989) because the velocity in the damping term is evaluated at the old time $n\Delta t$. During our digital simulation, which follows the course of time, the old velocity v^n is known, but the new velocity v^{n+1} is not. We can solve equation (3.56) as

$$v^{n+1} = v^n \left(1 - R\Delta t\right) \tag{3.57}$$

Usually, $R\Delta t$ will be much less than 1. But how will this method of integration behave if $R\Delta t$ happens to be large? For resistances R greater than $1/\Delta t$, the multiplication factor (i.e., the factor by which v is multiplied during each time step) is negative, violating desirable property (3). For resistances between $1/\Delta t$ and $2/\Delta t$, the multiplication factor increases with increasing R, violating desirable property (2). And for resistances greater than $2/\Delta t$, the absolute value of the multiplication factor is greater than 1, which makes this method unstable (violating desirable property 1).

The decay time associated with this method is given by

$$\frac{1}{e} = \left(\left|1 - R\Delta t\right|\right)^{\frac{t_{decay}}{\Delta t}} \qquad \rightarrow \qquad t_{decay} = \frac{-\Delta t}{\ln\left|1 - R\Delta t\right|} \tag{3.58}$$

which is zero for $R = \frac{1}{\Delta t}$, and negative (catastrophe) for $R > \frac{2}{\Delta t}$. Figure 3.13 shows the multiplication factor and the decay time as a function of $R\Delta t$.

3.9.3 The first-order implicit method

A better way to write the difference equation is

$$\frac{v^{n+1} - v^n}{\Delta t} = -Rv^{n+1} \tag{3.59}$$

This differencing scheme is called ***first-order implicit*** or ***backward Euler*** (Press, Flannery, Teukolsky & Vetterling 1989) because the velocity in the damping term is

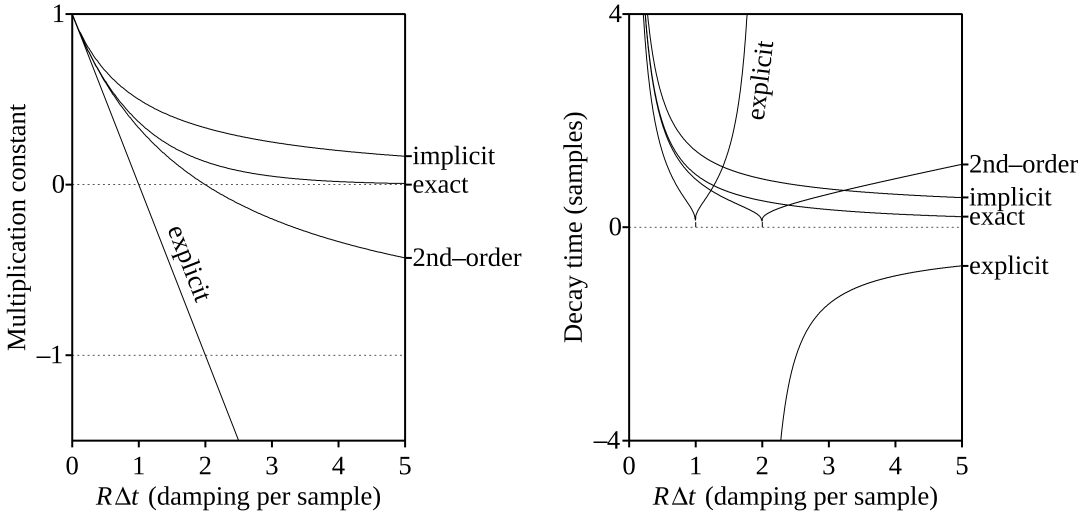


Fig. 3.13 The multiplication constant per sample, and the decay time in samples, for high resistances in the damping equation. As is seen from the figure, only the implicit and exact methods show behaviour that is both stable (absolute multiplication constant less than 1) and physically correct (multiplication constant positive and monotonically falling with $R\Delta t$).

evaluated at the new time $(n+1)\Delta t$. Fortunately, we can easily solve (3.59) by rearranging the terms in v^{n+1} and v^n:

$$v^{n+1} = \frac{v^n}{1 + R\Delta t} \tag{3.60}$$

The multiplication factor now fulfils all three desirable properties; however, it is larger than the multiplication factor in the exponential solution, so that the decay time simulated with the first-order implicit method is greater than the true decay time:

$$t_{decay} = \frac{\Delta t}{\ln(1 + R\Delta t)} > \frac{1}{R} \tag{3.61}$$

3.9.4 The second-order method

Instead of using the velocity at the old or new time as representative of the velocity between the times $n\Delta t$ and $(n+1)\Delta t$, it is more accurate to use the average of these two velocities. Thus, the differential equation (3.51) is rewritten as

$$\frac{v^{n+1} - v^n}{\Delta t} = -R\frac{v^n + v^{n+1}}{2} \tag{3.62}$$

This can be solved to give

$$v^{n+1} = v^n \frac{1 - \frac{1}{2}R\Delta t}{1 + \frac{1}{2}R\Delta t} \tag{3.63}$$

which is second-order accurate in $R\Delta t$. The absolute value of the multiplication factor is always less than 1, which makes this method stable in the Von Neumann sense. However, it shows unphysical behaviour for $R > \frac{2}{\Delta t}$: the multiplication becomes negative and its absolute value increases with increasing *R*. For very large resistances, the multiplication factor approaches –1, instead of 0, and the decay time approaches infinity, also instead of 0:

$$t_{decay} = \frac{\Delta t}{\ln\left|\frac{1+\frac{1}{2}R\Delta t}{1-\frac{1}{2}R\Delta t}\right|} \tag{3.64}$$

3.9.5 Which method should we use?

Which of the three approximations is the best one for our purposes? Figure 3.13 shows the multiplication factors and simulated decay times as functions of *R*Δ*t*. It is seen that, though the second-order method is the most accurate for moderate decay times, the first-order implicit method is the only one of the three approximations that honours the three desirable properties; moreover, for this method, large deviations are only found for decay times much smaller than one sampling period. Therefore, **we will use the first-order implicit differencing scheme for all resistances and dampings in our model**.

3.10 The harmonic part of the myo-elastic equations

The myoelastic equation of motion (2.1) contains a harmonic part (the first term of 2.2). The basic form of a harmonic differential equation reads

$$\frac{d^2y(t)}{dt^2} = -\omega^2 y(t) \tag{3.65}$$

For constant *ω*, we can find the exact solution:

$$y(t) = A\cos(\omega t + \varphi) \tag{3.66}$$

which is a harmonic oscillation with amplitude *A*, frequency $\omega/2\pi$ and starting phase *φ*, determined by the boundary values *y*(0) and $dy(0)/dt$. In general, however, *ω* depends on time in an unpredictable way and we have to integrate (3.65) numerically, preferably with a method that has the following properties:

1. All simulated solutions are stable.
2. If *ω* is constant, the simulated solution is periodic with little damping.
3. The angular oscillation frequency ω' of these simulated periodic solutions is a monotonically increasing function of *ω*.
4. The initial conditions $y(0) = 1$ and $\dot{y}(0) = 0$ represent a spring that is held fixed in the position 1 and released at the time $t = 0$. Therefore, the simulation with these initial conditions should show a harmonic motion with an amplitude of 1.

The second-order differential equation (3.65) is first split up into the two coupled first-order differential equations

$$\frac{dy(t)}{dt} = \dot{y}(t) \qquad ; \qquad \frac{d\dot{y}(t)}{dt} = -\omega^2 y(t) \tag{3.67}$$

As with the resistance equation, we have four strategies at our disposal for the numerical integration of these equations. In contrast with the resistance equation, however, all these strategies are second-order accurate.

3.10.1 The "explicit" method

We can rewrite the differential equations (3.67) with the finite-difference approximations

$$\begin{aligned} \dot{y}^{n+\frac{1}{2}} &= \dot{y}^{n-\frac{1}{2}} - \omega^2\, \Delta t\, y^n \\ y^{n+1} &= y^n + \dot{y}^{n+\frac{1}{2}}\, \Delta t \end{aligned} \tag{3.68}$$

These formulae look well-enough balanced in time, they are second-order accurate, and there seems to be no overt forward integration. We will show that the difference equations (3.68) have a periodic solution if ω is constant. If we try the solution

$$y^n = \cos \omega' n \Delta t \qquad \text{(all } n\text{)} \tag{3.69}$$

we see from the second part of (3.68) that

$$\dot{y}^{n+\frac{1}{2}} = \frac{y^{n+1} - y^n}{\Delta t} = \frac{\cos \omega'(n+1)\Delta t - \cos \omega' n \Delta t}{\Delta t} = \frac{-2 \sin \omega'\left(n + \frac{1}{2}\right)\Delta t \sin \frac{1}{2}\omega'\Delta t}{\Delta t} \tag{3.70}$$

Because this equation is valid for all n, we can compute

$$\begin{aligned} \dot{y}^{n+\frac{1}{2}} - \dot{y}^{n-\frac{1}{2}} &= \frac{-2 \sin \frac{1}{2}\omega'\Delta t\left(\sin \omega'\left(n+\frac{1}{2}\right)\Delta t - \sin \omega'\left(n-\frac{1}{2}\right)\Delta t\right)}{\Delta t} = \\ &= \frac{-2 \sin \frac{1}{2}\omega'\Delta t\left(2 \cos \omega' n \Delta t \sin \frac{1}{2}\omega'\Delta t\right)}{\Delta t} = \frac{-4 \sin^2 \frac{1}{2}\omega'\Delta t}{\Delta t} y^n \end{aligned} \tag{3.71}$$

According to the first part of (3.68), this must be equal to $-\omega^2 \Delta t\, y^n$. This is true if

$$\omega' = \frac{\arcsin \frac{1}{2}\omega\Delta t}{\frac{1}{2}\Delta t} \tag{3.72}$$

Thus, the simulated solution to (3.65) is an undamped harmonic oscillation with an angular frequency ω', given by (3.72), that is different from the angular frequency ω of the real solution (3.66). Figure 3.14 (on the left, curve "explicit") shows the dependency of ω' on ω. We notice an appreciable ***frequency warping*** for high frequencies; the angular frequency for which $\frac{1}{2}\omega\Delta t = 1$, i.e., a periodicity frequency equal to the sampling frequency divided by π, is mapped on the Nyquist frequency, i.e., the sampling frequency divided by 2. For higher ω, the integration is unstable, thus violating desirable property (1).

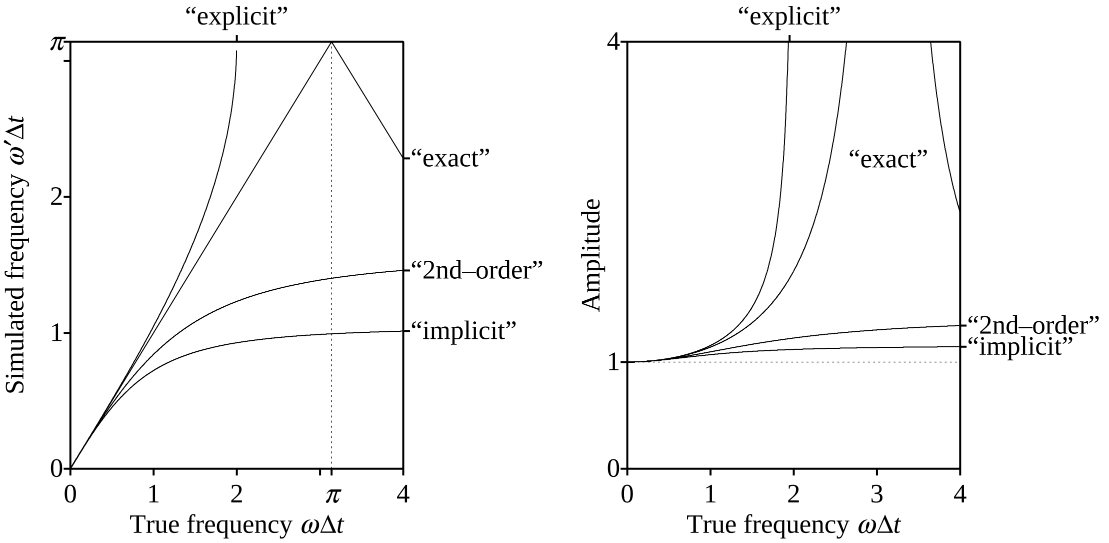


Fig. 3.14 The simulated angular frequency of periodicity, and the simulated amplitude of the motion relative to the real amplitude, for high spring constants in the harmonic equation. Only the implicit and second-order methods show both stable and physically plausible behaviour, apart from the appreciable frequency warping.

3.10.2 The "exact" method

We can make the relationship between ω' and ω exact, if we replace the multiplication factor in (3.68):

$$\omega^2 \Delta t \quad \to \quad \frac{4 \sin^2 \frac{1}{2} \omega \Delta t}{\Delta t} = \frac{2(1 - \cos \omega \Delta t)}{\Delta t} \tag{3.73}$$

Figure 3.14 (left) shows the relationship between ω' and ω. For $\omega \Delta t > \pi$, i.e., for underlying frequencies just above the Nyquist frequency, the frequency of periodicity decreases linearly with increasing ω. This ***aliasing*** is an unwanted side effect of the "exact" method, violating desirable property (3).

3.10.3 The "implicit" method

The multiplication factor for the "explicit" method can be rewritten as

$$-\omega^2 \Delta t^2 = \left(1 - \omega^2 \Delta t^2\right) - 1 \tag{3.74}$$

where the expression between parentheses reminds us of the multiplication factor of the explicit method for the resistance equation (3.57). This suggests that we use as the multiplication factor for our "implicit" method the expression

$$\frac{1}{1 + \omega^2 \Delta t^2} - 1 = \frac{-\omega^2 \Delta t^2}{1 + \omega^2 \Delta t^2} \tag{3.75}$$

As with the resistance equation, this "implicit" multiplication factor honours all four desirable properties. However, as we can see in figure 3.14, the frequency warping is quite strong:

$$\omega' = \frac{2}{\Delta t} \arcsin \tfrac{1}{2} \sqrt{\frac{\omega^2 \Delta t^2}{1 + \omega^2 \Delta t^2}} \tag{3.76}$$

giving a maximum periodicity frequency equal to one third of the Nyquist frequency.

3.10.4 The "second-order" method

With the resistance equation, subsequent values of the solution should not change sign; this is why we could not use the second-order method to simulate damping. With the harmonic equation, the restriction is looser: the multiplication factor must not be less than –2. Analogously to the "implicit" method, the multiplication factor suggests itself as

$$\frac{1 - \frac{1}{2}\omega^2 \Delta t^2}{1 + \frac{1}{2}\omega^2 \Delta t^2} - 1 = \frac{-\omega^2 \Delta t^2}{1 + \frac{1}{2}\omega^2 \Delta t^2} \tag{3.77}$$

This multiplication factor is a monotonic function of ω, approaching –2 for very large ω. The "second-order" method has all four properties that we desired, though it does show a certain amount of frequency warping, as shown in figure 3.14:

$$\omega' = \frac{2}{\Delta t} \arcsin \tfrac{1}{2} \sqrt{\frac{\omega^2 \Delta t^2}{1 + \frac{1}{2}\omega^2 \Delta t^2}} \tag{3.78}$$

giving a maximum periodicity frequency equal to one half of the Nyquist frequency.

3.10.5 The amplitude of the periodic motion

If we try to model with our system a spring that is held fixed out of equilibrium and released at the time $t = 0$, the initial conditions for our difference equations are

$$y^1 = 1 \qquad ; \qquad \dot{y}^{\frac{1}{2}} = 0 \tag{3.79}$$

According to desirable property (4), a simulation starting from these values should yield a periodic motion with an amplitude not much different from 1. The amplitude depends on the manifest angular frequency ω' as

$$amplitude = \left| \frac{1}{\cos \frac{1}{2} \omega' \Delta t} \right| \tag{3.80}$$

The right-hand side of figure 3.14 shows this amplitude as a function of the normalized underlying frequency $\omega \Delta t$, for all four integration methods. The "explicit" method yields

$$amplitude = \frac{1}{\cos \arcsin \frac{1}{2} \omega \Delta t} = \frac{1}{\sqrt{1 - \frac{1}{4}\omega^2 \Delta t^2}} \tag{3.81}$$

which is only defined for $\omega\Delta t < 2$. The “exact” method gives

$$amplitude = \left| \frac{1}{\cos \frac{1}{2}\omega\Delta t} \right| \tag{3.82}$$

which is infinite for $\omega\Delta t = (2n+1)\pi$. The “implicit” method gives

$$amplitude = \frac{1}{\cos \arcsin \frac{1}{2} \sqrt{\dfrac{\omega^2\Delta t^2}{1+\omega^2\Delta t^2}}} = \sqrt{\frac{1+\omega^2\Delta t^2}{1+\frac{3}{4}\omega^2\Delta t^2}} \tag{3.83}$$

which has a maximum of $\frac{2}{\sqrt{3}}$. The “second-order” method gives

$$amplitude = \frac{1}{\cos \arcsin \frac{1}{2} \sqrt{\dfrac{\omega^2\Delta t^2}{1+\frac{1}{2}\omega^2\Delta t^2}}} = \sqrt{\frac{1+\frac{1}{2}\omega^2\Delta t^2}{1+\frac{1}{4}\omega^2\Delta t^2}} \tag{3.84}$$

which has a maximum of $\sqrt{2}$.

3.10.6 Which method should we use?

The best choice seems to be the “second-order” method, because:

1. It is stable, unlike the “explicit” method.
2. It shows no aliasing, in contrast with the “exact” method.
3. It has less frequency warping than the “implicit” method.

Because our myoelastic problem will not give rise to the very high frequencies discussed here, however, we can and will use the simpler “explicit” method. The purpose of §3.10 was not to find the best integration method for the myoelastic equations, but to pave the way for the analysis of the hyperbolic part of the aerodynamic equations, which do show the phenomena of instability and frequency warping in our synthesizer with the “low” sample rates (500 kHz) that we will use.

3.11 The hyperbolic part of the aerodynamic equations

The “acoustic” part of (3.44), with its two coupled equations, requires a third method of integration. The basic form of a one-dimensional wave equation reads

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{3.85}$$

where c is a positive constant. We can easily check that its general solution is

$$u = f(x-ct) + g(x+ct) \tag{3.86}$$

where *f* and *g* are arbitrary twice-differentiable functions. Equation (3.86) represents the summation of a wave *f* travelling to the right with velocity *c* and a wave *g* travelling to the left, also with velocity *c*.

The second-order differential equation (3.85) can be written as two coupled first-order diferential equations:

$$\frac{\partial u}{\partial t} = -c\frac{\partial p}{\partial x} \qquad ; \qquad \frac{\partial p}{\partial t} = -c\frac{\partial u}{\partial x} \tag{3.87}$$

where we introduced the auxiliary variable *p*. For most physical problems, the equations are put in a more general form than (3.87). For conservation laws, this more general form is

$$\frac{\partial \mathbf{u}(x,t)}{\partial t} = -\frac{\partial \mathbf{F}(x,t)}{\partial x} \tag{3.88}$$

where:

- **u** is a vector of conserved quantities: in our case (eq. 3.44), the "momentum" *p* and the "mass" *e*;
- **F** is a vector of fluxes: in eq. (3.44), it is the mass flow *J* and the continuous pressure *Q*.

Equation (3.88) can be integrated numerically with the ***Lax-Wendroff*** method (Mitchell 1969; Press, Flannery, Teukolsky & Vetterling 1989; LeVeque 1992; Hirsch 1988, 1990), which is a second-order accurate integration scheme. As we will see in §3.11.1 and §3.11.2, this method requires:

(1) that **F** be known as a function **F**(**u**) of **u**;
(2) that the distances Δx between the space points be equal and constant;
(3) that both **u** and **F** be continuous.

Though requirement (1) is met by equation (3.45), the physics of our problem refuses to meet requirement (2), and as *e* and *p* (eqs. 3.42 and 3.43) are discontinuous at tube interfaces, requirement (3) cannot be met either. Therefore, we will extend the Lax-Wendroff method in such a way that these two requirements are dropped (§3.11.3).

3.11.1 The Lax-Wendroff method

The basic Lax-Wendroff method (Richtmyer strategy) consists of the following steps (Mitchell 1969; Press, Flannery, Teukolsky & Vetterling 1989; Hirsch 1988, 1990), deriving the values at the new time $(n+1)\Delta t$ from the values at the old time $n\Delta t$, for a one-dimensional array of space points that can be thought of as interfaces between tube sections. The space points *m*–1 and *m* are a fixed distance Δx apart; the centre point $m-\frac{1}{2}$ can be thought of as the centre of a tube.

Step 1: approximate the old centre values of **u** by averaging in space:

$$\mathbf{u}^n_{m-\frac{1}{2}} = \tfrac{1}{2}\left(\mathbf{u}^n_{m-1} + \mathbf{u}^n_m\right) \tag{3.89}$$

This is called the Lax step and puts in some numerical viscosity.

Step 2: approximate the centre values of **u** at the half-way time $\left(n+\frac{1}{2}\right)\Delta t$, from the values at time $n\Delta t$, using first-order-accurate explicit integration:

$$\mathbf{u}_{m-\frac{1}{2}}^{n+\frac{1}{2}} = \mathbf{u}_{m-\frac{1}{2}}^{n} + \tfrac{1}{2}\Delta t \frac{\mathbf{F}_{m-1}^{n} - \mathbf{F}_{m}^{n}}{\Delta x} \tag{3.90}$$

This step *could* have been second-order accurate, according to

$$\mathbf{u}_{m-\frac{1}{2}}^{n+\frac{1}{2}} = \mathbf{u}_{m-\frac{1}{2}}^{n-\frac{1}{2}} + \Delta t \frac{\mathbf{F}_{m-1}^{n} - \mathbf{F}_{m}^{n}}{\Delta x} \tag{3.91}$$

thus giving the ***staggered-leapfrog*** method, but this method can be unstable, and a simulation has shown that it would indeed be unstable in our case.

Step 3: compute the centre half-way values of the fluxes:

$$\mathbf{F}_{m-\frac{1}{2}}^{n+\frac{1}{2}} = \mathbf{F}\left(\mathbf{u}_{m-\frac{1}{2}}^{n+\frac{1}{2}}\right) \tag{3.92}$$

Step 4: use these half-way values for approximating the values of **u** at the boundaries at the new time $(n+1)\Delta t$ to second-order accuracy:

$$\mathbf{u}_{m}^{n+1} = \mathbf{u}_{m}^{n} + \Delta t \frac{\mathbf{F}_{m-\frac{1}{2}}^{n+\frac{1}{2}} - \mathbf{F}_{m+\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} \tag{3.93}$$

Step 5: finally, compute the new values of the fluxes at the space points:

$$\mathbf{F}_{m}^{n+1} = \mathbf{F}\left(\mathbf{u}_{m}^{n+1}\right) \tag{3.94}$$

3.11.2 Stability, numerical damping, and frequency warping

In this section, we shall determine to what extent the Lax-Wendroff method is capable of simulating a simple one-dimensional sinusoidal wave that travels from left to right. A solution to (3.87) is

$$u(x,t) = p(x,t) = e^{ik(x-ct)} \tag{3.95}$$

where k is the *wave number* (2π divided by the wavelength). Because we will be simulating a system of ducts with a known shape, we are interested in the performance of the algorithm for known wavelengths. The resonances (formants) that we will simulate in our finite-difference approach, will typically have the correct wavelengths, but may have the wrong frequencies because the speed of propagation in our disctretized system may be different from the real velocity of sound. Thus, let’s try to solve (3.95) as

$$u_{m}^{n} = p_{m}^{n} = e^{ik(m\Delta x - nc'\Delta t)} \tag{3.96}$$

which is a wave with a correct wavelength but with a different velocity c'. We must note that (3.96) is valid for all mesh positions $m\Delta x$, but only for one moment in time: $n\Delta t$. In other words, we will require a solution that evolves as

$$u_m^{n+1} = Ae^{-ikc'\Delta t}u_m^n \tag{3.97}$$

where A is a real constant (the absolute multiplication factor), which is not necessarily equal to 1. We will consider the method usable if our solution of (3.97) more or less features the following desirable properties:

(1) Stability: the solution is unstable (exponentially increasing in amplitude) if A is greater than 1; therefore, we require $A \le 1$
(2) Evolution: if A is small, our solution dies out quickly. Therefore, we would like to have $A \approx 1$.
(3) Faithfulness: the solution should resemble the solution of the underlying physical problem. In our case, we would like to have a faithful velocity $c' \approx c$, so that resonances in our tubes show up with the correct frequencies.

We will now determine the values of A and c' as functions of k, Δx, c, and Δt. Step 1 of the Lax-Wendroff method (eq. 3.89) gives

$$u_{m-\frac{1}{2}}^n = \tfrac{1}{2}\left(u_{m-1}^n + u_m^n\right) = \cos\tfrac{1}{2}k\Delta x \cdot e^{ik\left(\left(m-\frac{1}{2}\right)\Delta x - nc'\Delta t\right)} \tag{3.98}$$

Step 2 (3.90) gives (making use of the fact that (3.96) goes for every m)

$$u_{m-\frac{1}{2}}^{n+\frac{1}{2}} = u_{m-\frac{1}{2}}^n + \tfrac{1}{2}\alpha\left(p_{m-1}^n - p_m^n\right) = \left(\cos\tfrac{1}{2}k\Delta x - \alpha\sin\tfrac{1}{2}k\Delta x\right)e^{ik\left(\left(m-\frac{1}{2}\right)\Delta x - nc'\Delta t\right)} \tag{3.99}$$

where $\alpha \equiv c\Delta t/\Delta x$ is the sampling period relative to the time in which the wave travels from one mesh point to the next. Since the expression for $p_{m-\frac{1}{2}}^{n+\frac{1}{2}}$ must be identical to (3.99), step 4 (3.93) gives

$$u_m^{n+1} = u_m^n + \alpha\left(p_{m-\frac{1}{2}}^{n+\frac{1}{2}} - p_{m+\frac{1}{2}}^{n+\frac{1}{2}}\right) = \left(1 - \alpha\left(\cos\tfrac{1}{2}k\Delta x - \alpha i\sin\tfrac{1}{2}k\Delta x\right)2i\sin\tfrac{1}{2}k\Delta x\right)u_m^n =$$

$$= \left(1 - \alpha^2(1-\cos k\Delta x) - i\alpha\sin k\Delta x\right)u_m^n \tag{3.100}$$

This satisfies (3.97) if

$$1 - \alpha^2(1-\cos k\Delta x) - i\alpha\sin k\Delta x = Ae^{-ikc'\Delta t} \tag{3.101}$$

We can easily see that if the sampling period is chosen equal to the time in which the wave propagates from one mesh point to the next ($c\Delta t = \Delta x$), the multiplication factor A equals 1 and the simulated velocity of propagation c' equals the true velocity c: the ideal situation. Nevertheless, we will have to look into the behaviour for other sampling periods, because our synthesizer works with a non-uniform mesh.

Equation (3.101) really consists of two equations, from which we have to solve A and c'. We solve A from the absolute value of the left-hand side:

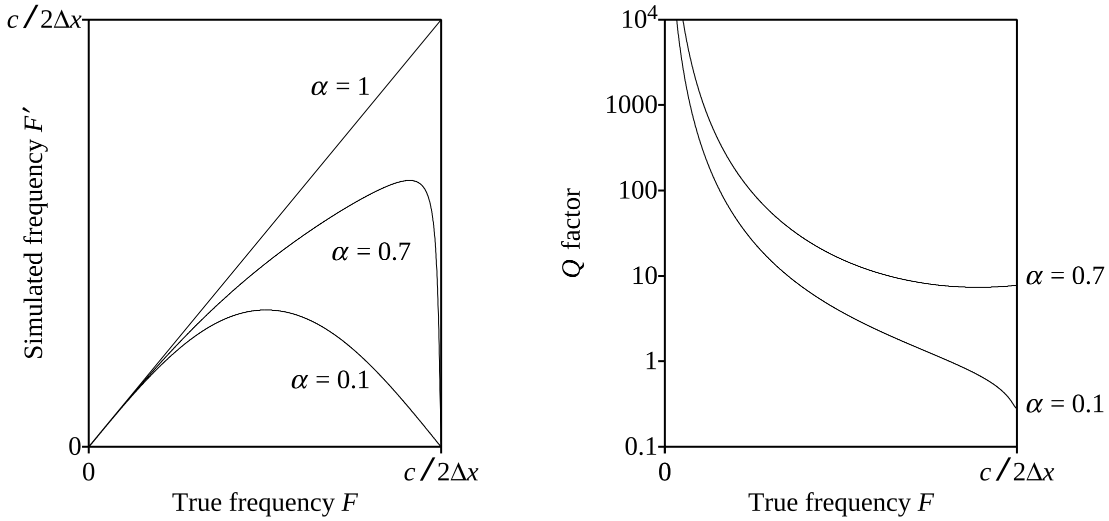


Fig. 3.15 Frequency warping and numerical damping in the Lax-Wendroff method.

$$A^2 = \left(1-\alpha^2(1-\cos k\Delta x)\right)^2 + \alpha^2 \sin^2 k\Delta x = 1-\alpha^2\left(1-\alpha^2\right)(1-\cos k\Delta x)^2 \tag{3.102}$$

which is equivalent to Hirsch’s (1988) formula (E8.3.14) for the dissipation (or diffusion) error of the Lax-Wendroff scheme for the convection equation.

We solve c' from the argument (angle) of the left-hand side:

$$c' = \frac{1}{k\Delta t}\arctan\frac{\alpha \sin k\Delta x}{1-\alpha^2(1-\cos k\Delta x)} \tag{3.103}$$

which is equivalent to Hirsch’s (1988) formula (E8.3.15) for the phase (or dispersion) error of the Lax-Wendroff scheme for the convection equation.

We can now assess our three desirable properties,

Stability. Stability is ensured if $A \le 1$, i.e., if $\alpha \le 1$. For longer sampling periods, the method is unconditionally unstable.

Frequency behaviour. The sinusoidal wave has an underlying frequency $F = kc/2\pi$. Because the velocity of propagation is simulated incorrectly (3.103), the simulated frequency will underestimate the real frequency:

$$F' = F\frac{c'}{c} = \frac{1}{2\pi\Delta t}\arctan\frac{\alpha \sin 2\pi F\Delta x/c}{1-\alpha^2(1-\cos 2\pi F\Delta x/c)} \tag{3.104}$$

The left-hand side of figure 3.15 shows the dependency of F' on F, for several relative sampling periods $\alpha = c\Delta t/\Delta x$.

Numerical damping. The multiplication factor A is valid for every consecutive sample. This causes an exponential decrease in the amplitude of the wave. The numerical damping of a sinusoid can be characterized by the ***bandwidth*** B, according to

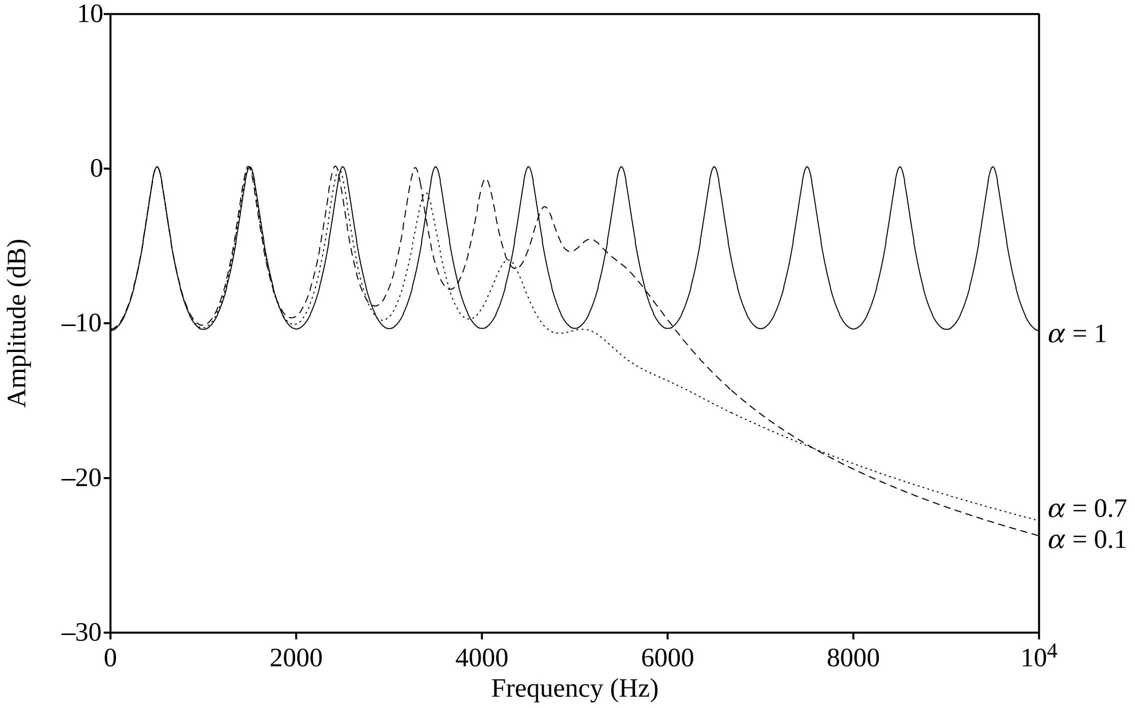


Fig. 3.16 Frequency warping and numerical damping in a Lax-Wendroff integration with three different sampling periods. The velocity of propagation is 353 m/s; the mesh length is 1 cm.

$$e^{-\pi Bt} = A^{t/\Delta t} \to B = \frac{-\ln A}{\pi \Delta t} = \frac{-1}{2\pi\Delta t}\ln\left(1-\alpha^2\left(1-\alpha^2\right)\left(1-\cos 2\pi F\Delta t/\alpha\right)^2\right) \qquad (3.105)$$

The right-hand side of figure 3.15 shows the Q factor ($Q \equiv F/B$) for several relative sample periods.

For our model, these results determine an upper bound on the tube lengths. Fig. 3.16 shows the theoretical distortion of the simulation of a vocal-tract-like tube (formants at 500, 1500, 2500, ..., 11500 Hz with a constant bandwidth of 100 Hz for every formant), with a sound velocity c of 353 m/s and a section length Δx of 1 cm. We see an appreciable frequency warping, increasing with the sampling frequency, and a larger bandwidth (this effect is strongest for $\alpha = \frac{1}{2}\sqrt{2}$).

3.11.3 Four extensions to the Lax-Wendroff method

In our case, the Lax-Wendroff method needs some modifications, because:

1. the lengths Δx of the tube sections are not equal;
2. these lengths are not constant either;
3. the quantities **u** ("mass" and "momentum") are not continuous at tube boundaries;
4. the quantity Q (belonging to **F**) is not continuous at boundaries with turbulence.

This leads to the following steps. We shift the spatial coordinate system so that the subscript index $m = 1 \ldots M$ is always a whole number that refers to the m-th tube section.

Step 1a: the quantities **u** are not continuous at the tube boundaries. The left- and right-limit values of **u** at the boundary between the m-th and (m+1)-th tube section are denoted by $\mathbf{u}_{m+}$ and $\mathbf{u}_{m+1-}$ respectively. They have to be computed from **F** at tube boundaries, which requires that we know the functions $\mathbf{u}_+(\mathbf{F})$ and $\mathbf{u}_-(\mathbf{F})$:

$$\mathbf{u}_{m-}^{n} = \mathbf{u}_-\left(\mathbf{F}_{m-}^{n}\right) \qquad ; \qquad \mathbf{u}_{m+}^{n} = \mathbf{u}_+\left(\mathbf{F}_{m+}^{n}\right) \tag{3.106}$$

Note that this notation allows for discontinuities in **F** as well as in **u**.

Step 1b: the original step 1 is now expressed as

$$\mathbf{u}_{m}^{n} = \tfrac{1}{2}\left(\mathbf{u}_{m-}^{n} + \mathbf{u}_{m+}^{n}\right) \tag{3.107}$$

Step 2: the lengths of the tubes are not equal, nor are they constant:

$$\mathbf{u}_{m}^{n+\frac{1}{2}} = \mathbf{u}_{m}^{n} + \tfrac{1}{2}\Delta t\,\frac{\mathbf{F}_{m-}^{n} - \mathbf{F}_{m+}^{n}}{\Delta x_{m}^{n}} \tag{3.108}$$

which is Lax-Wendroff. The staggered-leapfrog method would be

$$\mathbf{u}_{m}^{n+\frac{1}{2}} = \mathbf{u}_{m}^{n-\frac{1}{2}} + \Delta t\,\frac{\mathbf{F}_{m-}^{n} - \mathbf{F}_{m+}^{n}}{\Delta x_{m}^{n}} \tag{3.109}$$

Step 3 still computes the values of **F** from the values of **u** found in step 2, as in equation (3.92):

$$\mathbf{F}_{m}^{n+\frac{1}{2}} = \mathbf{F}\left(\mathbf{u}_{m}^{n+\frac{1}{2}}\right) \tag{3.110}$$

Steps 4 and 5 will have to be combined because of the discontinuities at the tube boundaries. Instead of the differential equation (3.88), we should start from the integral equation

$$\int \frac{\partial \mathbf{u}}{\partial t}\,dx = -\Delta_x \mathbf{F} \tag{3.111}$$

which is equivalent to (3.88), but makes the left-hand side continuous. Note that this equation looks like the hyperbolic part of the equation of motion (3.44b). Integrating (3.111) between the centres of two adjacent tubes yields

$$\begin{aligned}\mathbf{F}_{m}^{n+\frac{1}{2}} - \mathbf{F}_{m+1}^{n+\frac{1}{2}} &= \int_{\text{centre of } m\text{th tube}}^{\text{centre of } (m+1)\text{th tube}} dx\,\frac{\partial \mathbf{u}\left(x,\left(n+\frac{1}{2}\right)\Delta t\right)}{\partial t} = \\ &= \tfrac{1}{2}\Delta x_{m}^{n+\frac{1}{2}}\,\frac{\mathbf{u}_{m+}^{n+1} - \mathbf{u}_{m+}^{n}}{\Delta t} + \tfrac{1}{2}\Delta x_{m+1}^{n+\frac{1}{2}}\,\frac{\mathbf{u}_{m+1-}^{n+1} - \mathbf{u}_{m+1-}^{n}}{\Delta t}\end{aligned} \tag{3.112}$$

which leads to the implicit formula

$$\Delta x_{m}^{n+\frac{1}{2}}\,\mathbf{u}_{m+}^{n+1} + \Delta x_{m+1}^{n+\frac{1}{2}}\,\mathbf{u}_{m+1-}^{n+1} = \Delta x_{m}^{n+\frac{1}{2}}\,\mathbf{u}_{m+}^{n} + \Delta x_{m+1}^{n+\frac{1}{2}}\,\mathbf{u}_{m+1-}^{n} + 2\Delta t\left(\mathbf{F}_{m}^{n+\frac{1}{2}} - \mathbf{F}_{m+1}^{n+\frac{1}{2}}\right) \tag{3.113}$$

If, as in case of the mass-continuity equation, the length dx appears in the time derivand, the integral equation becomes (instead of 3.111)

$$\frac{\partial}{\partial t}\int \mathbf{u}dx = -\Delta_x \mathbf{F} \tag{3.114}$$

which is no longer equivalent to (3.88), does make the left-hand side continuous, and looks like the equation of conservation of mass (3.44a). Integrating (3.114) between the centres of two adjacent tubes yields

$$\begin{aligned} \mathbf{F}_m^{n+\frac{1}{2}} - \mathbf{F}_{m+1}^{n+\frac{1}{2}} &= \frac{\partial}{\partial t} \int_{\text{centre of } m\text{th tube}}^{\text{centre of } (m+1)\text{th tube}} dx\, \mathbf{u}\Big(x, \big(n+\tfrac{1}{2}\big)\Delta t\Big) = \\ &= \frac{\Big(\tfrac{1}{2}\Delta x_m^{n+1}\mathbf{u}_{m+}^{n+1} + \tfrac{1}{2}\Delta x_{m+1}^{n+1}\mathbf{u}_{m+1-}^{n+1}\Big) - \Big(\tfrac{1}{2}\Delta x_m^{n}\mathbf{u}_{m+}^{n} + \tfrac{1}{2}\Delta x_{m+1}^{n}\mathbf{u}_{m+1-}^{n}\Big)}{\Delta t} \end{aligned} \tag{3.115}$$

which leads to the implicit formula

$$\Delta x_m^{n+1}\, \mathbf{u}_{m+}^{n+1} + \Delta x_{m+1}^{n+1}\, \mathbf{u}_{m+1-}^{n+1} = \Delta x_m^{n}\, \mathbf{u}_{m+}^{n} + \Delta x_{m+1}^{n}\, \mathbf{u}_{m+1-}^{n} + 2\Delta t\left(\mathbf{F}_m^{n+\frac{1}{2}} - \mathbf{F}_{m+1}^{n+\frac{1}{2}}\right) \tag{3.116}$$

In the implementation of formulas (3.113) and (3.116), $\mathbf{u}_{m+}^{n+1}$ and $\mathbf{u}_{m+1-}^{n+1}$ will have to be written explicitly as functions of **F**, in a way that allows us to determine $\mathbf{F}_{m+}^{n+1}$ and $\mathbf{F}_{m+1-}^{n+1}$; we need a little luck for this to succeed.

If **F** has at the boundary a discontinuity that does not affect $\partial\mathbf{u}/\partial t$, e.g., a pressure drop that causes turbulence instead of acceleration, the last term in (3.113) or (3.116) has to be replaced by

$$2\Delta t\left(\mathbf{F}_m^{n+\frac{1}{2}} - \mathbf{F}_{m+}^{n+\frac{1}{2}} + \mathbf{F}_{m+1-}^{n+\frac{1}{2}} - \mathbf{F}_{m+1}^{n+\frac{1}{2}}\right) \tag{3.117}$$

where the middle two terms represent the discontinuity, which must be given or implied.

At the left and right edges of the tube array (m=0 and m=M), the integration can only be performed over half a tube length, so that instead of (3.113) and (3.116) we have

$$\begin{aligned} \Delta x_1^{n+\frac{1}{2}}\, \mathbf{u}_{1-}^{n+1} &= \Delta x_1^{n+\frac{1}{2}}\, \mathbf{u}_{1-}^{n} - 2\Delta t\left(\mathbf{F}_1^{n+\frac{1}{2}} - \mathbf{F}_{1-}^{n+\frac{1}{2}}\right) \\ \Delta x_M^{n+\frac{1}{2}}\, \mathbf{u}_{M+}^{n+1} &= \Delta x_M^{n+\frac{1}{2}}\, \mathbf{u}_{M+}^{n} + 2\Delta t\left(\mathbf{F}_M^{n+\frac{1}{2}} - \mathbf{F}_{M+}^{n+\frac{1}{2}}\right) \end{aligned} \tag{3.118}$$

and

$$\begin{aligned} \Delta x_1^{n+1}\, \mathbf{u}_{1-}^{n+1} &= \Delta x_1^{n}\, \mathbf{u}_{1-}^{n} - 2\Delta t\left(\mathbf{F}_1^{n+\frac{1}{2}} - \mathbf{F}_{1-}^{n+\frac{1}{2}}\right) \\ \Delta x_M^{n+1}\, \mathbf{u}_{M+}^{n+1} &= \Delta x_M^{n}\, \mathbf{u}_{M+}^{n} + 2\Delta t\left(\mathbf{F}_M^{n+\frac{1}{2}} - \mathbf{F}_{M+}^{n+\frac{1}{2}}\right) \end{aligned} \tag{3.119}$$

where the outermost boundary values $\mathbf{F}_{1-}^{n+\frac{1}{2}}$ and $\mathbf{F}_{M+}^{n+\frac{1}{2}}$ will have to be given or implied by boundary conditions.

3.11.4 Stability, frequency warping, and numerical damping

We can approximately translate the results of §3.11.2 for a non-uniform mesh as follows:

1. The method is stable if the sampling period is not greater than the time needed for a wave to travel the shortest tube. For instance, if our shortest tube is 0.7 mm long, and the velocity of sound is 350 m/s, the sample rate must be at least 500 kHz.
2. For high sample rates, a frequency cut-off is found at approximately one sixth of the inverse of the time needed for a wave to travel the longest tube. For instance, if our longest tube is 10 mm long, the cut-off frequency is just below 6 kHz, and underlying frequencies above 3 kHz will surface with a lower frequency.
3. For high sample rates, the numerical damping is small for frequencies below the cut-off frequency.

Thus, in our model we will have to work with a minimum sample rate of approximately 500 kHz, and a maximum tube length of approximately 7 mm.

3.11.5 Accuracy

If the two mesh lengths in (3.113) are different, the equation is not second-order accurate any longer: for a continuous system, the value of **F** in the centre of the shorter tube is probably a better predictor for the change in the value of **u** at the boundary than the value of **F** in the centre of the longer tube, because the former represents a situation closer to the boundary. In (3.113) and (3.116), by contrast, both values of **F** are given equal weight.

Thus, we can rewrite (3.112) as

$$\begin{aligned}
&\Delta x_{m+1}^{n+\frac{1}{2}}\left(\mathbf{F}_{m}^{n+\frac{1}{2}}-\mathbf{F}_{m+\frac{1}{2}}^{n+\frac{1}{2}}\right)-\Delta x_{m}^{n+\frac{1}{2}}\left(\mathbf{F}_{m+1}^{n+\frac{1}{2}}-\mathbf{F}_{m+\frac{1}{2}}^{n+\frac{1}{2}}\right)= \\
&=\Delta x_{m+1}^{n+\frac{1}{2}}\int_{\text{centre } m}^{\text{boundary}} dx\,\frac{\partial\mathbf{u}\left(x,\left(n+\frac{1}{2}\right)\Delta t\right)}{\partial t}+\Delta x_{m}^{n+\frac{1}{2}}\int_{\text{boundary}}^{\text{centre } m+1} dx\,\frac{\partial\mathbf{u}\left(x,\left(n+\frac{1}{2}\right)\Delta t\right)}{\partial t}= \\
&=\Delta x_{m}^{n+\frac{1}{2}}\Delta x_{m+1}^{n+\frac{1}{2}}\,\frac{\mathbf{u}_{m+}^{n+1}-\mathbf{u}_{m+}^{n}+\mathbf{u}_{m+1-}^{n+1}-\mathbf{u}_{m+1-}^{n}}{2\Delta t}
\end{aligned} \tag{3.120}$$

and (3.113) as

$$\begin{aligned}
\mathbf{u}_{m+}^{n+1}+\mathbf{u}_{m+1-}^{n+1}=\mathbf{u}_{m+}^{n}+\mathbf{u}_{m+1-}^{n}+ \\
+2\Delta t\,\frac{\Delta x_{m+1}^{n+\frac{1}{2}}\left(\mathbf{F}_{m}^{n+\frac{1}{2}}-\mathbf{F}_{m+\frac{1}{2}}^{n+\frac{1}{2}}\right)-\Delta x_{m}^{n+\frac{1}{2}}\left(\mathbf{F}_{m+1}^{n+\frac{1}{2}}-\mathbf{F}_{m+\frac{1}{2}}^{n+\frac{1}{2}}\right)}{\Delta x_{m}^{n+\frac{1}{2}}\Delta x_{m+1}^{n+\frac{1}{2}}}
\end{aligned} \tag{3.121}$$

This couple of equations would be more accurate than (3.112) and (3.113), if only we knew how to compute those **F** at the boundary.

3.12 The algorithm

This section describes how our finite-differencing methods implement the time evolution of the aerodynamics and myoelastics of our problem. The quantities **u** of the previous section are the "mass" *e* and the "momentum" *p*, and the fluxes **F** are the mass flow *J* and the continuous pressure Q. Diverging from eq. (3.42) for computational reasons, we introduce a new *e*, which is the old e multiplied by $\Delta x\, c^2$ and has the dimensions of an energy, so that at tube boundaries

$$\begin{aligned} J(e,p) &= pA \qquad ; \qquad Q(e,p) = \frac{e}{V} + \frac{p^2}{2\rho_0} \\ e_{m\pm}(J_{m\pm},Q_{m\pm}) &= \left(Q_{m\pm} - \frac{1}{2\rho_0}\left(\frac{J_{m\pm}}{A_m}\right)^2 \right) V_m \qquad ; \qquad p_{m\pm}(J_{m\pm},Q_{m\pm}) = \frac{J_{m\pm}}{A_m} \\ J^n_{m+1-} &= J^n_{m+} \qquad ; \qquad Q^n_{m+1-} = Q^n_{m+} + P^n_{turb,m,m+1} \end{aligned} \tag{3.122}$$

where $V \equiv A\Delta x$ is the volume of a tube.

The initial state of the system is defined as

$$\begin{aligned} J^0_{m\pm} &= 0 && \text{(no airflow)} \\ Q^0_{m\pm} &= \rho_0 c^2 && \text{(atmospheric pressure)} \\ \Delta x^0_m &= \Delta x^0_{eq,m} && \\ \Delta y^0_m &= \Delta y^0_{eq,m} && \text{(walls in equilibrium)} \\ \Delta \dot{y}^{-\frac{1}{2}}_m &= 0 && \text{(walls in rest)} \\ \Delta z^0_m &= \Delta z^0_{eq,m} && \end{aligned} \tag{3.123}$$

From this, we compute the cross sections A^0_m according to equations (2.4) and (2.5), and the starting values of the momentum densities *p* and kinetic pressures $K \equiv Q - P$ at tube boundaries, and the initial volumes *V* of the tube sections:

$$p^0_{m\pm} = 0 \qquad ; \qquad K^0_{m\pm} = 0 \qquad ; \qquad V^0_m = A^0_m\, \Delta x^0_m \tag{3.124}$$

For every sampling period *n*, starting at *n* = 0, we proceed by the following steps:

Step 1: the Lax step (3.89) averages the values for *e* and *p* inside the tubes from *J* and *Q* at the boundaries.

Step 1a: compute for every tube *m* (eq. 3.122):

$$e^n_{m\pm} = \left(Q^n_{m\pm} - K^n_{m\pm}\right) V^n_m \tag{3.125}$$

Step 1b: compute the mean "mass" and "momentum" for every tube *m* (3.107):

$$e^n_m = \tfrac{1}{2}\left(e^n_{m-} + e^n_{m+}\right) \qquad ; \qquad p^n_m = \tfrac{1}{2}\left(p^n_{m-} + p^n_{m+}\right) \tag{3.126}$$

The excess pressure ΔP, which we will need in the mass-spring equations, is from (3.42) and (3.28)

$$\Delta P_m^n = \frac{e_m^n}{V_m^n} - \rho_0 c^2 \tag{3.127}$$

The particle velocity v, which we will need when computing resistances and turbulence, is (combining 3.42 and 3.43):

$$v_m^n = \frac{p_m^n}{\rho_0 + \frac{\Delta P_m^n}{c^2}} \tag{3.128}$$

Step 1c: perform the integration of the myoelastic equation (2.1), combining the second-order-accurate "explicit" scheme (3.68) for the harmonic part with the first-order-accurate implicit scheme (3.60) for the damping (dissipative) term:

$$\begin{aligned} \dot{y}_m^{n+\frac{1}{2}} &= \frac{\dot{y}_m^{n-\frac{1}{2}} + \frac{\Delta t}{m_m^n}\left(tension_m^n + \Delta P_m^n \Delta z_m^n \Delta x_m^n\right)}{1 + \frac{B_m^n \Delta t}{m_m^n}} \\ y_m^{n+1} &= y_m^n + \dot{y}_m^{n+\frac{1}{2}} \Delta t \end{aligned} \tag{3.129}$$

where the tension is computed from eqs. (2.2), (2.9), and (2.15), and the damping B is computed from eqs. (2.11) to (2.14); the aerodynamic term is equation (2.16). Equation (3.129) is computed for both masses, and an analogous formula (with Δy_m^n in the pressure term) is used for z and $\dot{z}$. The new widths and depths are obtained by adding the displacements of each pair of masses:

$$\Delta y_m^{n+1} = y_{m,top}^{n+1} + y_{m,bottom}^{n+1} \quad ; \quad \Delta z_m^{n+1} = z_{m,left}^{n+1} + z_{m,right}^{n+1} \tag{3.130}$$

The new values of the cross section A_m^{n+1} are derived from these with the help of equations (2.4) and (2.5), and the half-way value of the cross section is interpolated as

$$A_m^{n+\frac{1}{2}} = \tfrac{1}{2}\left(A_m^{n+1} + A_m^n\right) \tag{3.131}$$

Step 1d: the half-way values of the tube lengths and volumes are now

$$\begin{aligned} \Delta x_m^{n+\frac{1}{2}} &= \tfrac{1}{2}\left(\Delta x_m^{n+1} + \Delta x_m^n\right) \\ V_m^{n+\frac{1}{2}} &= A_m^{n+\frac{1}{2}} \Delta x_m^{n+\frac{1}{2}} \end{aligned} \tag{3.132}$$

We can compute the resistances R_m^n from (3.24) and the resistance factors as

$$r_m^n \equiv \left(1 + \frac{R_m^n \, \Delta t}{\rho_0}\right) \frac{\Delta x_m^{n+\frac{1}{2}}}{A_m^{n+1}} \tag{3.133}$$

Step 2: the half-way aerodynamics inside tube *m* is

$$e_m^{n+\frac{1}{2}} = e_m^n + \tfrac{1}{2} c^2 \Delta t \left(J_{m-}^n - J_{m+}^n \right) \qquad ; \qquad p_m^{n+\frac{1}{2}} = \frac{p_m^n + \tfrac{1}{2} \Delta t \, \dfrac{Q_{m-}^n - Q_{m+}^n}{\Delta x_m^n}}{1 + \dfrac{1}{2\rho_0} \Delta t \, R_m^n} \tag{3.134}$$

Step 3:

$$J_m^{n+\frac{1}{2}} = p_m^{n+\frac{1}{2}} A_m^{n+\frac{1}{2}} \qquad ; \qquad Q_m^{n+\frac{1}{2}} = \frac{e_m^{n+\frac{1}{2}}}{V_m^{n+\frac{1}{2}}} + \frac{\left(p_m^{n+\frac{1}{2}} \right)^2}{2\rho_0} \tag{3.135}$$

Steps 4 and 5 between two tube sections: first, we compute the turbulence pressure (loss plus noise) from (3.29) and (3.31), after which the equation of motion becomes

$$\begin{aligned} &\left(1 + \frac{R_m^n \Delta t}{\rho_0} \right) \Delta x_m^{n+\frac{1}{2}} \, p_{m+}^{n+1} + \left(1 + \frac{R_{m+1}^n \Delta t}{\rho_0} \right) \Delta x_{m+1}^{n+\frac{1}{2}} \, p_{m+1-}^{n+1} = \\ &= \Delta x_m^{n+\frac{1}{2}} \, p_{m+}^n + \Delta x_{m+1}^{n+\frac{1}{2}} \, p_{m+1-}^n + 2 \Delta t \left(Q_m^{n+\frac{1}{2}} - Q_{m+1}^{n+\frac{1}{2}} + P_{turb,m,m+1}^n \right) \end{aligned} \tag{3.136}$$

Note that the resistance and the pressure discontinuity are approximated by the “old” values. The left-hand side of this equation can be written as

$$r_m^n \, J_{m+}^{n+1} + r_{m+1}^n \, J_{m+1-}^{n+1} \tag{3.137}$$

Thanks to the continuity of *J*, we can solve

$$J_{m+}^{n+1} = J_{m+1-}^{n+1} = \frac{\Delta x_m^{n+\frac{1}{2}} \, p_{m+}^n + \Delta x_{m+1}^{n+\frac{1}{2}} \, p_{m+1-}^n + 2 \Delta t \left(Q_m^{n+\frac{1}{2}} - Q_{m+1}^{n+\frac{1}{2}} + P_{turb,m,m+1}^n \right)}{r_m^n + r_{m+1}^n} \tag{3.138}$$

Thus, as far as the equation of motion is concerned, the pressure discontinuity is equivalent to a resistance in the tube with the smaller cross section. The factor of 2 is correct: the viscous resistance is also counted twice in these equations.

We can now compute the new limit values of the momentum density *p* and the kinetic pressure *K* at the tube boundaries, as well as the new section volumes:

$$\begin{gathered} p_{m\pm}^{n+1} = \frac{J_{m\pm}^{n+1}}{A_m^{n+1}} \\ K_{m\pm}^{n+1} = \frac{\left(p_{m\pm}^{n+1} \right)^2}{2\rho_0} \qquad ; \qquad V_m^{n+1} = A_m^{n+1} \Delta x_m^{n+1} \end{gathered} \tag{3.139}$$

The continuity equation (3.116) becomes

$$e_{m+}^{n+1} + e_{m+1-}^{n+1} = e_{m+}^n + e_{m+1-}^n + 2c^2 \Delta t \left(J_m^{n+\frac{1}{2}} - J_{m+1}^{n+\frac{1}{2}} \right) \tag{3.140}$$

The left-hand side must be written as

$$\left(Q_{m+}^{n+1}-K_{m+}^{n+1}\right)V_m^{n+1}+\left(Q_{m+1-}^{n+1}-K_{m+1-}^{n+1}\right)V_{m+1}^{n+1} \tag{3.141}$$

so that we can solve

$$Q_{m+}^{n+1}=Q_{m+1-}^{n+1}-P_{turb,m,m+1}^{n+1}=$$
$$=\frac{e_{m+}^{n}+e_{m+1-}^{n}+2c^2\Delta t\left(J_m^{n+\frac{1}{2}}-J_{m+1}^{n+\frac{1}{2}}\right)+K_{m+}^{n+1}V_m^{n+1}+\left(K_{m+1-}^{n+1}-P_{turb,m,m+1}^{n+1}\right)V_{m+1}^{n+1}}{V_m^{n+1}+V_{m+1}^{n+1}} \tag{3.142}$$

Steps 4 and 5 at the lungs: at a boundary closed at the left side

$$J_{1-}^{n+1}=0 \qquad ; \qquad Q_{1-}^{n+1}=\frac{e_{1-}^{n}-2c^2\Delta t\,J_1^{n+\frac{1}{2}}}{V_1^{n+1}} \tag{3.143}$$

Steps 4 and 5 at the lips and at the nostrils: at a boundary open to the atmosphere at the right side

$$\left(1+\frac{R_M^n\,\Delta t}{\rho_0}\right)\Delta x_M^{n+\frac{1}{2}}\,p_{M+}^{n+1}=\Delta x_M^{n+\frac{1}{2}}\,p_{M+}^{n}+2\,\Delta t\left(Q_M^{n+\frac{1}{2}}-Q_{M+}^{n+\frac{1}{2}}\right) \tag{3.144}$$

If we approximate the quantity $Q_{M+}^{n+\frac{1}{2}}$ as the average of Q_{M+}^{n} and Q_{M+}^{n+1}, integrating equation (3.47) to second-order precision leads to

$$\begin{aligned}0&=Q_{M+}^{n+1}-Q_{M+}^{n}-cp_{M+}^{n+1}+cp_{M+}^{n}+\frac{c\Delta t}{a}\frac{Q_{M+}^{n}+Q_{M+}^{n+1}}{2}=\\&=g_{rad}\,Q_{M+}^{n+1}-r_{rad}\,Q_{M+}^{n}-cp_{M+}^{n+1}+cp_{M+}^{n}=\\&=2g_{rad}\,Q_{M+}^{n+\frac{1}{2}}-2\,Q_{M+}^{n}-cp_{M+}^{n+1}+cp_{M+}^{n}\end{aligned} \tag{3.145}$$

where

$$g_{rad}\equiv 1+\frac{c\Delta t}{a} \qquad ; \qquad r_{rad}\equiv 1-\frac{c\Delta t}{a} \tag{3.146}$$

so that the new flow and pressure at the lips and at the nostrils are computed from

$$\begin{aligned}p_{M+}^{n+1}&=\frac{\left(\frac{\Delta x_M^{n+\frac{1}{2}}}{\Delta t}+\frac{c}{g_{rad}}\right)p_{M+}^{n}+2Q_M^{n+\frac{1}{2}}-\frac{2Q_{M+}^{n}}{g_{rad}}}{\left(\left(1+\frac{R_M^n\,\Delta t}{\rho_0}\right)\frac{\Delta x_M^{n+\frac{1}{2}}}{\Delta t}+\frac{c}{g_{rad}}\right)}\\J_{M+}^{n+1}&=p_{M+}^{n+1}\,A_M^{n+1}\\Q_{M+}^{n+1}&=\frac{r_{rad}\,Q_{M+}^{n}+c\left(p_{M+}^{n+1}-p_{M+}^{n}\right)}{g_{rad}}\end{aligned} \tag{3.147}$$

Steps 4 and 5 at a three-way boundary. At a three-way boundary, e.g., at the velopharyngeal port, we have a formula for Q that is analogous to (3.131):

$$Q_{1+}^{n+1} = Q_{2-}^{n+1} = Q_{3-}^{n+1} =$$
$$\frac{e_{1+}^{n} + e_{2-}^{n} + e_{3-}^{n} + 2c^2 \Delta t \left(J_1^{n+\frac{1}{2}} - J_2^{n+\frac{1}{2}} - J_3^{n+\frac{1}{2}} \right) + K_{1+}^{n+1} V_1^{n+1} + K_{2-}^{n+1} V_2^{n+1} + K_{3-}^{n+1} V_3^{n+1}}{V_1^{n+1} + V_2^{n+1} + V_3^{n+1}} \tag{3.148}$$

where we numbered the tubes as in figure 3.11. If the three tubes involved all have equal lengths Δx, the formula for the flow J_{2-} through the left boundary of tube 2, becomes

$$\left(\frac{r_2^n}{A_2^{n+1}} + \frac{1}{\dfrac{A_1^{n+1}}{r_1^n} + \dfrac{A_3^{n+1}}{r_3^n}} \right) J_{2-}^{n+1} = \tag{3.149}$$
$$= \left(\frac{1}{A_2^n} + \frac{1}{A_1^n + A_3^n} \right) J_{2-}^{n} - \frac{2\Delta t}{\Delta x_{av}^{n+1/2}} \left(Q_2^{n+\frac{1}{2}} - \frac{A_1^{n+\frac{1}{2}} Q_1^{n+\frac{1}{2}} + A_3^{n+\frac{1}{2}} Q_3^{n+\frac{1}{2}}}{A_1^{n+\frac{1}{2}} + A_3^{n+\frac{1}{2}}} \right)$$

The formula for J_{3-} is exactly analogous, with the subscripts 2 and 3 exchanged. The flow through the right boundary of tube 1 is

$$J_{1+}^{n+1} = J_{2-}^{n+1} + J_{3-}^{n+1} \tag{3.150}$$

Acoustic output. Finally, the acoustic result is

$$sound(t) = \frac{4\pi}{0.4} \left(\sum_{M=\text{nose,lip}} \frac{J_{M+}^{n} - J_{M+}^{n-1}}{\Delta t} + \sum_{m=\text{every tube}} 1000 \rho_0 \Delta x_m^n \Delta z_m^n \Delta \dot{y}_m^{n-\frac{1}{2}} \right) \tag{3.151}$$

This expresses the sound pressure in Pa (N/m^2) at 40 centimetres from the head.

3.13 Conclusion

In this chapter, we developed a new articulatory synthesizer, which should be able to simulate faithfully more speech phenomena than any previous algorithm could. That these expectations are met, will be shown in chapter 5. Subsequently, the synthesizer will be used to corroborate our explanations for sound structures (parts II and III). Before testing our model in chapter 5, however, we will need some methods to analyse the acoustic output of our model. This will be the subject of chapter 4.

4 Perception models

Abstract. This chapter describes some modest models of auditory perception, including possible ways to determine the perceptual confusion between sounds that differ along more than one dimension.

Besides an articulation model (chapters 2 and 3), a model of peripheral auditory perception would also contribute to our understanding of the relations between articulation and perception. Unlike the situation with articulation models, however, existing models of most perceptual processes seem good enough for a simplified account of peripheral perception, and for stating quantitative predictions about perceptual contrast. In this chapter, I will discuss briefly how I will model the peripheral perception of pitch, intensity, and spectrum. Finally, §4.4.2 describes a way to combine the results of these various perceptual features into a single measure of perceptual contrast.

4.1 Pitch

In a good approximation (for speech sounds), the perceived pitch can be modelled as the acoustic ***periodicity*** of the signal. Boersma (1993b) describes a particularly accurate method for determining the periodicity of a sampled signal from the autocorrelation of the original signal segment, which is estimated as the autocorrelation of a windowed signal, divided by the autocorrelation of the window. This method has proved accurate enough to allow a determination of the ***harmonics-to-noise ratio*** (periodicity divided by noisiness) up to values of 60 dB, which is 30 dB higher than any other method described in the literature.

For a scale of perceptual distance, we should probably express pitch not in Hertz but in Mel units (Fant 1968: p. 206), and to allow comparison of the pitch scale with other perceptual features with respect to perceptual confusion, we should calibrate the pitch scale in ***difference-limen units***, i.e., the distance between two pitch values that are one just-noticeable difference (JND) apart, should be 1.

4.2 Perceptual spectrum

Our purpose in deriving an auditory spectrum from an acoustic signal is the explanation of two universal phenomena in the languages of the world (Crothers 1978):

- The height dimension is used far more exhaustively than the place dimension, i.e., languages tend to have many more vowels between [a] and [i] or between [ɑ] and [u] than between [i] and [u].
- Languages tend to use more front than back vowels, i.e., the average language has a little more vowels between [a] and [i] than between [ɑ] and [u].

Let's assume, with almost everyone else, that these phenomena are related to constraints of auditory perception. We would then like to have a language-independent model to derive the perceptual spectrum from an acoustic signal. The output of our model that does this, will be the excitation pattern of the basilar membrane in the inner ear, as a function of time. From this output, we may be able to derive a measure for the perceptual spectral distance between two sounds.

This has often been tried before; the techniques either used complete spectra (Plomp 1970; Bladon & Lindblom 1981; Klatt 1982), principal components of band-filtered spectra (Klein, Plomp & Pols 1970; Pols, Tromp & Plomp 1973), formant analyses (Peterson & Barney 1952; Pols, Van der Kamp & Plomp 1969; Liljencrants & Lindblom 1972; Kewley-Port & Atal 1989; Ten Bosch 1991; Schwartz, Boë, Vallée & Abry 1997) or large-scale integrations (Chistovich 1985; Schwartz & Escudier 1989). For instance, Ten Bosch (1991), defined a spectral distance measure for vowels from the difference between formant values. This measure contained the following terms:

$$\left(F_{1,A}-F_{1,B}\right)^2+\alpha\cdot\left(F_{2,A}-F_{2,B}\right)^2+\ldots \tag{4.1}$$

where the formants were measured in Bark units. Kewley-Port & Atal (1989) suggest that the perceptual vowel space is two-dimensional and resembles the two-dimensional acoustic space of the first and second formants (expressed in Bark), with a Euclidean distance measure (α = 1). With the measure (4.1), Ten Bosch simulated vowel systems within a spectral space that was bounded by articulatory constraints. The constant α was fitted so that the distribution of vowel inventories emerging from the model matched that of the vowel systems that actually occur in the languages of the world, and the result was that α should be set to 0.2. However, if our assumptions about independent perceptual dimensions are valid for the two-formant case, we could interpret (4.1) in difference-limen units. Fortunately, the difference limens for the first and second formants are known. According to Flanagan (1955), they are 26 Hz (for F_1 = 500 Hz) and 60 Hz (for F_2 = 1500 Hz), which amounts to 0.22 and 0.26 Barks, respectively. If the formants in (4.1) were given in JND units, the constant α would have been $0.2\cdot(0.26/0.22)^2 = 0.3$. This is not anywhere near 1, which should have been the value if the first two formants were the independent perceptual dimensions of the vowel space.

Obviously, a more comprehensive spectral measure is called for. Bladon & Lindblom (1981), for instance, make an [i] that has its third and fourth formants very close together and find that the two-formant vowel that matches this [i] best, must have a second formant F_2' that lies between the third and fourth formants of their [i]. They conclude that this can be described by a model in which the distance between two vowels is determined by the shape of the auditory spectrum as a whole. With them, an auditory spectrum is: the loudness (in Sone units) as a function of frequency (in Bark units), taking into account the filtering by the basilar membrane. In formulas:

$$f=650\sinh\frac{z}{7} \qquad ; \qquad z=7\cdot\left(\frac{f}{650}+\sqrt{1+\left(\frac{f}{650}\right)^2}\right) \tag{4.2}$$

where *f* is the frequency in Hertz and *z* is the frequency in Bark units. The critical bandwidth is 1 Bark everywhere, which is expressed in Hertz as

$$\frac{df}{dz} = \frac{650}{7}\cosh\frac{z}{7} = \frac{650}{7}\sqrt{1+\left(\frac{f}{650}\right)^2} \tag{4.3}$$

The intensity density in Watt/Hertz is

$$\frac{dI(f)}{df} \tag{4.4}$$

where *I*(*f*) is the power in the signal between the frequencies 0 and *f*. The Sound Pressure Level (SPL) “density” is defined in weird units as

$$\mathrm{SPL}(\mathrm{dB}\,/\,\mathrm{Hertz}) \equiv 10\log_{10}\frac{dI(f)}{df} \tag{4.5}$$

The intensity in Watt/Bark must be

$$\frac{dI(z)}{dz} = \frac{dI(f)}{df}\cdot\frac{df(z)}{dz} \tag{4.6}$$

and we can rewrite the Sound Pressure Level “density” in other units as

$$\mathrm{SPL}(\mathrm{dB}\,/\,\mathrm{Bark}) = 10\log_{10}\frac{dI(z)}{dz} = \mathrm{SPL}(\mathrm{dB}\,/\,\mathrm{Hz}) + 10\log_{10}\frac{650}{7} + 10\log_{10}\cosh\frac{z}{7} \tag{4.7}$$

One of the consequences of this formula is that white noise, which has a constant SPL as a function of frequency if measured in dB/Hz, shows a 6 dB/octave high-pass slope for high frequencies if it is measured in dB/Bark. From their figures, it looks as though Bladon & Lindblom did not integrate the intensity density over their one-Bark bands, but averaged it instead, thus effectively incorrectly leaving out the second factor on the right-hand side in (4.6) and the last two terms in (4.7), and finding a constant SPL for white noise, even if measured in dB/Bark.

Bladon & Lindblom went on to use an auditory filter *B*(*z*) to account for masking effects caused by the spreading of excitation due to mechanical properties of the cochlea (formula by Schroeder, Atal & Hall 1979):

$$10\log_{10} B(z) = 15.81 + 7.5(z+0.474) - 17.5\sqrt{1+(z+0.474)^2} \quad (\mathrm{dB}) \tag{4.8}$$

This filter has an area of 1.58 Bark. The Bark power spectrum

$$S^2(z) \equiv \left|\frac{dI(z)}{dz}\right|^2 \tag{4.9}$$

is convolved with this filter, giving the “basilar” spectrum:

$${S_b}^2(z) = \int_0^{26} S^2(\zeta)B(z-\zeta)d\zeta \tag{4.10}$$

and the basilar intensity level:

$$\mathit{Intensity\ level}\ (dB) = 10 \log_{10} S_b^2(z) \tag{4.11}$$

To compute the perceptual intensity (loudness) level (sensation level, SL) in phon units, this has to be corrected for the dependence of the sensitivity of the ear on frequency and intensity according to the equal-loudness curves published by Fletcher & Munson (1933).

Bladon & Lindblom convert the values of the loudness level (in phon) into “loudness” values (in sones) according to

$$L(z) = 2^{(\mathrm{SL}(z) - 40\,\mathrm{phon})/10} \tag{4.12}$$

and express the perceptual distance between two vowels *i* and *j* as

$$D_{ij} = \left(\int_0^{26} \left| L_i(z) - L_j(z) \right|^p dz \right)^{\frac{1}{p}} \tag{4.13}$$

However, this last conversion is uncalled for, as loudness is defined as a property of the sound as a whole, not as a function of frequency, and, more importantly, the loudness in sones bears no simple relationship to difference limens, whereas the loudness level in phon does, as clarified in the next section. The ultimate reason why Bladon & Lindblom decided to use conversion (4.12), must be the prominence that listeners lend to spectral peaks in speech. This, however, is related to the fact that speech should function in noisy environments, and could probably be modelled better by the use of simulated natural noise (which will fill up the valleys), not by an arbitrary transformation that happens to favour peaks.

The left side of figure 4.1 shows the basilar excitation patterns for the vowels at the three corners of the vowel triangle (male Dutch speaker). We see the following features:

- [a] has two prominent peaks around 7.66 Bark (854 Hz) and 10.15 Bark (1299 Hz), and minor peaks at 14.63 Bark (2569 Hz) and 18.79 Bark (4701 Hz), which would drown in a 50-phon background noise, but not in a 40-phon noise. These are very typical, reproducible properties of an [a] spoken in isolation.
- [i] has prominent peaks at 2.71 Bark (252 Hz) and 16.07 Bark (3172 Hz), and a minor peak at 13.14 Bark (2058 Hz), which again would drown in a 50-phon noise. In a typical [i] spectrum (of which this one is an example), the perceptually relevant second formant is the peak above 3000 Hz, not the weaker peak between 2000 and 2500 Hz, which usually makes its appearance in the role of “F2” in two-dimensional pictures of acoustical vowel spaces. The *perceptual* second formant is usually called F_2', and can usually be found between the acoustical F_3 and F_4, which are so close together that they are represented as a single peak in the basilar spectrum.
- [u] has a peak at 2.66 Bark (248 Hz) and one at 5.93 Bark (612 Hz). Higher peaks are so weak that they do not even emerge from a 40-phon background noise. This result, which is quite reproducible again, suggests that higher formants are irrelevant for the perception of [u].

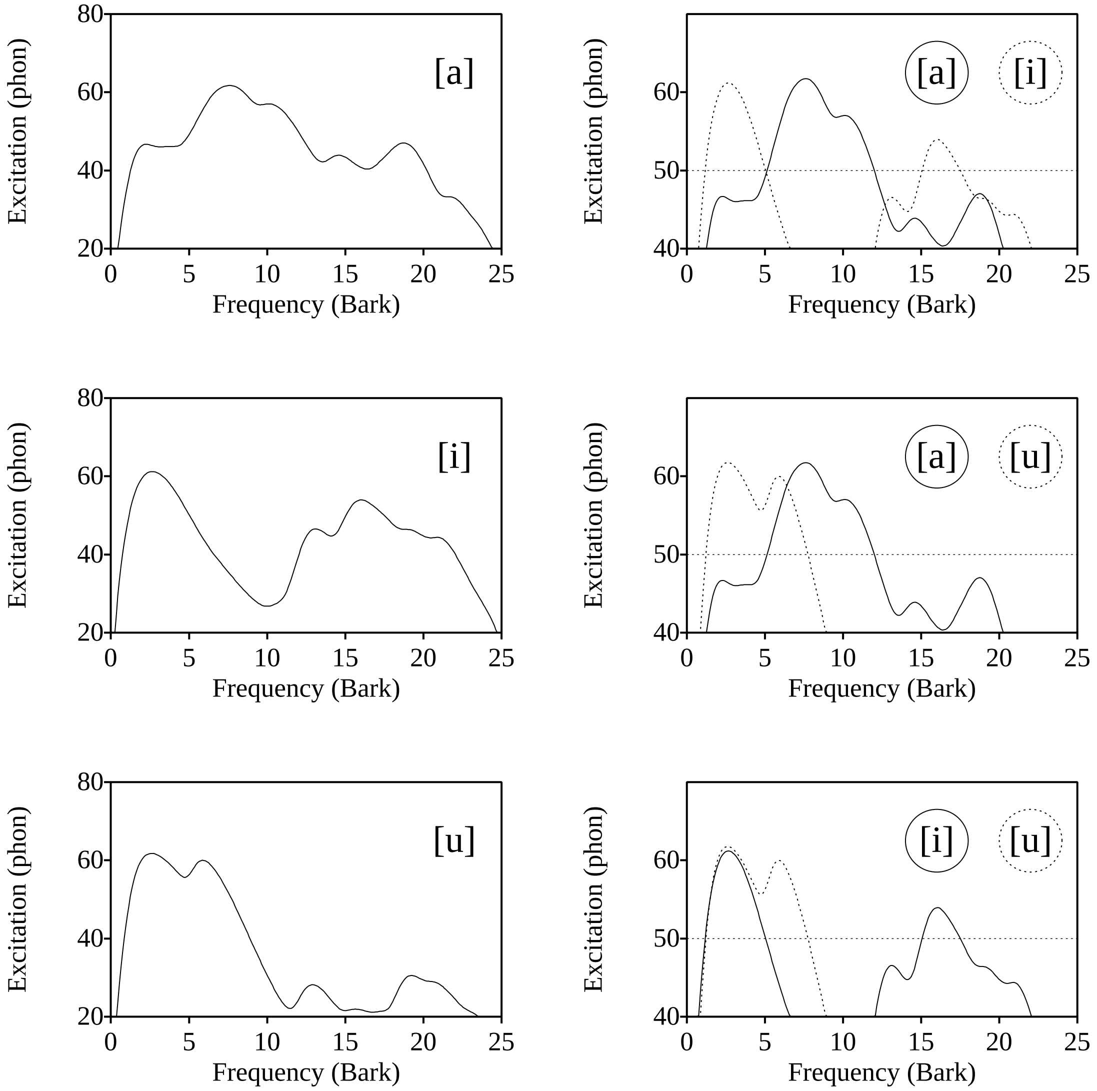


Fig. 4.1 On the left: the basilar excitation patterns of the three vowels at the corners of the vowel triangle. On the right: the amount of overlap between the basilar spectra of the three peripheral vowels, with a uniformly exciting background noise of 40 or 50 phon.

The right side of figure 4.1 pictures the overlap between the three spectra:

- In the upper figure, we see that the basilar spectra of [a] and [i] show no overlap at all in the environment of a 50-phon background noise. Their perceptual distinctivity must be optimal.
- The middle figure shows a partial overlap between the first formant of [a] and the second formant of [u]. Distance measures based on the distances between formant pairs do not take into account such overlap, though it is clear that the region of overlap cannot contribute to the identification of [a] versus [u]. Indeed, it is not difficult to pronounce both [ɑ] and [u] with their two formants so close together that they form a single peak on the basilar membrane (in fact, such vowels would be the most peripheral back vowels possible); in that case, the concept of ordered formants becomes meaningless. Thus, the perceptual space between [a] and [u] is somewhat smaller than the space between [a] and [i]; this would explain the fact that languages possess, on

the average, more vowels between [a] and [i] than between [a] and [u] (there is only a slight skewing, judging from Maddieson 1984).

- The lower figure shows that the first peaks of [i] and [u] completely fall together. Therefore, the [i] - [u] pair is perceptually much closer than either the [a] - [i] or the [a] - [u] pair. Generally, for vowel pairs on the same height but with a different place of articulation, the F_1 values are equal; such a spectral relation is not found for pairs that have the same place but vary along the height dimension. This explain the fact that languages tend to have more height distinctions than place distinctions for vowels.

Here is a comparison with several previous attempts to explain the language data:

- Bladon & Lindblom (1981), while using the above-mentioned distance measure based on “loudness”, found too many vowel places and too few vowel heights when trying to simulate vowel inventories. Lindblom (1990) tried to remedy this by taking into account the proprioceptive contrast between various degrees of *jaw closure*, which, he argues, is greater than the proprioceptive contrast between various tongue places; in his words, vowels should not only sound different, but also *feel different*.
- Ten Bosch (1991) used a distance measure for vowels that involved the difference between the F_1 values of the pair, and the difference between their F_2 values. As this method, too, overestimated the perceptual front/back distinction with respect to height ditinctions, Ten Bosch used a magic factor of 0.3 by which the F_2 distance was devaluated with respect to the F1 distance. This approach is continued by Schwartz, Boë, Vallée & Abry (1997).
- In this section, I derived a correct representation of the ratio of the distance between front and back vowels and the distance between high and low vowels, without using parameter fitting or resorting to otherwise unsupported theories about perceptual contrast in the *speaker*.

I will not continue to try to derive possible vowel inventories from the distance measure developed in this section, because I think, in contrast with Ten Bosch and Bladon & Lindblom, that constraints of minimal number of articulatory and perceptual ‘tricks’ (coordination and categorization, ch. 7, 8, 9) cause symmetries to arise along the dimensions of vowel height and vowel place.

We will now see how to transform the loudness levels in phon into loudness levels along a scale based on difference limens. Jesteadt, Wier & Green (1977) give the following formula for the difference limen of intensity:

$$\frac{\Delta I}{I} = \alpha \left(\frac{I}{I_0} \right)^{-\beta} \tag{4.14}$$

where: I = sensation intensity; I_0 = sensation-level reference (threshold); ΔI = just noticeable intensity difference; $\alpha = 0.463$; $\beta = 0.072$.

The number of difference limens above threshold is thus

$$\mathrm{DLI-level(I)} = \int_{I_0}^{I} \frac{dx}{\Delta I(x)} = \frac{1}{\alpha} \int_{I_0}^{I} dx\, I_o^{-\beta} x^{\beta - 1} = \frac{1}{\alpha\beta} \left(\left(\frac{I}{I_0} \right)^{\beta} - 1 \right) \tag{4.15}$$

The sensation level in phon is defined as

$$\mathrm{SL} = 10\log_{10}\frac{I}{I_0} \tag{4.16}$$

From this, it follows that the number of difference limens above threshold is

$$\mathrm{DLI} - \mathrm{level}(\mathrm{SL}) = \frac{1}{\alpha\beta}\left(10^{\frac{\beta}{10}\left(10\log\frac{I}{I_0}\right)} - 1\right) = 30\cdot\left(1.0167^{\mathrm{SL}} - 1\right) \tag{4.17}$$

We can now use an equation like (4.13) with p=2 (“root-mean-square”) to compute the distance between two perceptual spectra. If this is tried on a digital simulation of three vowels [a], [i], and [u] of Fant’s (1960) Russian speaker, with as a voice source light pink noise (white below 2 Bark, –6 dB/octave above, so as not to get any interactions with harmonics of F_0), the distances between the vowels are:

- distance between [a] and [i]: 18 JNDs
- distance between [a] and [u]: 18 JNDs
- distance between [i] and [u]: 12 JNDs

This means that it would be equally suitable for a language to have three vowel places, as it would be for it to have four vowel heights. This explains the fact that the languages of the world use more height oppositions than place oppositions, *on the average*.

4.3 Intensity

The best smoothing method for a periodic signal is convolution with a Gaussian window (Boersma & Weenink 1996), and the method for the measurement of an acoustic intensity contour from a windowed signal is related to this: band-limit the signal between 0 Hz and one half of the Nyquist frequency (probably by upsampling by a factor of two), square all the samples, and convolve with a Gaussian window (probably via multiplication by a Gaussian in the frequency domain). This method is used in §5.7 for measuring the relationship between lung pressure and intensity.

However, the perceptual correlate of intensity is not the acoustic intensity, but the excitation pattern, integrated along the basilar membrane (eqs. 4.11-4.13). Thus, the perceived loudness is expected to be higher for sounds with flat basilar spectra than for sounds that have most of their energy in the lower frequency range. Indeed, Sluijter (1995) found that syllables with flat spectra were perceived as more “accented” than syllables with falling spectra; she did not determine how much of this effect is related to basilar integration and how much has to be ascribed directly to the perception of an independent perceptual feature of spectral balance; probably mainly the latter, because the Strouhal number (Stevens 1971) is a good indication of the relative produced power.

4.4 Contrast and confusion

I will now review some possible ways of measuring contrast or confusion.

4.4.1 Discrete measures

A rough measure of the contrast between two utterances is the number of differing features. For instance, the difference between [v] and [p] is larger than the distance between [b] and [p]: two features (voicing and frication) versus one feature (voicing).

More precision can be achieved if we recognize the fact that the existence of a salient feature may partially obscure another contrast. Thus, the voicing contrast between [b] and [p] will probably be larger than the contrast between [f] and [v], because the presence of frication noise distracts the attention from other features. This statement has its roots in intuitive knowledge about the workings of the human ear. If not, we could equally well have brought forward that "the voicing contrast between [b] and [p] will probably be *smaller* than the contrast between [f] and [v], because the *absence* of frication noise distracts the attention from other features". We know, however, of two properties of the auditory mechanism: firstly, the presence of noise may mask spectral information from other sources; secondly, periodic noise bursts (as in [z]) have a lower degree of periodicity than a truly periodic signal (as in [b]), thus giving a smaller periodicity contrast for the fricatives than for the plosives. A large say in the matter comes from perception experiments (though these are heavily influenced by language-specific categorization), which agree that [b] and [p] are perceptually farther apart than [f] and [v] (for Dutch: Pols 1983). The unmarkedness of plosives as compared to fricatives, as can be induced from the data of the languages of the world, can partly be traced back to this asymmetry.

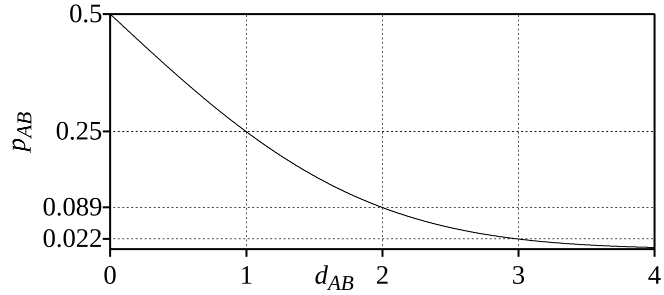
We can achieve a little more precision yet by taking into account some asymmetries of the speech organs. Thus the voicing contrast between [k] and [g] will be smaller than the one between [p] and [b], because of the different volumes of expandable air involved in helping to maintain the contrast (§5.12, §11.9, §16.4.1, §16.4.2, §17.1.2, §17.3.15).

4.4.2 Combining various perceptual dimensions to a global contrast measure

There exists a universal measure for the perceptual contrast between any two events (e.g., sounds) A and B. This measure is the ***confusion probability*** of A and B, and is defined as the probability that event A will be perceived as event B, which need not be equal to the probability that event B will be perceived as A. If this confusion probability *is* symmetric with respect to A and B (i.e., if there is no ***bias*** for either A or B), and A and B differ along only one acoustic/perceptual dimension, the confusion probability often bears a monotonic relationship with the ***distance*** between A and B along that dimension. This distance can then be expressed as a number of ***difference limens*** (units of just noticeable differences), and, if the variation along the scale is small in comparison with the total length of the scale, this number of difference limens may well exhibit an almost universal relationship with the confusion probability. Thus, if the distance between A and B is one difference limen, the confusion probability is 25% (this is one definition of a difference limen); if the perceptual measurements are drawn from a Gaussian distribution, and the distance is two difference limens, the confusion probability is 10%; for three difference limens, it is 2.4%; for four, 0.47%. The confusion probability is given by the formula

$$p_{AB} = \tfrac{1}{2}\left(1 - \operatorname{erf}\left(d_{AB} \cdot \operatorname{inv}\operatorname{erf}\left(\tfrac{1}{2}\right)\right)\right) \qquad (4.18)$$

or by the curve



(4.19)

where d_{AB} is the difference between A and B, expressed in difference limens, and erf is related to the primitive of the Gaussian distribution function. If there are three events A, B, and C, there are two special cases. The first special case is if all three events differ along the same dimension, and B is perceptually somewhere between A and C. The distance between A and C can then be expressed as

$$d_{AC} = d_{AB} + d_{BC} \qquad (4.20)$$

The second special case is if B and C differ along a dimension that is perceptually independent of the dimension along which A and B differ. The confusion probability between B and C can then be expressed as

$$p_{AC} = p_{AB} \cdot p_{BC} \qquad (4.21)$$

To derive an equation for the distance between A and C, we approximate (4.18) by

$$p_{AB} \approx e^{-\left(\frac{d_{AB}}{\alpha}\right)^2} \quad \text{or} \quad d_{AB}^2 \approx -\alpha^2 \log p_{AB} \qquad (4.22,23)$$

We can now rewrite (4.20) as

$$d_{AC}^2 \approx -\alpha^2 \log p_{AC} = -\alpha^2 \log\left(p_{AB} \cdot p_{BC}\right) = -\alpha^2 \log p_{AB} - \alpha^2 \log p_{BC} \approx d_{AB}^2 + d_{BC}^2 \quad (4.24)$$

which is the perceptual counterpart of the global articulatory-effort measure that we will see later in equation (7.4).

If we realize that both equations (4.20) and (4.24) are Euclidean distance measures (for one dimension and two independent dimensions, respectively), we can conclude that the distance in the perceptual space can be measured as if this were a Euclidean space, provided that it is calibrated in units of one difference limen along every independent dimension. For instance, if the intensities of two sounds differ by 3 difference limens, and their pitches differ by 4 difference limens, the perceptual distance between these sounds can be expressed as “5 difference limens”.

To sum up, measuring every perceptual dimension with a dimensionless difference-limen scale allows us to compare distances along very different kinds of dimensions, and to compute in a natural way the total distance between any pair of events, provided that the Gaussian hypothesis and the strong hypothesis of separability (4.21) holds. And, of course, they do not normally hold. For instance, the total confusion probability may depend only on the maximum constituent confusion probability (a case of strict ranking):

$$p_{AC} = \max\left(p_{AB}, p_{BC}\right) \tag{4.25}$$

or, in the other direction, (4.20) might hold even if the pairs AB and BC differ along perceptually independent dimensions (city-block distance), so that the two sounds of our example differ by 7, instead of 5, difference limens.

In chapters 7 and 9, we will see that global measures of perceptual confusion are linguistically irrelevant, because each language is free to choose the relative importance that it assigns to the various perceptual dimensions.

4.4.3 Perceptual salience versus dissimilarity

Kawasaki (1982) draws our attention to the acoustic correlates of two aspects of the maximization of contrast. First, she points out that languages tend to disfavour contrasting, but acoustically very similar, sounds: poorly distinguishable sequences such as [gla] and [dla] tend not to co-occur in languages; Kawasaki calls this ***maximization of dissimilarity***. Secondly, sequences of acoustically similar sounds such as [wu] or [ji] are avoided in the world's languages in favour of sequences with a greater acoustical dynamic variation like [wi] or [ju]. Kawasaki calls this ***maximization of perceptual salience***.

Kawasaki defines perceptual salience as the amount of change of the perceptual features within an utterance. Her formula is

$$\sum_i \int \left(\frac{dP_i(t)}{dt} \right)^2 dt \tag{4.26}$$

where P_i are perceptual features (in Kawasaki’s case, formants in mel). This is consistent with (4.24); the use of the squares causes (4.26) to be sensitive to the rate of change of the parameter, interpreting rapid changes as more salient than slow ones.

An analogous formula for the perceptual contrast between the utterances *a* and *b* is

$$\sum_i \int \left(P_{a,i}(t) - P_{b,i}(t) \right)^2 dt \tag{4.27}$$

In §4.4.2, we saw how perceptual features of different origin (voicing, tone, spectrum, loudness) can be combined in such a formula if we know all of their difference limens.

4.5 Conclusion

The models discussed in this chapter will be used in chapter 5 to evaluate the articulation model, and in parts II and III to underpin any claims with respect to the perceptual features of speech sounds and their relation to the speech production mechanism.

5 Test of the articulation model[1]

Abstract. This chapter shows that our articulation model can faithfully simulate various phenomena that occur in speech.

In this chapter, we will investigate how our articulation model performs in simulating some aerodynamic and myoelastic phenomena that occur in the production of speech. In each of the simulations, the speaker modelled is a female speaker with vocal folds consisting of two masses each. First, we will look at what happens when our speaker tries to exhale by reducing the equilibrium width of her lungs, with her glottis open and her lips open or closed. After that, we will look into what happens if the vocal folds are brought closely together while the supralaryngeal passage is not obstructed. From §5.9 on, we will see what happens if there *is* an obstruction.

In contrast with the descriptions in chapter 2, the simulations of the current chapter were all performed, as far as the oral and pharyngeal walls were concerned, with a constant Δz of 15 mm, a wall mass of 6 grams, a linear wall tension of 30 N/m, a cubic wall tension of zero, and a relative damping of 1.

5.1 Silence

If the activities of all the muscles in the equations of chapter 2 is 0, the walls of the tubes stay still, the air flow stays 0, and the “adiabatic” air pressure stays constant at $\rho_0 c^2 = 142054.26\ \mathrm{Pa}$, which is 1.4 times higher than the true atmospheric pressure of 101467 Pa (100 Pascal = 1 millibar ≈ 1 cm H_2O).

5.2 Sigh

If our speaker wants to exhale, she could act as shown in figure 5.1:[2]

1. Reduce the equilibrium width of the lungs from 132 to 120 millimetres in the first 30 milliseconds, by reducing the *lungs* parameter (eq. 2.26) linearly from 0.1 at $t = 0$, to 0 at $t = 0.03$ seconds. This is only barely visible in figure 5.1, but can also be seen as a dotted line in figure 5.3a (left).
2. Keep the pharyngeal and oral cavities open, by keeping the activities of all supralaryngeal muscles (eqs. 2.40-2.67) at 0 during the entire utterance.
3. Keep the glottis open (at 5 mm) by keeping both the *interarytenoid* parameter and the *posteriorCricoarytenoid* parameter relaxed at 0 (eq. 2.38).

[1] Parts of this chapter appeared in Boersma (1993a) and Boersma (1995).

[2] The Praat scripts that create the articulations in this chapter, or compute the sounds, tube widths, air pressures, and air velocities, or draw the pictures, are available from **http://fonsg3.hum.uva.nl/paul/**, so that the reader can replicate everything in this chapter and find by herself the answers to the questions not touched upon here. The generated sounds are available as web-playable **.au** files.

What happens in the vocal tract, depends on whether the velopharyngeal port is closed (by keeping the *levatorPalatini* parameter at 1, see eq. 2.40) or open. Figure 5.2 shows spectrograms of nasalized and oral sighs: the nasalized version shows less high-frequency noise, because the air velocity in the oral cavity is so low that no turbulence noise is generated there, so that the spectrum will be that of a glottal noise source filtered by the supralaryngeal cavities.

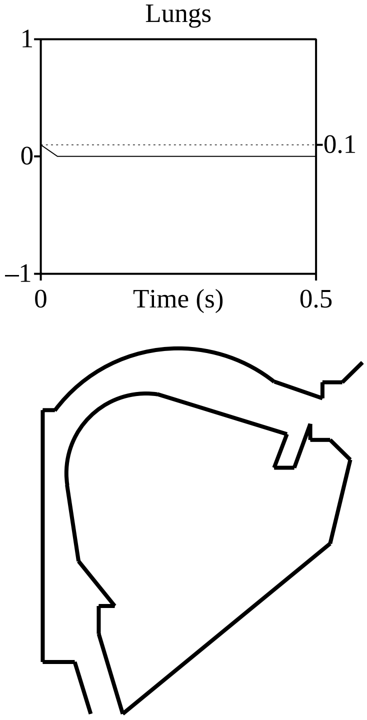


Fig. 5.1 Expiration.

Figure 5.3 shows what happens inside the tract if the velopharyngeal port is closed. During the first 30 ms, the lung pressure (fig. 5.3a) quickly rises to 440 Pa (relative to the atmospheric pressure, as all pressures in the rest of this chapter), so that the realized width of the lungs approaches the rest width of 120 mm much slower than does the equilibrium width, which is shown by a dotted line. The lung pressure diminishes with time, since air starts to flow from the lungs through the bronchi into the vocal tract (fig. 5.3b). The pressure in the bronchi does not exceed 330 Pa. As figure 5.3c shows, the air flows through the glottis fast enough (more than v_{crit} = 10 m/s, see §3.4) to generate turbulence noise, between t = 15 ms and t = 270 ms. This gives the auditory impression of a sigh. The pressure in the glottis first rises to +80 Pa because of the extra amount of air arriving from the lungs, and then falls to –460 Pa due to the Bernoulli effect (§3.2.2) caused by the large air velocity: for v = 37 m/s, the kinetic pressure $-\frac{1}{2}\rho v^2$ is –780 Pa. Much the same happens near the uvula (fig. 5.3d): since the velocity rises above 10 m/s, noise is generated in the oral cavity as well, with an audible impact on the sound. The macroscopic volume velocity of the air is approximately equal to the volume velocity in the glottis. Finally, the evolution of the widths of the bronchi, the glottis, and the vocal tract near the uvula (fig. 5.3e) follows the air pressures, except that at the uvula, the movement of the walls is smoothed as a result of their inertia.

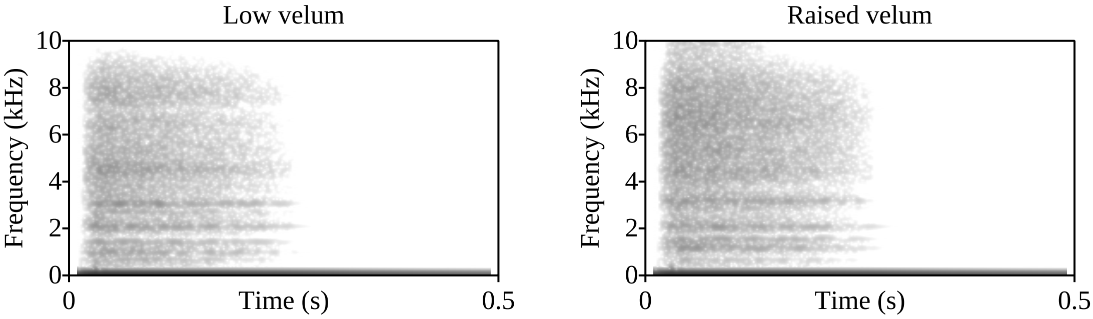


Fig. 5.2 Spectrograms of nasalized and oral sighs (100 dB dynamic range; Gaussian window with an effective length of 10 milliseconds; no pre-emphasis) .

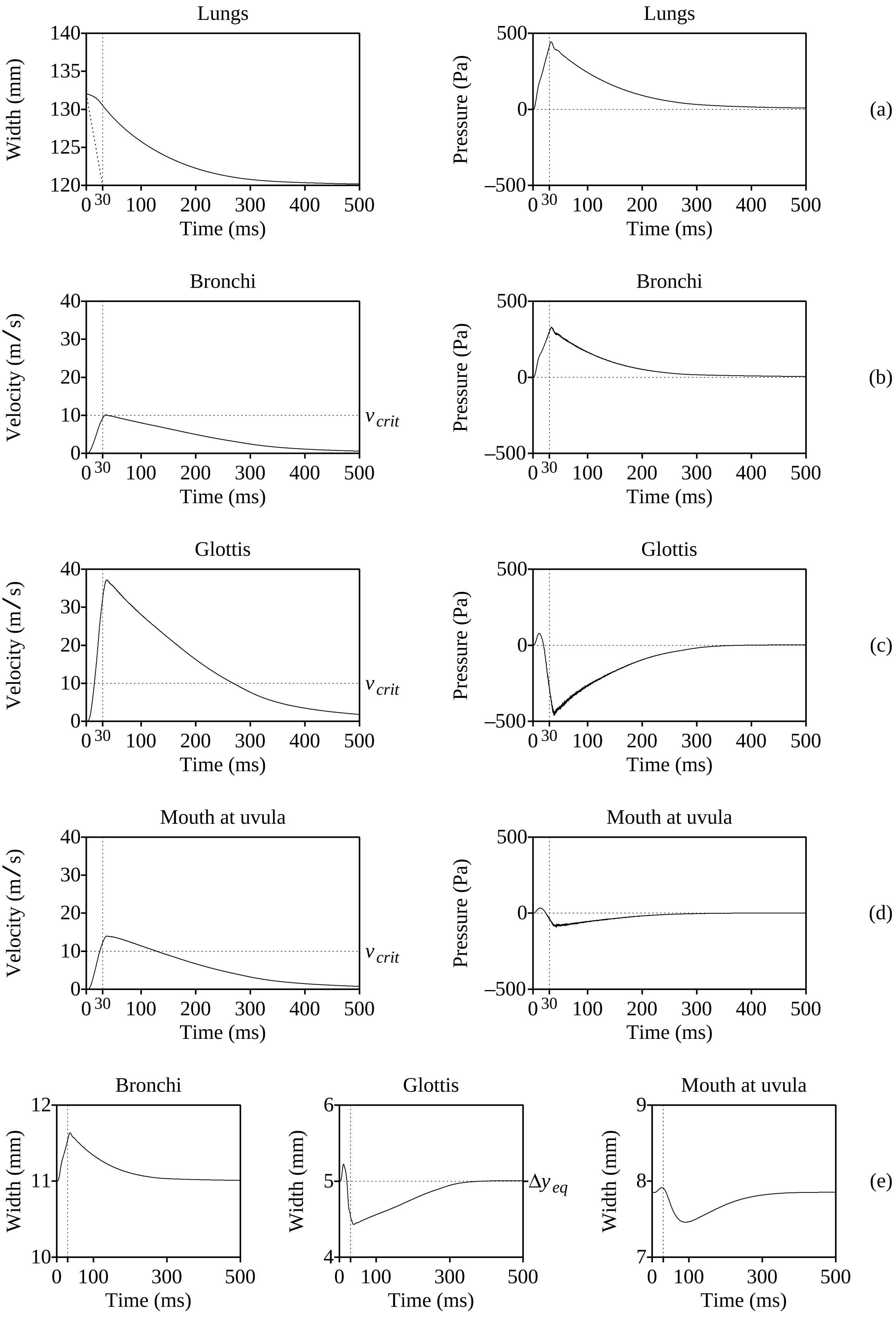


Fig. 5.3 Expiration: tube widths, particle velocities and air pressures at several positions in the vocal tract, if the equilibrium width of the lungs is given by the falling dotted line in the upper left figure. In each figure, a vertical dotted line is shown at t = 30 ms.

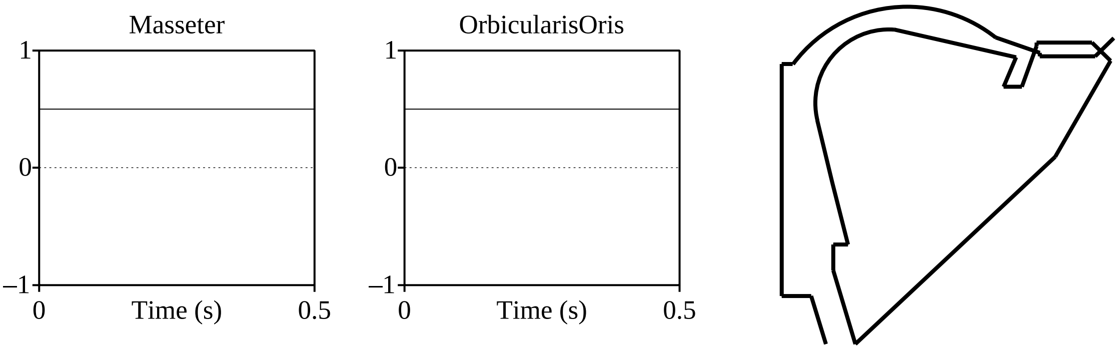


Fig. 5.4 Inflation of the vocal tract.

To model inspiration instead of expiration, we could set the initial *lungs* parameter to –0.1 instead of +0.1, and keep the *posteriorCricoarytenoid* parameter constant at 1 to simulate the glottal abduction that occurs during inspiration. According to Hirose (1997), the abduction during deep inspiration is achieved by activity of the posterior cricoarytenoid muscle (though Fink & Demarest (1978) attribute it to an “unfolding” caused by a tracheal pull).

5.3 Balloon

To inflate her oral cavity, our speaker could proceed as in figure 5.4:

1. Reduce the equilibrium width of the lungs as for expiration (§5.2).
2. Keep the glottis open as for breathing (§5.2).
3. Keep the velopharyngeal port closed (§5.2).
4. Keep the oral cavity closed, by keeping the *masseter* and *orbicularisOris* parameters (eqs. 2.49 and 2.66) constant at 0.5.

During the first 30 ms, the pressure in the lungs (fig. 5.5a) quickly rises to 570 Pascal (above the atmospheric pressure). In contrast with the case of breathing, the pressure does not diminish with time. Instead, the pressures throughout the vocal tract become equal after 100 ms or so, as shown by the right sides of figures 5.5a-d. After 0.5 seconds, the pressure in the lungs is 553.757 Pa, in the glottis it is 553.756 Pa, and in the oral cavity it is 553.551 Pa. During these 0.5 seconds, the total volume of the vocal tract dropped from 3.803443 litres to 3.788767 litres. This predicts an equilibrium pressure of (3.803443 / 3.788767 – 1) · 142054.26 Pa = 550.26 Pa. This 1% difference of 3 Pa is probably due to the approximate nature of the first part of (3.134), which is not explicitly mass-conserving; however, this should not worry us, since the error of neglecting heat conduction and convection is much larger (40%, see §3.3).

Figure 5.5a shows that in contrast with the case of breathing, the lungs never reach their target position of 100 mm; instead, they do not contract much. Figure 5.5e shows that the widths of the tubes in all other parts of the vocal tract become greater than the equilibrium positions of their walls.

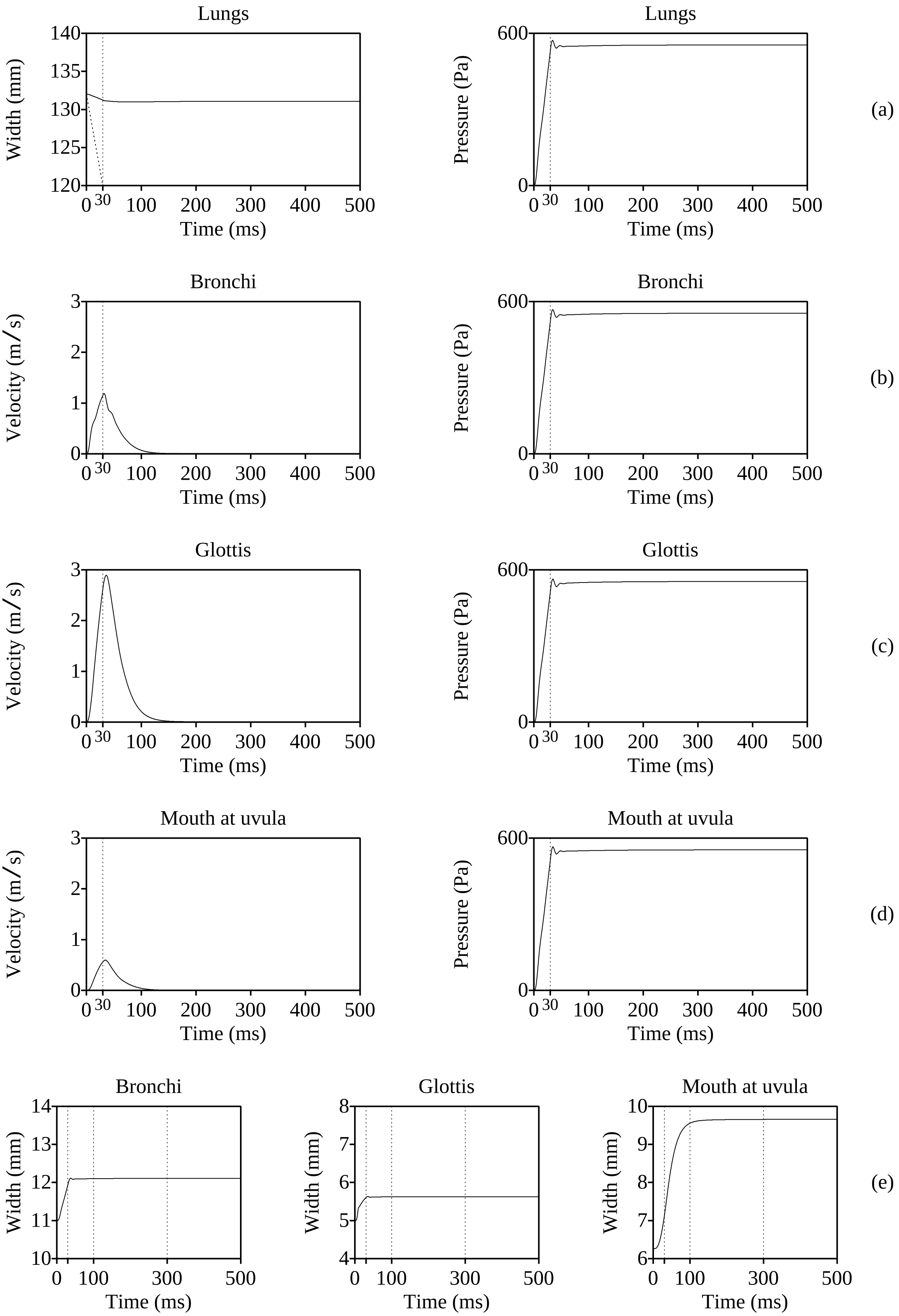


Fig. 5.5 Inflation: tube widths, particle velocities and air pressures at several positions in the vocal tract, if the lungs contract, the glottis is held open, but the lips and velopharyngeal port are kept closed.

5.4 The onset of phonation

If our speaker wants to phonate, she could act as follows:

1. Reduce the equilibrium width of the lungs as for breathing (§5.2).
2. Keep the pharyngeal and oral cavities open, as for breathing.
3. Keep the glottis closed, with an equilibrium width of 0 mm, achieved by setting the *interarytenoid* parameter (eq. 2.38) to 0.5.

Further, we take a two-mass model for each vocal fold, close the velopharyngeal port (as in §5.2), and set the *cricothyroid* parameter (eqs. 2.35, 2.37) to 1 for a slightly raised F_0.

In figure 5.6, we see that the vocal folds start vibrating. After the fourth cycle, at 25 ms, the upper parts of the vocal folds close for the first time (fig. 5.6a). One cycle later, at 28 ms, the lower parts also close (fig. 5.6b). After this, the vibration is quite regular.

The widths shown in figures 5.6a and 5.6b are the same as the quantity Δy introduced in chapter 2 (fig. 2.2), and can therefore be negative, although, of course, the actual area of the orifice is always positive (eq. 2.8).

Figure 5.6c shows the build-up of tracheal pressure. It fluctuates with vocal-fold vibration around 500 Pa. The pressure in the lower part of the glottis (fig. 5.6d) fluctuates pitch-synchronously between –400 and +400 Pa, with sharp peaks up to +2.7 kPa (§5.5.3). The acoustic result (fig. 5.6e) shows a characteristic initial pressure drop that is not usually found in speech recorded with a microphone.

5.5 During phonation

With figure 5.7, we can look into the myoelastic and aerodynamic phenomena that occur in the glottis during phonation, and into their relative timing.

5.5.1 The motion of the vocal folds during phonation

As figure 5.7a shows, the amplitude of the vibration in the lower part of the glottis is greater than the amplitude in the upper part. This is remarkable in a sense, since Ishizaka & Flanagan (1972) had to model the lower parts with a lower damping ratio (0.1) than the upper parts (0.6), while we model all of them with a damping ratio of 0.2 (§2.7.2).

We see that the closing gesture is much faster than the opening gesture (the fall of the curves is much steeper than the rise), and that the upper parts of the vocal folds close and open somewhat later than the lower parts; we also see that the glottis is closed from $t = 163.2$ ms to 165.3 ms, and open from 165.3 ms to 167.7 ms, from which we can compute that the glottis is open 54 percent of the time. Both phenomena agree well with measurements of areas of live vibrating vocal folds (Koike & Imaizumi 1988).

If we represent each vocal fold by a *single* mass, the vocal folds will still vibrate (this was modelled by Flanagan & Landgraf 1968). In this case, reported in Boersma (1991), the width of the glottis as a function of time roughly resembles the dotted curve in figure 5.7a, which means that the glottis is open about 85 percent of the time. So here the one-mass model deviates from reality (or is a faithful model of the falsetto register).

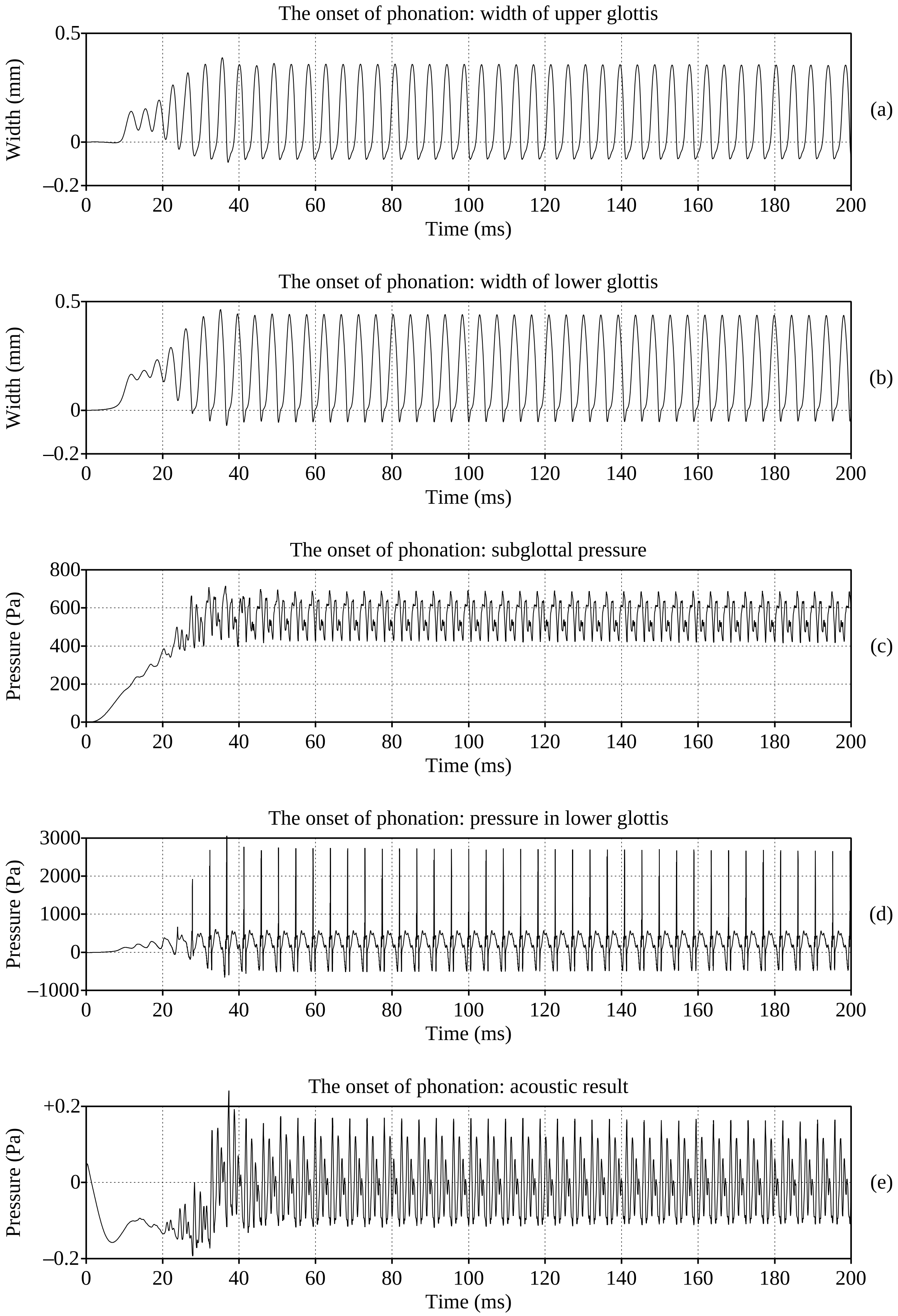


Fig. 5.6 The motions of the vocal folds and some relevant pressures, as functions of time.

5.5.2 Air velocity in the glottis during phonation

In figure 5.7b, we see the velocity of the air particles in the lower and upper parts of the glottis. After the upper glottis opens (at 165.3 milliseconds), the velocity starts rising, attaining a maximum of 39 metres per second.

During the first part of the open interval, the velocity tends to oscillate. This is due to the phase difference between the subglottal formant and the supraglottal formant (figure 5.7d). The supraglottal formant starts to die out after the glottis has been open for one millisecond or so, because of the damping effect of the lungs, and so does the velocity oscillation. This pitch-synchronous variation in the damping and, therefore, in the externally measurable bandwidth of the supraglottal formants, is a well-established phenomenon found in live human speakers (Flanagan 1972: 65).

As the lower parts of the vocal folds close (at 167.7 ms), the velocity quickly drops to zero. Though the air has to escape very quickly from between the walls, it can flow away in two directions, making the average velocity zero. Because the amount of air displaced is very small, no traces of it are seen in the velocity in the upper part of the glottis. However, when the upper parts of the folds close (at 167.9 ms), we see a positive velocity peak in the upper part of the glottis, because the little amount left between the walls can only escape well into the supraglottal direction. Also, the even tinier amount of air pumped from the upper glottis into the “closed” lower glottis (there is a leaking width of 0.01 mm) shows up as a negative peak of –1 m/s in the lower glottis at 167.9 ms.

The reverse effects are seen when the lower parts of the vocal folds start opening (at 168.4 ms): the air sucked between the folds must come for the larger part from the subglottal end of the glottis, which gives rise to a positive velocity peak in the lower part of the glottis. At the same time, a very small amount of air is sucked from between the almost-closed upper parts of the vocal folds, giving rise to a negative velocity peak there.

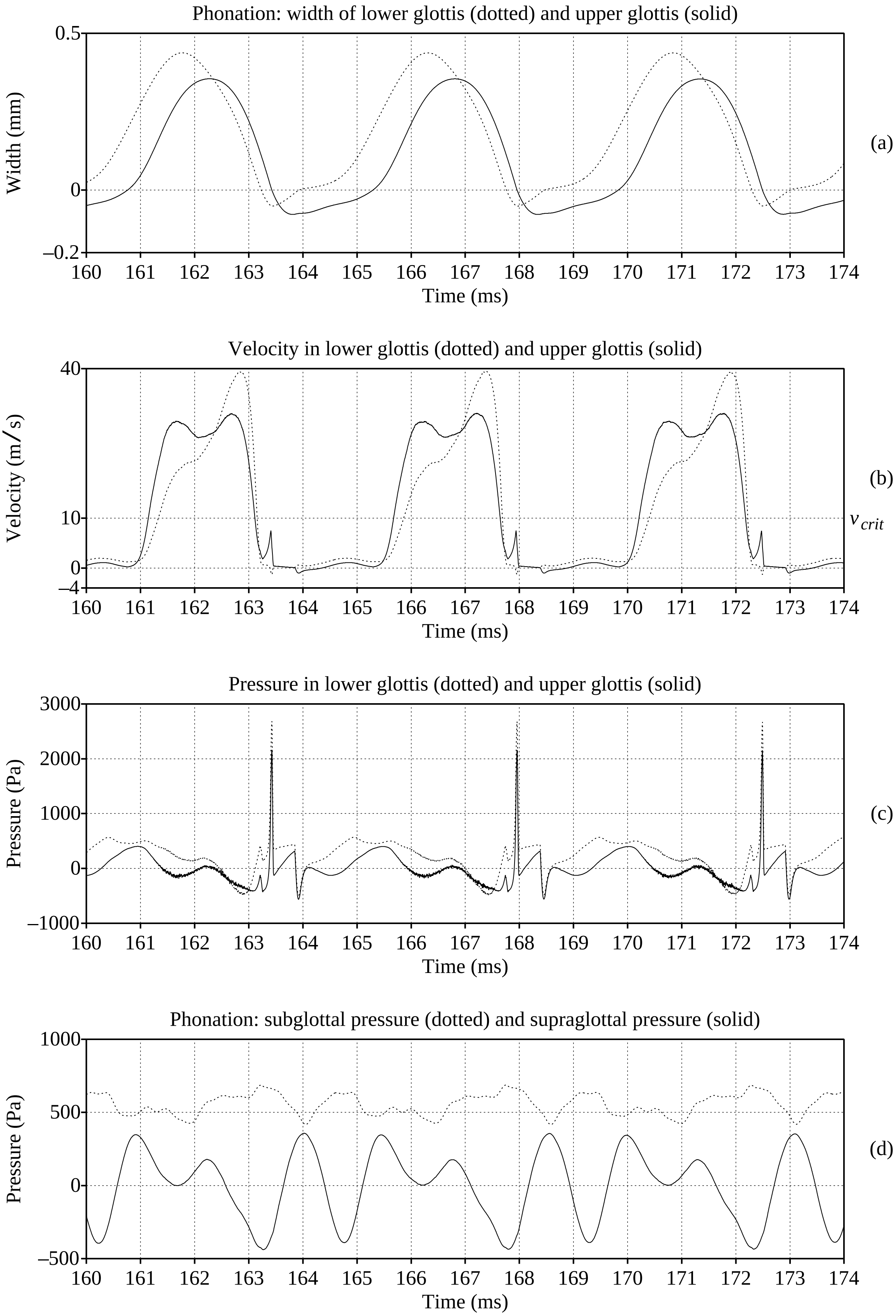
We have seen that we can get a detailed understanding of the velocity contour with physical arguments.

5.5.3 Air pressure in and around the glottis during phonation

In figure 5.7c, we see that the pressure in the lower half of the glottis is negative during the closing gesture of the vocal folds (e.g. around 163.0 ms). This negative pressure, which is due to the air velocity being high (Bernoulli effect), causes the slope of the closing gesture to be much steeper than the slope of the opening gesture (fig. 5.7a).

After this, the lower halves of the vocal folds close; this gives rise to a positive pressure peak of almost 400 Pa (at 163.2 ms). This must be due to the air being compressed between the rapidly approaching walls.

An even larger positive peak of 2700 Pa appears to arise in the lower glottis when the *upper* halves of the vocal folds close. This pressure peak has no acoustic result below the glottis, as we can see in figure 5.7d, which shows that both the subglottal formant and the supraglottal formant (700 Hz) are triggered by the first, smaller, peak. The fact that a large pressure may have no acoustic effect, is caused by the smallness of the amount of air that can be displaced if the glottal width is 0.01 mm (leakage); this is comparable to the phenomenon that static electricity causes no bodily harm even at 10,000 Volts thanks to the low amount of electric charge contained by a small capacitance.


Phonation: width of lower glottis (dotted) and upper glottis (solid)

Velocity in lower glottis (dotted) and upper glottis (solid)

Pressure in lower glottis (dotted) and upper glottis (solid)

Phonation: subglottal pressure (dotted) and supraglottal pressure (solid)

Fig. 5.7 The synchronicity of vocal-fold motion and aerodynamics in the glottis during phonation. Because of the sharp features in figure (c), the sampling rate was 551250 Hz.

The fourth phenomenon that we can see in figure 5.7c, is a negative pressure peak arising when the lower part of the glottis opens (around 163.9 seconds). This is a sticky reaction mirroring the first positive peak. The fact that the upper glottis is still closed, enhances this effect, because air can be sucked in from one end only; however, the stickiness is also seen in the one-mass model of the vocal folds, though less conspicuous (Boersma 1991).

Finally, figure 5.7c shows us the noise generated by turbulence when the air velocity is greater than 10 m/s, in both parts of the glottis between 166.0 and 167.5 ms.

5.6 Sustained phonation

Phonation cannot go on forever: when the glottis is open, some air moves through it from the lungs to the pharyngeal and oral cavities, and from there to the outer air. Figure 5.8a shows how the subglottal pressure of our model speaker (who, remember, is a woman with two-mass vocal folds) decreases as time goes by. In the first 5 seconds, the average pressure slowly falls from 500 to 240 Pa. Around 5.5 seconds, some irregularities emerge. After this, the pressure drops further, until another phenomenon (not visible in the figure) shows up at 9.5 seconds.

The interesting things at 5.5 and 9.5 seconds are ***register break***s. We can see this by comparing the movements of the vocal folds at different times. During the first 5 seconds, both parts of the glottis periodically close, as seen in figure 5.7a.

At 5.3 seconds (not shown), there are alternating cycles: in the odd-numbered cycles, the lower part of the glottis closes only for a very short time; in the even-numbered cycles, it does not close at all. This ***period doubling*** (seen in fig. 5.8d as an octave drop) is known from chaos theory as being characteristic of the first stage of the transition from a periodic movement to a chaotic movement. It has been reported for the vocal break from the modal to the falsetto register by Švec & Pĕšák (1994).

Around 5.6 seconds, the behaviour is irregular, as we see in figure 5.8b: whether the lower parts of the vocal folds close or not during a certain cycle, is not predictable from what happened in the cycles before. At 5.8 seconds, there are alternating cycles again (not shown). The irregularity can be heard as a rough (aperiodic) sound. At 6.3 seconds, a different periodic motion sets in, shown in figure 5.8c: the lower part of the glottis does not close, and the upper halves of the vocal folds touch for a short time only, so that the glottis is open 80% of the time. This situation probably represents the falsetto register.

From 9.5 seconds on, the upper part of the glottis does not close any longer, so here is another break. If we had modelled each vocal fold with three parts instead of two, we would have found one more register break. Whether these breaks in our model are realistic phenomena, is difficult to say, as long as we cannot ask people to phonate for as long as possible without adjusting their chest muscles, and at the same time numb their respiratory reflexes; nevertheless, people do utter several breaks when trying to phonate their lungs empty.

Figure 5.8d shows the fundamental frequency during the 12 seconds of phonation. The fact that the fundamental frequency suddenly rises during a falsetto break, is a realistic phenomenon, and so is the irregularity that occurs during the transition.

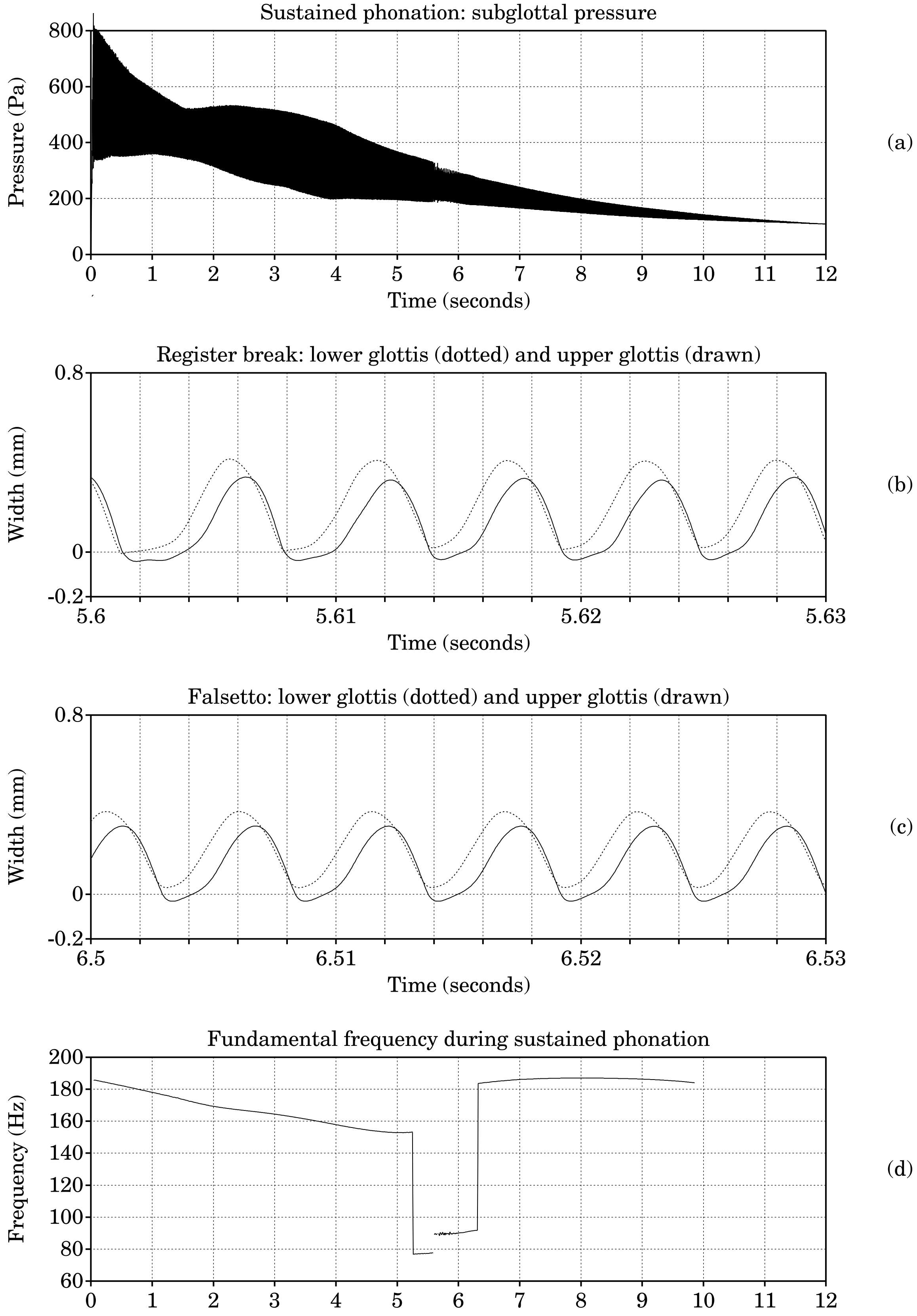


Fig. 5.8 Sustained phonation with register breaks at 6.5 and 9.5 seconds.

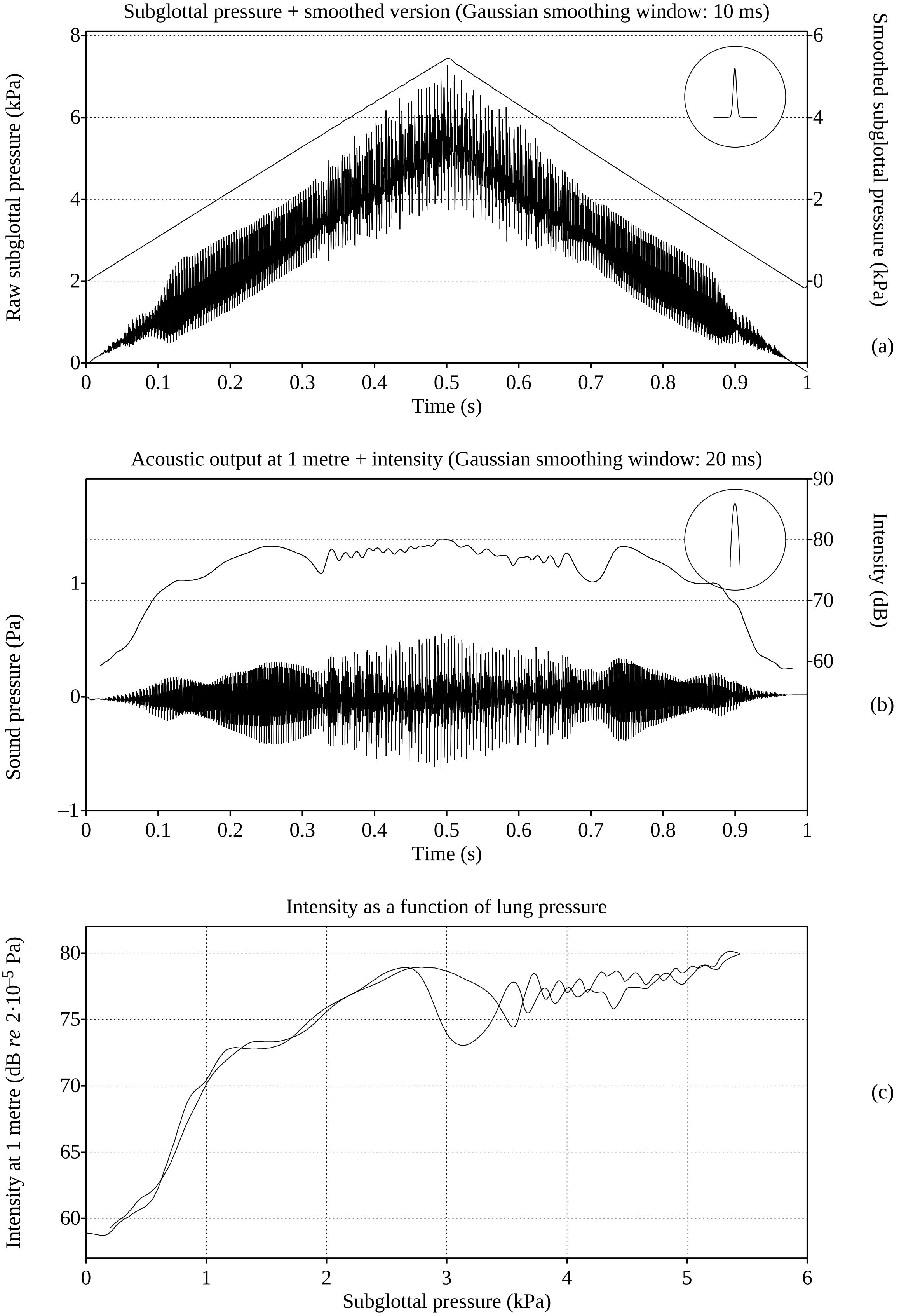
Male voice. While our female voice manages to phonate for 11 seconds without adjusting the equilibrium width of her lungs after the first 30 milliseconds, the corresponding male voice gives up after 6 seconds (and, again, a falsetto break). This is because our male speaker has much larger vocal folds, which open for more than a millimetre; thus, the male glottis forms much less of a constriction to the air sneaking out into the open, than the female glottis; the male's larger lung volume cannot compensate for that. This difference between female and male voices is not found in practice. Besides, both 11 and 6 seconds is much too long for these low subglottal pressures. The probable cause of the much quicker exhaustion found in reality, is the presence of an open duct between the arytenoid cartilages, parallel to the glottis; the width of this duct, which may correspond to the ***breathiness*** of the voice, happens to be smaller for male voices, so that a man can keep up with a woman although he spills more air through his glottis.

5.7 Varying lung pressure

This section describes what happens with the sound intensity and the vibration frequency of the vocal folds, if the lung pressure is varied (from figure 5.8, we can already guess that the F_0 drops with 120 Hz per kPa as a function of the subglottal pressure during normal phonation, and that F_0 does not depend on the subglottal pressure during falsetto). To this end, we simulated 1 second of [a], with a *lungs* parameter that falls linearly from +0.5 (at 0 seconds) to –0.5 (at 0.5 s), and rises again linearly to +0.5 (at 1 s); the vocal-fold muscles were kept inactive.

Figure 5.9a shows the resulting lung pressure as a function of time: it varies pitch-synchronously. The same figure also shows a smoothed (§4.3) version of this, along a scale shifted up by 1 kPa; it has a maximum of 5.3 kPa and depends almost linearly on the *lungs* parameter. Figure 5.9b shows the sound signal at 1 metre in front of the speaker's face, and the sound intensity contour (§4.3). We see that the higher the lung pressure, the higher the sound intensity; however, the pressure-intensity curve (parametrized by time) in figure 5.9c, obtained from interpolating 1000 points on the smoothed lung-pressure contour and on the intensity contour, shows that below 1.3 kPa, the intensity varies by 15 dB/kPa, and that it varies by less than 3 dB/kPa for higher lung pressures.

Figure 5.10a shows the fundamental frequency of the vocal-fold vibration (§4.1). For this particular (default) setting of the *cricothyroid* (vocal-fold tension) and *posterioCricoarytenoid* (glottal width) parameters, there are, besides the normal periodic mode of vibration shown in 5.10b and 5.10e (for lung pressures below 3.2 kPa), several chaotic modes of vibration (the voiceless regions in 5.10a, mainly at pressures between 3 and 5 kPa), and an irregular periodic mode above 5.0 kPa, at 500 ms; figures 5.10cdfg show the acoustic signals and vocal-fold motions associated with these modes. Figure 5.10h shows the trajectory of this frequency with the smoothed lung pressure. Between lung pressures of 1 and 3 kPa, the pitch rises by 20 to 40 Hz/kPa, which is equal to the range found by Hixon, Klatt & Mead (1971).



Subglottal pressure + smoothed version (Gaussian smoothing window: 10 ms)

Acoustic output at 1 metre + intensity (Gaussian smoothing window: 20 ms)

Intensity as a function of lung pressure

Fig. 5.9 Subglottal pressure and sound intensity as functions of time and of each other, when the lungs behave as in §5.7. Inside the circles we find the smoothing functions: Gaussian on a linear scale (a), parabolic on a logarithmic scale (b).

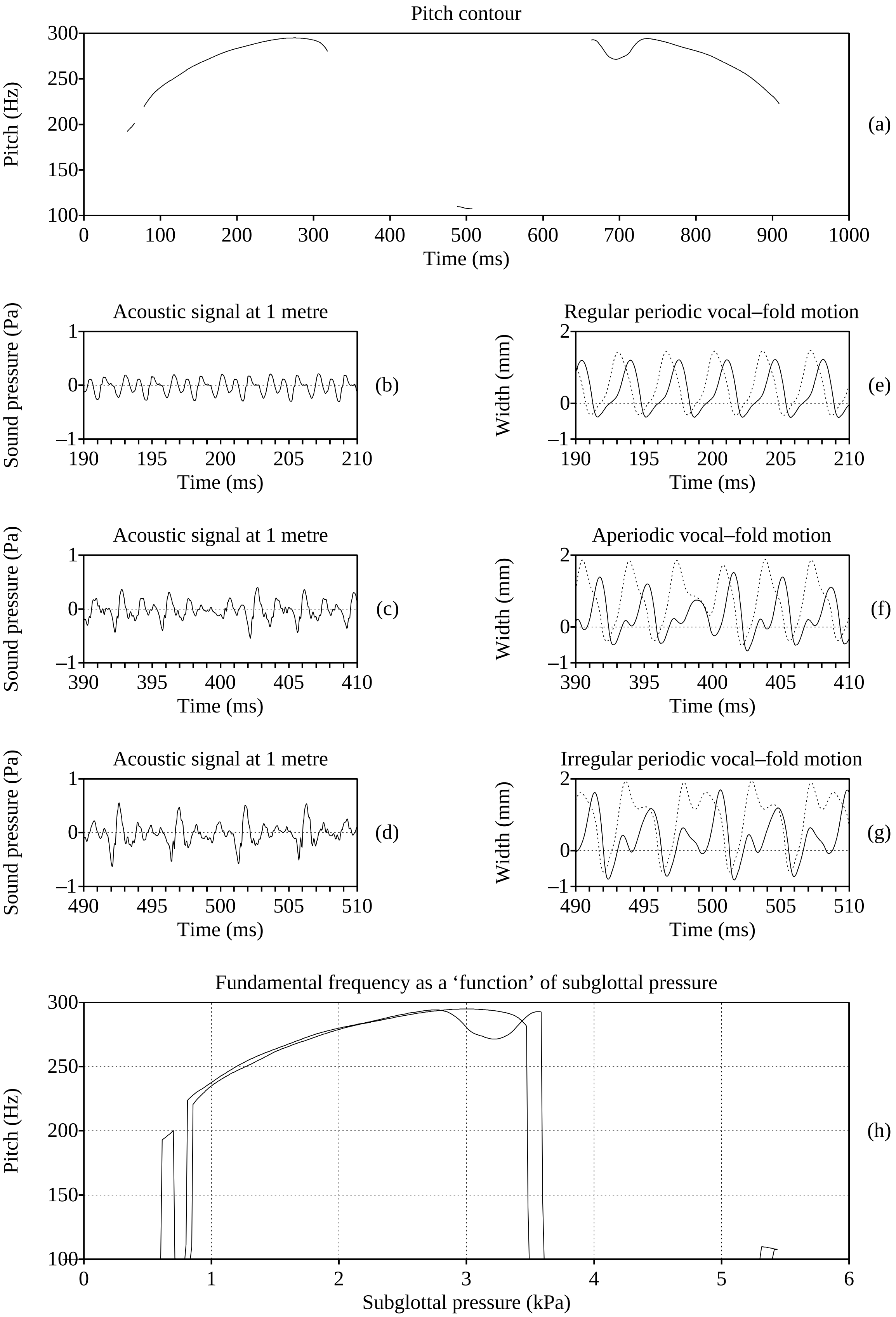


Fig. 5.10 Fundamental frequency of the acoustic signal, and modes of the vibration of the vocal folds, when the lungs behave as in §5.7. In the vocal-fold pictures, the dotted curves show the motions of the lower parts, the solid curves those of the upper parts of the vocal folds.

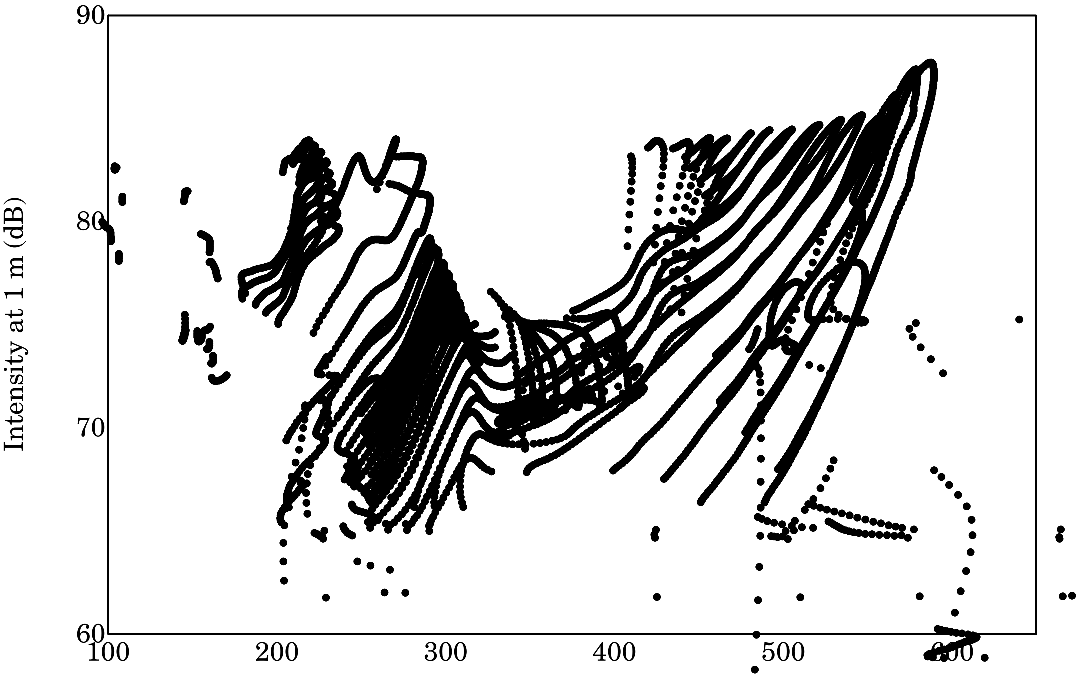


Fig. 5.11 Phonetogram: trajectories of fundamental frequency and intensity for various vocal-fold tensions, when the lungs behave as in figure 5.9, and the neutral glottal width is 0.

5.8 Phonetogram

If we repeat the experiment of §5.7 for different vocal-fold tensions, we find mode breaks as in figure 5.10, but for high values of the *cricothyroid* parameter, the fundamental frequency for the highest lung pressures is higher than the fundamental frequency for the lower lung pressures, in contrast with what we saw in figure 5.10. Also, the lung pressure associated with the main mode break becomes lower.

Figure 5.11 shows the trajectories of the fundamental frequency and the intensity for *cricothyroid* values between 0 and 1. The higher this parameter, the higher the fundamental frequency (with equal lung pressure).

The ***phonetogram*** of figure 5.11 shows three main modes of vibration:

- regular low, between 250 and 300 Hz (figure 5.10e); this occurs for combinations of low lung pressures and middle or low vocal-fold tensions.
- regular high, between 400 and 600 Hz; this occurs for a combination of high lung pressure and middle-high vocal-fold tensions, and for very high vocal-fold tensions with any lung pressure. As both parts of the vocal folds close, this mode is not a falsetto register.
- irregular, around 200 Hz (figure 5.10g); this occurs for combinations of high lung pressures and low or middle vocal-fold tensions.

5.9 Voicing in obstruents

While we discussed above the supralaryngeally simple phenomena of breathing and phonation, the rest of this chapter will show how our articulation model manages to simulate myoelastic-aerodynamic interactions above the glottis. This is the prime feature that distinguishes our model from other existing models of the articulatory-to-acoustic process, and was the rationale behind the very decision to implement the model.

In contrast with the production of vowels, there are various processes in the production of consonants that cannot be described exclusively in terms of an independent glottal source and supraglottal filter. First, we will look into the various ways in which our model speaker can cause her obstruent consonants to be voiced or voiceless. We will further look at our model's capabilities and limitations with respect to fricatives and nasals, at passive vibrations outside the glottis (trills), and at consonants that involve supralaryngeal pumping (ejectives) or suction (clicks).

As we saw earlier, the vocal folds vibrate if they are in the correct relative position and if the airflow between them is large enough. During a supraglottal closure, if everything else stays equal, the airflow decreases and the vocal folds are less likely to vibrate. Figure 5.12 shows what happens in our model speaker if the only active gesture she makes is a closing and subsequent opening of her lips. This gesture is superposed on the same [a]-like utterance that we used to investigate phonation in §5.4-8, i.e. a quick contraction of the lungs between 20 and 50 ms (fig. 5.12a), a 'hyoglossus' activity of 0.4 to pull the tongue down (fig. 5.12b), and a jaw-opening activity of 0.4 (represented as negative 'masseter' activity in fig. 5.12c). The lip closure is achieved by a simultaneous closing gesture of the jaw (fig. 5.12c) and a closing gesture of the lips (fig. 5.12d), both occurring between 100 and 200 ms. This gesture effectively closes the lips, though not as firmly as in figure 5.4. The muscle activity is continued between 200 and 300 ms, and a symmetric opening gesture occurs between 300 and 400 ms, after which the target positions are again those of the original [a]-like tongue shape. The actual width of the lips follows the target values with some delay, because of the inertia of the wall masses (fig. 5.12e; as before, the values can be negative).

As in §5.3, where we tried to 'inflate' the oral cavity by means of a bilabial closure, the supraglottal pressure becomes high when the mouth is closed (fig. 5.12f from 0.20 to 0.34 seconds). There seem to be two phenomena that inhibit voicing, and both are consequences of this rising supraglottal pressure:

- The intraglottal pressure, which, if we neglect the Bernoulli forces, is approximately the *average* of the sub- and supraglottal pressures, becomes high (fig. 5.12g from 0.20 to 0.34 seconds), thus pushing the vocal folds apart (fig. 5.12h from 0.20 to 0.34 seconds). So we see that even if the speaker does not actively widen her glottis, the glottis will still be widened by the changing air pressures, and to a width (0.6 mm in fig. 5.12h) that slightly surpasses the maximum amplitude during unobstructed phonation (0.4 mm in fig. 5.12h). It thus becomes more difficult for the Bernoulli forces to pull the vocal folds together.

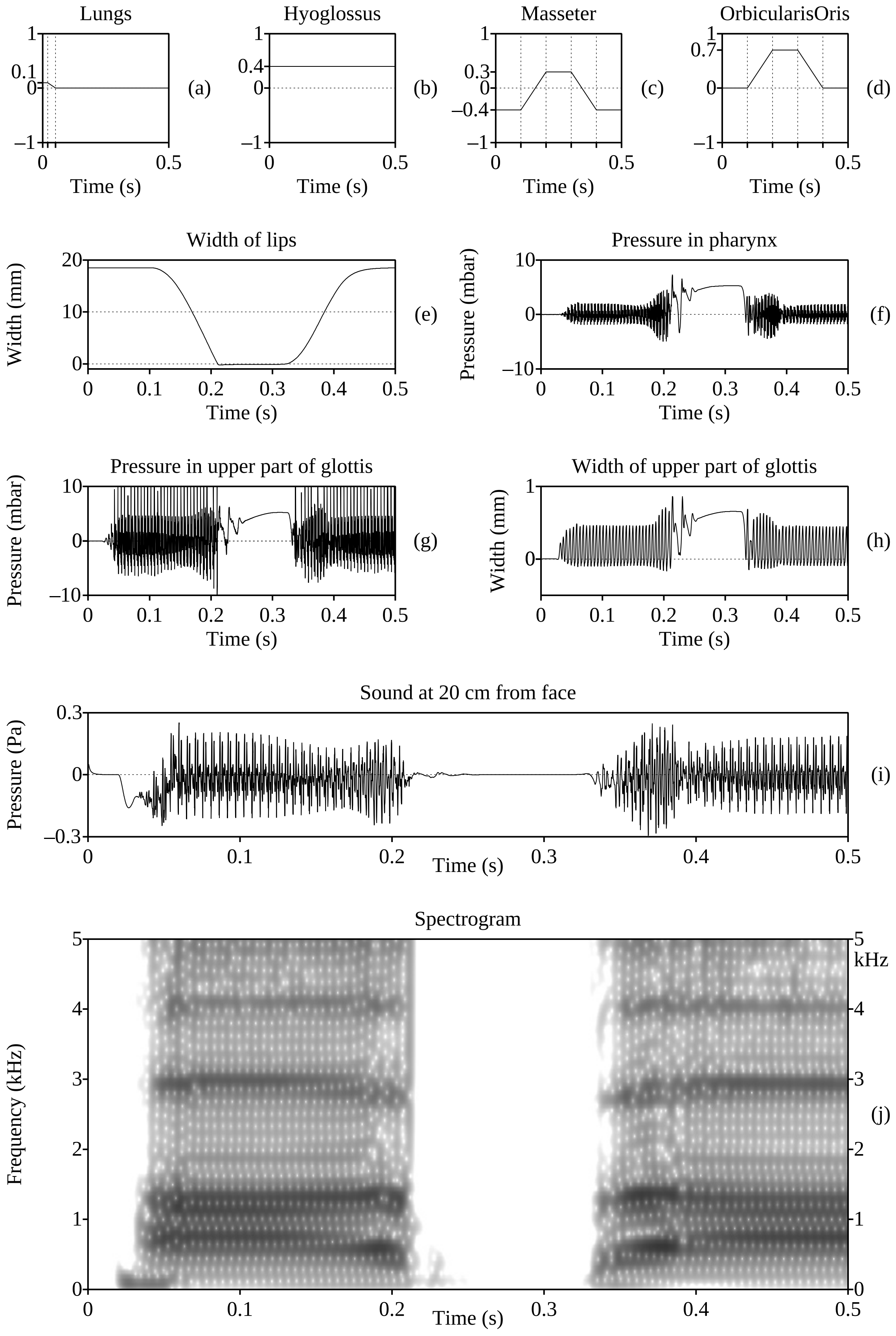


Fig. 5.12 An [ab̥a]-like utterance simulated with our model: its articulatory gestures (a-d), its myoelastics (e, h), its aerodynamics (f, g), and its acoustics (i, j).

- Because the transglottal pressure, which is the *difference* between the sub- and supraglottal pressures, decreases, the glottal flow diminishes. The Bernoulli forces, which are proportional to the square of the particle velocity through the glottis, are thus less capable of delivering energy to the vibration.

The resulting sound (oscillogram in fig. 5.12i, a spectrogram with a 10-ms window in fig. 5.12j) shows the formants associated with [a] (750 and 1350 Hz) and a hint of their transition from lower values in the neighbourhood of closure (at 0.35 s). Furthermore, the sound is voiceless throughout the bilabial closure, which suggests the transcription [apa]. The [p], however, is not exactly the *p* known from Dutch or French, which is actively devoiced to enhance the contrast with its voiced counterpart *b*; if it were this *p*, it would have shown a strong noise burst at the bilabial release. Rather, voicing starts immediately after the release, giving the sound written as *b* in southern German or English, which only have to contrast it with an aspirated [p^{h}]. Also, it is similar to the sound that young children produce in the babbling stage of speech development, when they cannot yet synchronize laryngeal with oral gestures; it is not by accident that their prototypical reduplicative utterance is traditionally written as *dada* in English but *tata* in Dutch.

The kind of plosive described here, with its ‘passive’ larynx, is traditionally called ***lenis voiceless***, and can be transcribed explicitly as [b̥]. The following sections describe various strategies that speakers can use to make stop consonants more voiced or more voiceless.

5.10 Voicing contrast using glottal width: aspiration

During a long supraglottal closure after pronouncing a vowel, the vocal folds will automatically stop vibrating after the airflow has fallen below a certain value. However, during the first part of the closure, while the supraglottal pressure is still relatively low, the vocal folds will vibrate, as was seen in figure 5.12h from 0.20 to 0.24 seconds. In order to make a genuine voiceless stop consonant, the speaker can actively make the folds stop vibrating by widening her glottis. She can do this by pulling her posterior cricoarytenoid muscles and at the same time relaxing the muscles that have brought together the vocal folds, like the interarytenoid muscles. The vocal folds will be too far apart to vibrate; the likely acoustic effect if this articulatory trick is called ***aspiration***.

Figure 5.13 shows our simulation of this phenomenon. All gestures are the same as in figure 5.12, except that the ‘posterior cricoarytenoid’ activity (fig. 5.13a) rises and falls with a timing exactly synchronous to the jaw and lip target movements. This laryngeal gesture is followed without much delay by the actual width of the glottis (fig. 5.13b). We see in fig. 5.13c that normal voicing stops soon after the opening gesture of the glottis is initiated: at 0.13 seconds, breathy voicing starts, and voiceless aspiration takes over from 0.16 seconds. In all, we can say that there is 70 ms of pre-aspiration. After the bilabial release, we see another 40 ms of aspiration (between 0.35 and 0.39 seconds).

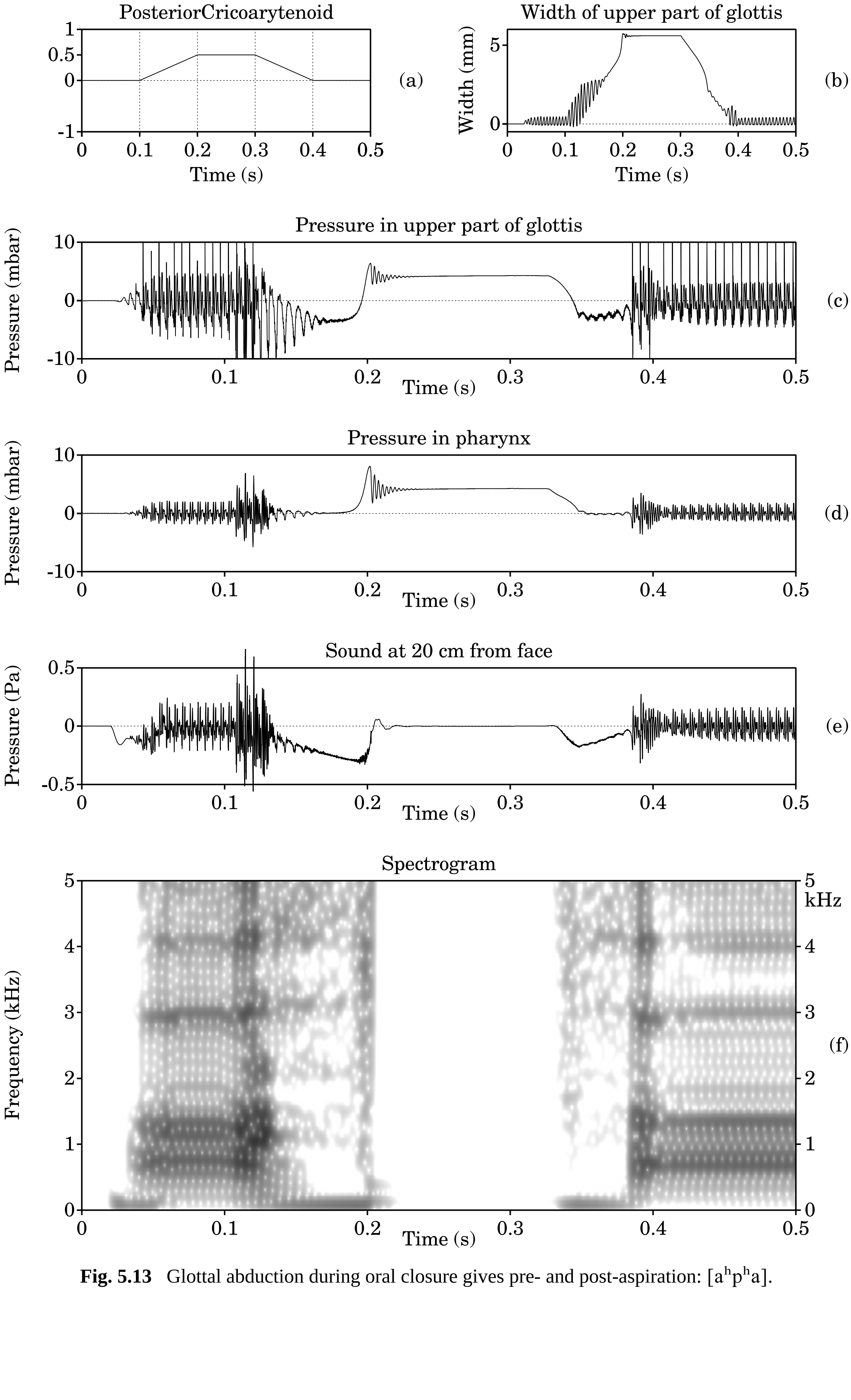


Fig. 5.13 Glottal abduction during oral closure gives pre- and post-aspiration: [a^{h}p^{h}a].

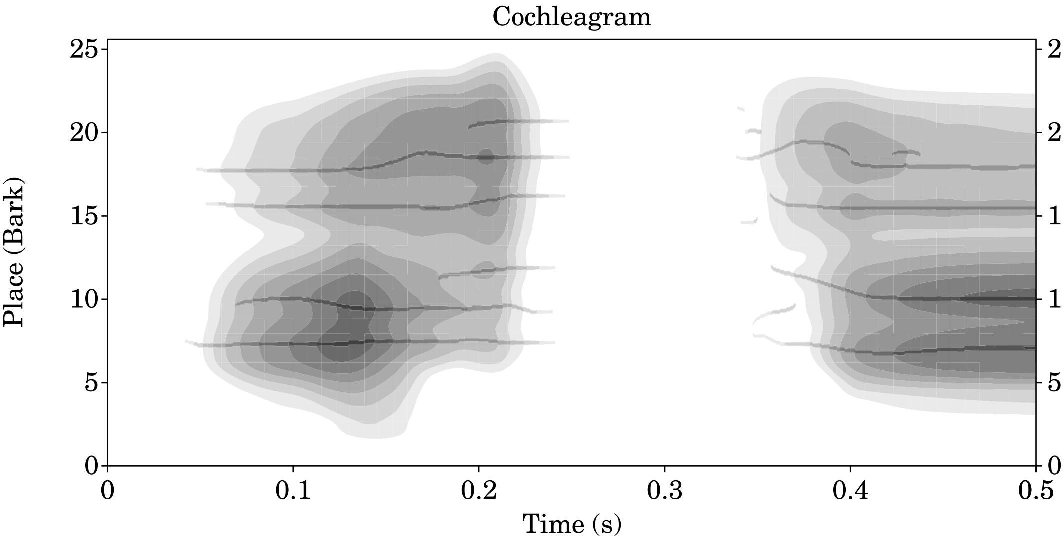


Fig. 5.14 A cochleagram (perceptually based spectrogram) of the simulated utterance of figure 5.14.

The post-aspiration lasts shorter than the pre-aspiration, because the lips are relatively slow in following the muscle commands, and the vocal folds are relatively fast, while the timing of the target positions is the same. This is an effect of the response times in our model being implicitly tied to the resonance frequencies of the walls, which are about 30 Hz and 170 Hz for the lips and vocal folds, respectively. In other words, the active restoring force (the gamma loop) equals the passive relaxation force, for the same deviation from the target position (or equilibrium position). In real languages, the timing of the lip gesture is often very different from that of the glottal gesture: if the plosive is at the beginning of a syllable, we mainly find post-aspiration.

As we can see in the spectrogram (fig. 5.13f), high frequencies are well represented in the aspiration noise (drawing pre-emphasis was 6 dB/octave), and some of the format structure is seen even in the completely voiceless part.

The fundamental frequency drops from 166 Hz before the bilabial closure to 159 Hz after. Between 0.20 and 0.25 seconds, we see (figs. 5.13c and d) a reverberation at 418 Hz, which is probably due to the sudden halt of the motion of the vocal folds (at 0.20 seconds in figure 5.13b), which should have been modelled in a smoother way. We are lucky that this tone does not make it into the outer air. However, it is superposed on a 17-Hz damped sine, visible in figs. 5.13c and d, but more clearly in fig. 5.13e. This is an oral wall vibration, not audible because of its low frequency, but still apparent in the spectrogram, together with the still lower frequency associated with the DC shift in the acoustic sound between 0.13 and 0.20 seconds (fig. 5.13e). The pre-emphasis seems to lift these very low tones up from the bottom edge in figure 5.13f. A spectrogram made from a natural speech utterance would not show these tones, because microphones do not normally transmit these low frequencies. We do show these tones here, because we are trying to imitate the capabilities of human speech, not the limitations of recording devices. Of course, more perceptually based representations of the sound, like the cochleagram of figure 5.14 (see §4.2), would not show these low frequencies, either, since the human ear is rather insensitive to them.

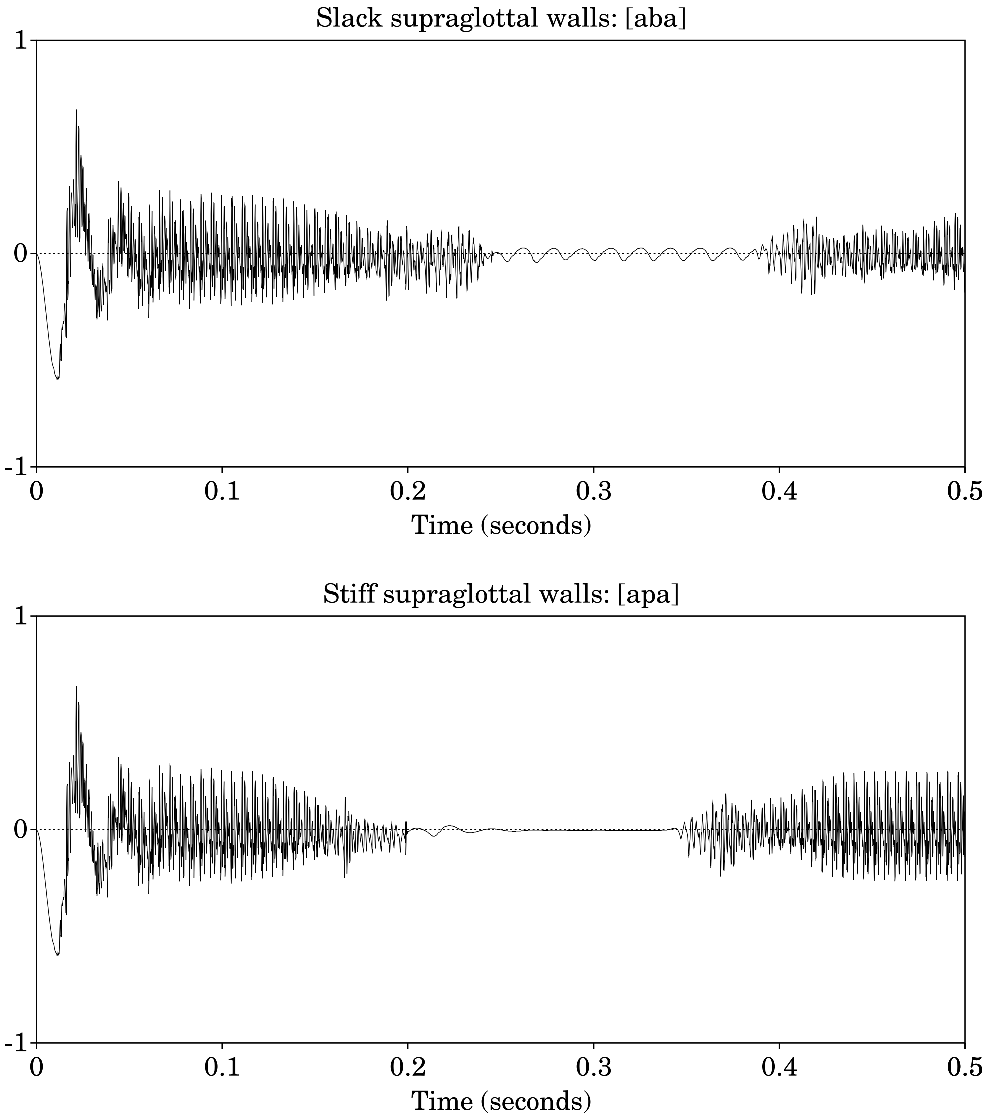


Fig. 5.15 The workings of a supraglottal wall-stiffness contrast. Cheek stiffness above: 0.4 bar/m; below: 3 bar/m. Notice the very low F_0 during labial closure.

5.11 Voicing contrast using tract-wall tension: fortis - lenis

In order to prolong voicing during a bilabial closure, a speaker can try to maintain a sufficient amount of airflow by enlarging the supraglottal cavities. One way of getting this done is letting the rising oral pressure inflate the walls of the oral cavity. Figure 5.15 shows what happens if we diminish the stiffness of the walls, including the lips, during the whole utterance from 300,000 N/m^3 (= 3 bar/m) to 40,000 N/m^3 (= 0.4 bar/m). Voicing continues throughout the slack-wall utterance, which can therefore be transcribed as [aba]; I will transcribe the utterance with less voicing as [apa].

We see that the first vowel in [aba] is much longer than that in [apa]. It is a known fact that in many languages vowels are somewhat longer before voiced obstruents than

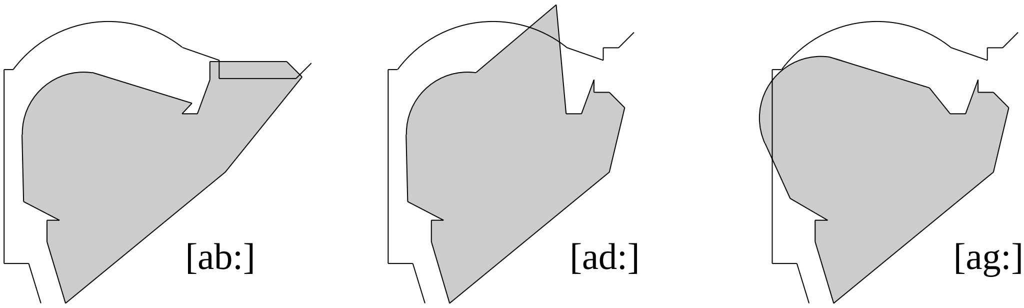


Fig. 5.16 Three oral occlusions with different volumes between them and the larynx.

before voiceless obstruents. Probably because in English this phenomenon is exceptionally clear, many explanations have been proposed, and we can now blithely add one. We note that for both utterances in figure 5.16 the muscle gestures were simulated as exactly synchronous. The timing difference in the acoustic output, therefore, must be a result of the lips being slower in [b] than in [p], in following the muscle commands. Our model has this peculiarity because the muscles and the walls are modelled by the same springs, so that if the wall stiffness changes, the target-seeking muscle force changes, too. In physiological terms, this would mean that the gain of the gamma loop is proportional to the passive relaxation force, and that if either of these is related to motoneuron activity, the other must also be, and in the same way. Though such a relation sounds plausible, it may not be an appropriate description of all muscles.

Another common difference between voiced and voiceless plosives is that voiceless plosives tend to be longer. Since figure 5.15 suggests that the closure durations are equal if the articulatory gestures are timed equally, the difference is probably externally planned, with the likely objective of enhancing the perceptual voicing contrast.

5.12 Place-dependent maintenance of voicing

After the speaker closes the lips in [abː], voicing will soon stop if she does not adjust her laryngeal settings, because the vocal folds will separate by a millimetre or so as a result of the rising pressure, as we saw in figure 5.12. To continue voicing, the speaker can adjust the equilibrium width of the vocal folds, perhaps by increasing the *interarytenoid* parameter from 0.5 to 0.53, to compensate for that millimetre. After this, we expect that voicing will continue until the average pressure in the cavities between the larynx and the constriction will have risen to a point where the transglottal pressure is too low for passive vibration to continue.

Because the cavity behind a labial plosive is larger than the one behind a dorsal plosive, we expect that voicing proceeds longer in [abː] than in [agː]. To test this, we compare the three possible oral occlusions in figure 5.16. The laryngeal and velopharyngeal settings are as in the previous examples, *hyoglossus* activity is set to 1 throughout the utterances, and after stable phonation has evolved, the *orbicularisOris, upperTongue*, or *styloglossus* parameter is raised from 0 to 1 within 0.1 seconds, while the *interarytenoid* parameter is simultaneously raised from 0.5 to 0.53.

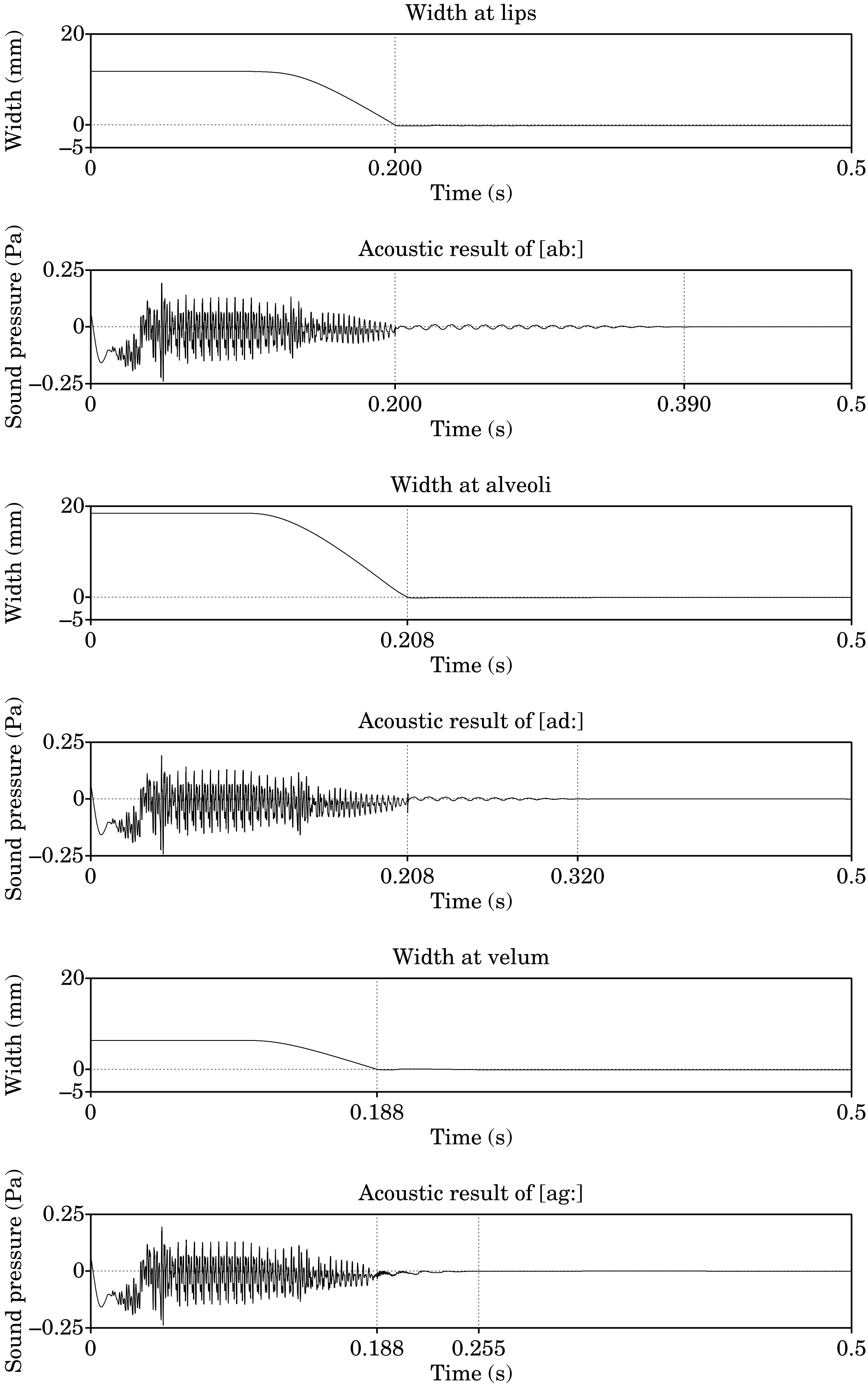


Fig. 5.17 The continuation of plosive voicing as a function of place.

Figure 5.17 shows the widths of the tube sections that experience the first closure in these three cases. These tube sections are the 26th, 22nd, and 11th sections in figure 2.10, respectively, counted from the larynx.

From figure 5.17, we can read that voicing continues for 190, 112, and 67 milliseconds after the closure, for [abː], [adː], and [agː], respectively. This confirms the findings of Ohala & Riordan (1979). These differences in voicing of [b], [d], and [g] lead to a hierarchy of articulatory effort for the implementation of the voicing feature for the various plosives, and to a hierarchy of perceptual confusability for voiced/voiceless pairs of dorsal, coronal, and labial plosives. In chapters 11 and 17, we will see linguistic evidence of these hierarchies.

5.13 Voiceless by glottal constriction: ejectives

Another way of making sounds voiceless is by firmly constricting the glottis (effort closure). A plosive stop that uses this can have a release burst either if the glottal constriction is released just before the oral constriction, so that the lung pressure may produce a burst, or if the glottal constriction is released after the oral constriction and there is an alternative mechanism to raise the pressure behind the constriction, like narrowing of the pharynx or raising of the larynx.

5.13.1 The production of ejectives

Ejective plosive stops are voiceless plosives with a simultaneous glottal stop and a raising of the larynx.

According to Westermann & Ward (1933), “the mouth closure is generally released half a second before the glottal closure”. Halle & Stevens (1971) amend this statement by telling us that “there is a delay of 50-odd ms before the adducting glottal musculature can be relaxed and the glottis can assume a configuration appropriate for the onset of vocal-cord vibration”. Catford (1977) reports having measured a great variation for this interval in Caucasian languages, “ranging from only 12 ms in Abkhaz, through 28 ms in Kabardian, and 70 ms in Chechen, to about 100 ms in Avar and in the Bzhedukh dialect of Adyghe”.

During the time that both the mouth and the glottis are closed, the larynx is raised (though Westermann & Ward (1933) did not mention this, as noted by Hayward (1993)), which causes a high pressure build-up in the pharynx. The empirical data of Pinkerton (1986) for several Mayan languages show that this pressure is typically between 1200 and 2000 Pa, whereas the pharyngeal pressure for plain voiceless plosives is between 600 and 1000 Pa. Ladefoged and Traill (1994) measure pharyngeal pressures up to 2000 Pa before the velar or uvular release of post-click ejectives in !Xóõ.

Because of the high pressure behind the oral constriction, the release burst of an ejective plosive is stronger than that of a plain plosive. However, this pressure, having only the pharyngeal cavity as its back-up reservoir, drops after the release more rapidly than the pressure of a plain release burst, which is maintained by the large reservoir of the lungs. Therefore, the noise burst will be shorter in ejectives. On the other hand, the

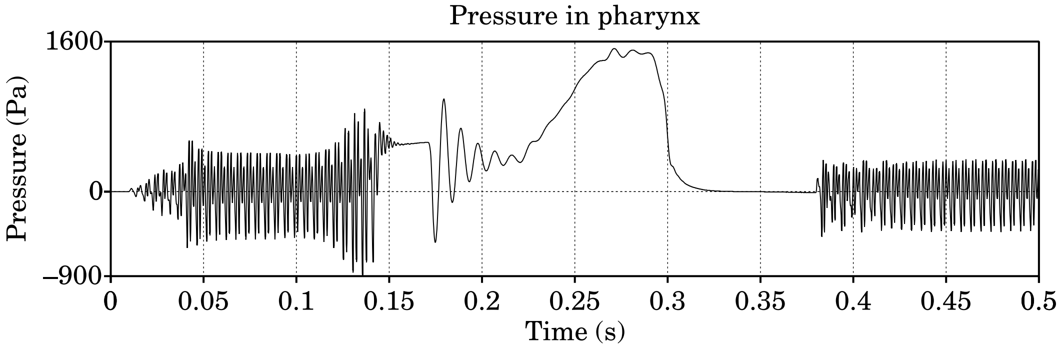


Fig. 5.18 The pressure in the pharynx with the production of an ejective dorsal stop.

supralaryngeal reverberations will last longer in ejectives than in plain plosives, because there is no glottal damping of the formants, the glottis still being closed. The perceptive impression thus raised is that of a “peculiarly sharp sound” (Westermann & Ward 1933). Catford (1977) describes a syllable in the following way: “in a sequence such as [p’a], the rather ‘hollow’ sounding ‘pop’ on release of the glottalic pressure stop [p’] is heard a moment before the glottal closure is released, in the form of an abrupt start of voicing”.

The high pharyngeal pressure could also be brought about by a sphincteric, instead of longitudinal, compression of the pharynx. Catford (1977) states that exactly this may occur as a secondary articulation. We should keep in mind, however, that some Caucasian languages contrast pharyngealized with non-pharyngealized ejectives (Kibrik & Kodzasov 1990).

5.13.2 Simulation of ejectives

Our articulatory synthesizer is designed to be strong in modelling automatically the interaction between tract shape and aerodynamics. In particular, it is supposed to correctly represent the aerodynamic effects of changes in the lengths of the tube sections. Therefore, it should have no trouble simulating the aerodynamic and acoustic properties of ejectives.

To test this, we synthesize from the following articulations:

Step 1. Phonate in the familiar way, using a *hyoglossus* of 0.5 and a *masseter* of –0.3 to simulate an open vowel.

Step 2. Oral closure. Move *styloglossus* from 0 to 1 between 0.1 and 0.15 s.

Step 3. Glottal closure. Move *interarytenoid* from 0.5 to 1 between 0.17 and 0.2 seconds. This step could also precede step 2.

Step 4. Raising of the larynx. Move *stylohyoid* from 0 to 1 between 0.22 and 0.27 seconds.

Step 5. Oral release. Move *styloglossus* back from 1 to 0 between 0.29 and 0.32 seconds.

Step 6. Glottal release during return of the larynx. Move *interarytenoid* back from 1 to 0.5 and *stylohyoid* back from 1 to 0 between 0.35 and 0.38 s.

Around t = 0.16 seconds, the pressure in the pharynx (fig. 5.18) equals the pulmonic pressure of 550 Pa. Between 0.22 and 0.27 seconds, the hyoid is pulled up and the pressure rises to 1500 Pa.

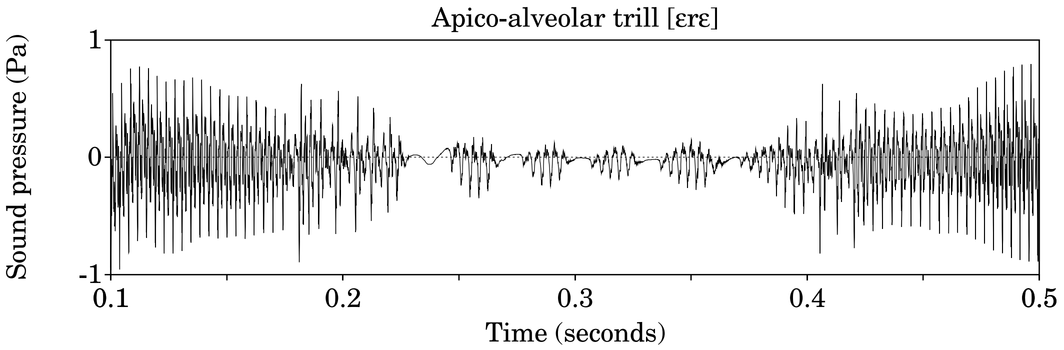


Fig. 5.19 Passive vibration outside the glottis.

5.14 Trills

Because the entire vocal tract has been modelled in the same way as the glottis, passively vibrating walls should be able to occur wherever the conditions are favourable. At the tongue tip, for instance, the “wall” mass is low, as is the wall stiffness because of the protruding shape. Also because of this shape, the damping is low. Thus, the conditions for strong myoelastic-aerodynamic interaction are fulfilled, although they are very different from those in the glottis. Figure 5.19 shows the acoustic output for a simulated utterance with an apico-alveolar trill superposed on a vowel, which can best be transcribed as [ɛrɛ].

In the example of figure 5.19, the trill is periodic, i.e., the trill frequency is a subharmonic of the vocal-fold frequency. This need not always be the case.

5.15 Clicks

We have already seen many cases of muscles controlling air pressure: contracting lungs, yielding cheeks, and rising larynx. In click consonants we see another example: the walls in between two constrictions are pulled apart, which causes the pressure in this cavity to fall, and an inward release burst powered by suction when one of the constrictions is opened.

The bilabial click involves simultaneous bilabial and velar closures. When the bilabial closure is released, we hear the characteristic ‘pop’ sound resulting from a bilabial burst reverberating in the oral cavity which is still closed at the velum. Figure 5.21 shows the gestures involved. Figure 5.20 shows some of the vocal-tract shapes and acoustics.

As we see in the spectrogram, the release initially causes a short burst with a large frequency content. After this, we hear a sine-like sound with a frequency that rises from 300 Hz to 1000 Hz.

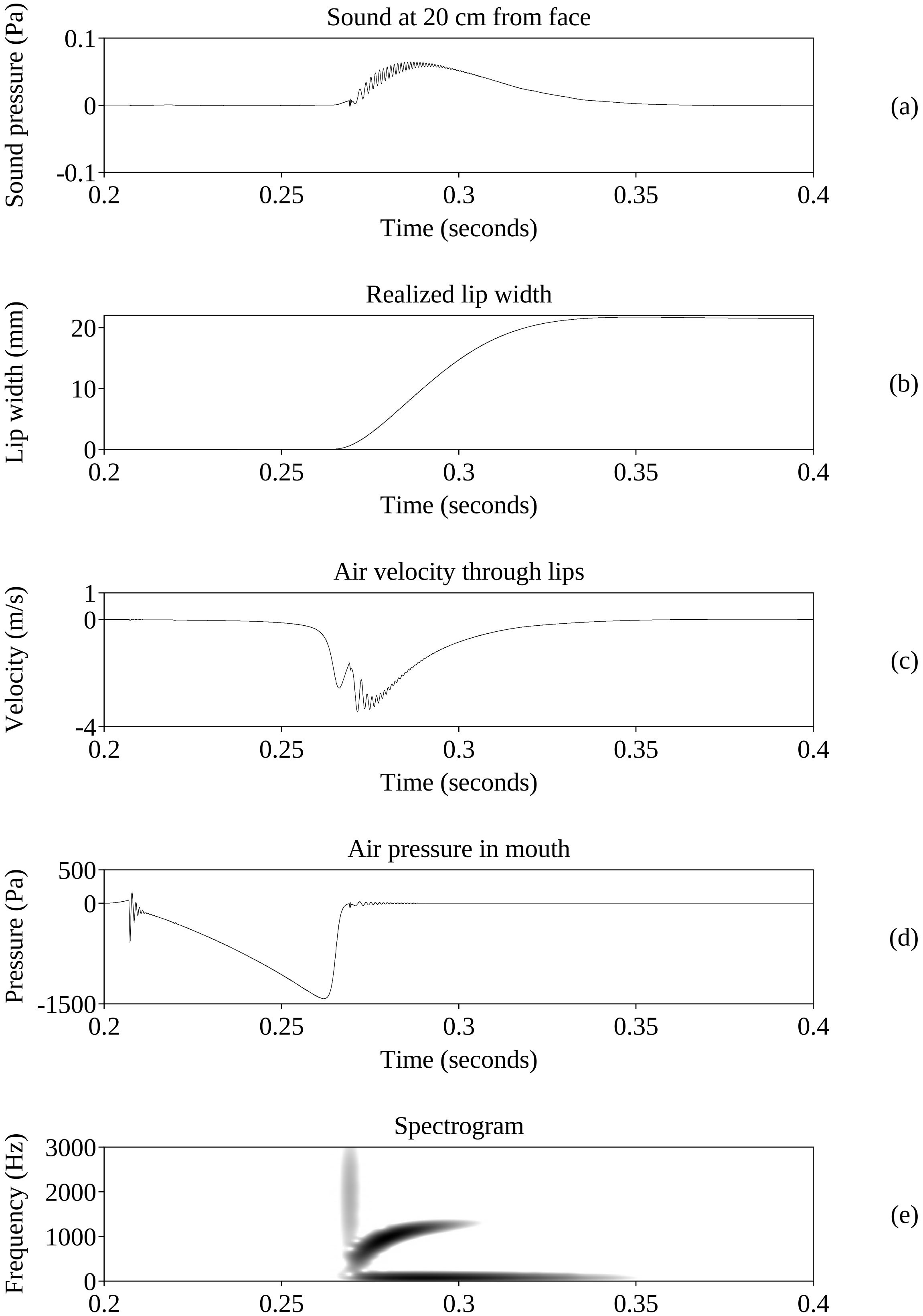


Fig. 5.20 Myoelastic, aerodynamic and acoustic results of a bilabial click.

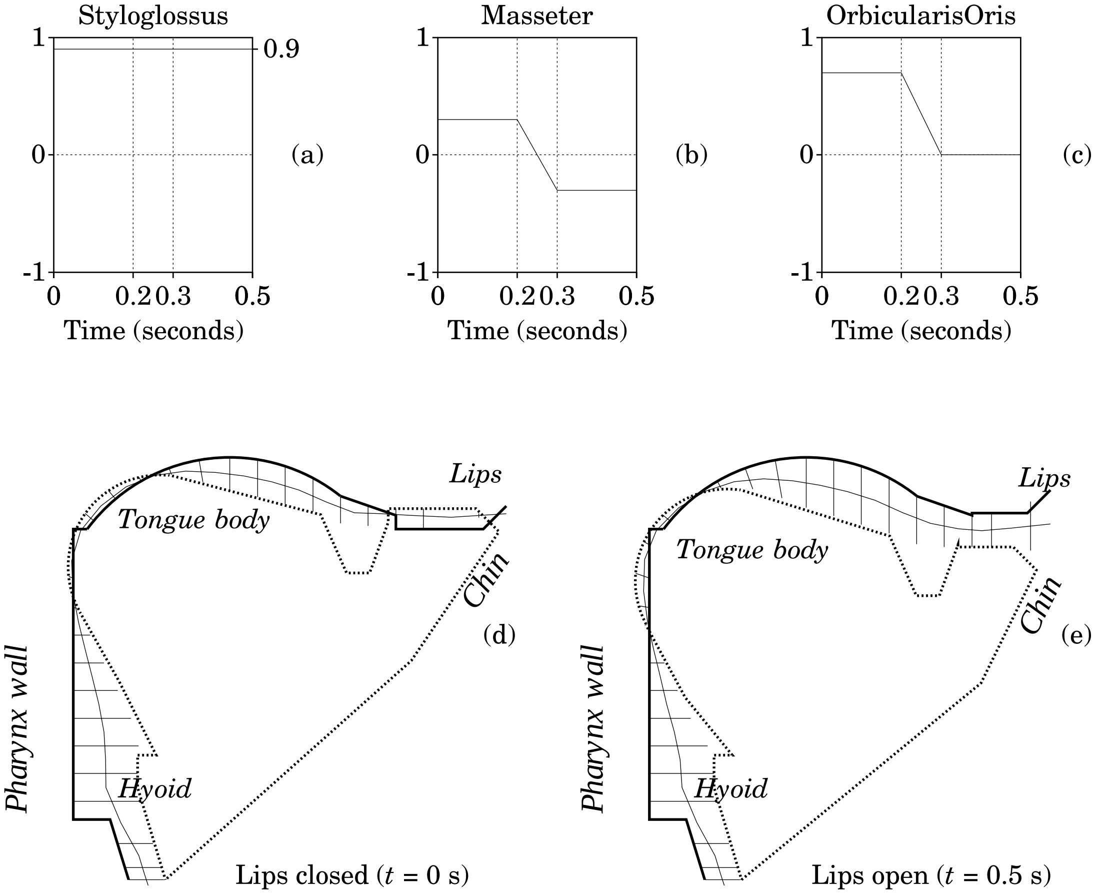


Fig. 5.21 The articulations involved in making a bilabial click consonant.

5.16 Conclusion

Our articulation model managed to simulate realistically several speech phenomena that could not be handled by previous models: the dependence of voicing on oral wall tension, the dependence of the maintenance of voicing on the size of the supralaryngeal cavity, pumping and sucking effects caused by length changes (ejectives), and pumping and sucking effects caused by wall movements (clicks). This gives us the reassurance that we can use the model profitably in parts II and III of this book. Among the speech phenomena that stay problematic, we find: effects of crucial two-dimensional shape of cross sections (laterals; some glottis configurations), and non-local noise generation (sibilant fricatives).

Part II

CONSTRAINTS

Chapter 6 will show that the constraint-ranking formalism of Optimality Theory is particularly suited for marrying “phonetic” explanation with “phonological” description. Chapters 7 to 10 derive several constraint families from common properties of human motor behaviour and perception. Chapters 11 to 13 discuss some consequences of the theory in the realms of typology, segmentality versus autosegmentality, and underspecification.

Together with chapter 1, chapters 7 to 13 appeared on Rutgers Optimality Archive in February 1997 (Boersma 1997a).

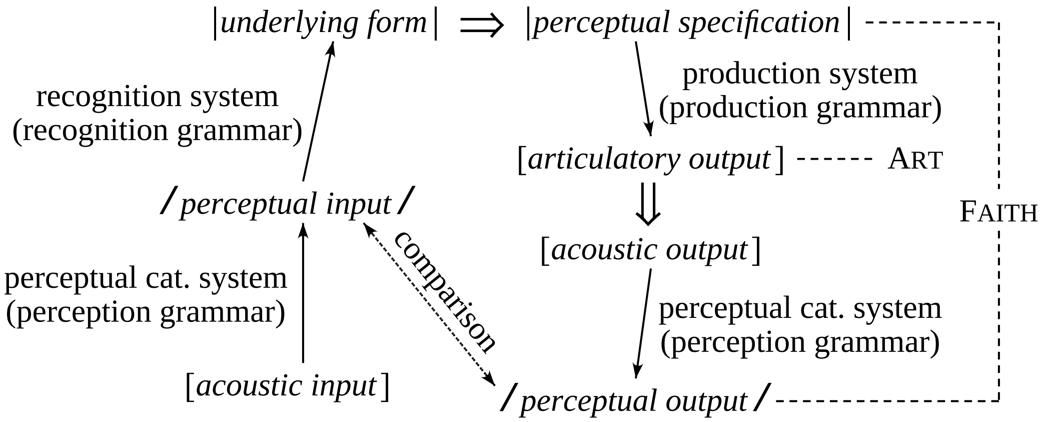
6 Functional Optimality Theory[1]

Abstract. This chapter briefly introduces Optimality Theory from the functionalist viewpoint.

The functional hypothesis for phonology (Passy 1891) maintains that sound structures reflect an interaction between the articulatory and perceptual principles of efficient and effective communication. The theory that I develop in this book maintains that this interaction is directly reflected in the grammar: it handles substance-related phonological phenomena within the constraint-ranking framework introduced by Optimality Theory (Prince & Smolensky 1993), but without the need for positing innate features and hierarchies; if restricted to gestural and faithfulness constraints, its scope equals that of autosegmental phonology and feature geometry.

6.1 Grammar model

As defended in chapter 1, Functional Phonology makes a principled distinction between articulatory and perceptual representations and features. The grammatical correlates of the speech production and perception processes illustrated in figure 1.1, are depicted in figure (6.1), which shows the concept of the linguistically relevant systems, processes, and representations of the speech production and perception systems of a single speaker-listener, to the level of precision that we will need in this book:



(6.1)

In (6.1), we see the following representations:

(1) The ***acoustic input*** of the speech uttered by another person, as presented to the ear of the listener; written between brackets because it is a language-independent representation.

[1] This chapter appeared as Boersma (1997e).

(2) The ***perceptual input***: the speech uttered by another person, as perceived by the listener, in terms of perceptual features (periodicity, noise, spectrum) and their combinations; written between slashes.
(3) A ***perceptual specification*** (§1.3.3) in terms of perceptual features, as stored in the language user's lexicon as an ***underlying form***; written between pipes.
(4) The ***articulatory output*** of the speaker (§1.3.3), in terms of articulatory gestures (articulator positions, muscle tensions) and their combinations; written between brackets.
(5) The ***acoustic output*** of the speaker: an automatic result of her articulatory output (ch. 2 & 3); also written between brackets.
(6) The ***perceptual output*** of the speaker (§1.3.3): her acoustic output, as perceived by herself; written between slashes.

Figure (6.1) also shows the following processing systems:

- The speaker's ***production system*** determines the surface form of the utterance from an underlying perceptual specification.
- The listener's ***perceptual categorization system*** determines how a listener converts the raw acoustic input to a more perceptual representation; she uses this system for the acoustic input from other speakers as well as for her own acoustic output.
- The listener's ***recognition system*** converts the perceptual input into an underlying form (and helps the categorization system).
- A ***comparison module*** on behalf of language acquisition. If the learner's output, as perceived by herself, differs from the adult utterance, as perceived by the learner, the learner will take a learning step (chs. 14-15).

The abbreviations ART and FAITH refer to ***articulatory*** and ***faithfulness*** constraints, as explained below.

6.2 Constraint-ranking grammars and functionalism

Consider the process of place assimilation of nasals in Dutch. The words /tʀɛin/ 'train' and /pɑkə/ 'catch' will often be concatenated as /tʀɛimpɑkə/. The process is confined to the coronal nasal: bilabial nasals, velar nasals, and non-nasals at any place, do not usually assimilate place.

6.2.1 Explanation versus description

A phonetic explanation for these facts can readily be given: as compared to the articulation [tʀɛinpɑkə], the articulation [tʀɛimpɑkə] saves the speaker one tongue-tip gesture, since the bilabial gesture for [m] was already needed for [p]; the perceptual loss of this assimilation is the neutralization of any specified |n| and |m|, which could lead to confusions between words that end in these sounds, and to extra required effort in the recognition system. The restriction to nasals can be explained by the fact that e.g. the nasals /m/ and /n/ are perceptually much more alike than the plosives /p/ and /t/, so that the listener will rely less on place information for nasals than for plosives, so that the speaker has more freedom to mispronounce a nasal than a plosive; the restriction to coronals can be explained by the fact that coronals are much more common than labials, so that the listener will have a bias towards recognizing a coronal instead of a labial, so that the speaker will assume that the listener will reconstruct the coronal even if she pronounces it as a labial (I will return to these explanations in chs. 9 and 15).

These explanations, however, do not tell us what a speaker does when she has to concatenate the words /tʀɛin/ and /pɑkə/, and this is why phonetic explanations have seldom met with enthusiasm on the part of linguists.

Until 1993, linguists tended to describe phonological processes with rules, e.g., they would describe nasal place assimilation with a structure-changing recipe like "n → m / _ p", or with a generalization over the various places, or with a feature-filling recipe like "[0place] → [αplace] / _ [αplace]", or in an autosegmental and/or feature-geometric formulation. All these notational variants, however, are still recipes and have little explanatory power. So the explanatory and descriptive accounts had been divorced for a long time.

6.2.2 Constraint-ranking grammars

The advent of Optimality Theory (Prince & Smolensky 1993; McCarthy & Prince 1993a, 1993b, 1994, 1995) changed this situation, by making ***constraints*** instead of rules central to the grammar. A traditional Optimality-Theoretic account of nasal place assimilation would have that a universal constraint NASSIM ("nasals have the same place as a following consonant") is ***dominating*** the universal constraint IDENT (place) ("the surface place is equal to the underlying place specification"). Since these constraints are ***violable***, the outcome depends on their ***rankings***, so that we have the following mini-typology: if NASSIM outranks IDENT (place), there will be assimilation; if, on the other hand, IDENT (place) dominates NASSIM, there won't.

A constraint like NASSIM still provides no explanation: it is still purely descriptive. But instead of these allegedly universal constraints, we can directly translate the phonetic principles of minimization of effort and perceptual confusion into the grammar, namely, into ***articulatory constraints*** ("ART" in figure 6.1), which evaluate articulatory outputs, and ***faithfulness constraints*** ("FAITH" in figure 6.1), which evaluate the similarity between the specification and the perceptual output.

For nasal place assimilation, the relevant articulatory and faithfulness constraints would be (the asterisk can be read as "no" or "don't", or simply "star"):

*GESTURE (tongue tip: close & open):

"do not make a tongue-tip opening and closing gesture"

*REPLACE (place: coronal, labial / nasal / _ C):

"do not implement a perceptual coronal place specification as something that will be heard as labial place, for a nasal, before a consonant"

The Dutch assimilation process can then be seen as the result of the following grammar of ranked constraints (I will freely abbreviate constraint names):

Nasal place assimilation

*REPLACE (cor / plosive)
*GESTURE (tip)
*REPLACE (cor / nasal)

(6.2)

Because plosives do not assimilate, the constraint *REPLACE (place: coronal, labial / plosive / _ C) must be ranked higher than *GESTURE (tongue tip). Note that the ranking of *REPLACE (place / plosive) above *REPLACE (place / nasal) reflects the asymmetry of perceptual confusion discussed above, so that we may well hypothesize that this ranking is nearly universal. Indeed, if we could find out what rankings are universal and what rankings can be set on a language-specific basis, we would have a typologically adequate account of possible and impossible sound systems, which, in my view, is an important goal of phonological theory.

Thus, violable constraints can be expressed in such a general way that they yield to the linguist's requirement of universality and simplicity, and to the phonetician's requirement of explicability in terms of the properties of the human speech mechanism. In part II of this book, I will identify these functional constraints and investigate their interactions; in part III, I will show their empirical adequacy.

6.3 The production grammar

So I will assume that the speaker's production system can be described by an Optimality-Theoretic production grammar.

A typical production process can thus be represented with the following Optimality-Theoretic ***tableau***:

|*spec*|	A	B
☞ [art_1] /$perc_1$/		*
[art_2] /$perc_2$/	*!	

(6.3)

This tableau shows the following representations, each of which can be identified in figure (6.1):[2]

(1) A perceptual specification *spec*.
(2) Many candidate articulatory outputs art_i.
(3) For each candidate articulatory output art_i: the corresponding perceptual output $perc_i$.

In tableau (6.3), the two constraints A and B both issue a ***protest*** against a certain candidate, as shown by the asterisks (the ***marks***). Because A is ranked higher than B, the

[2] Where there is no change of confusion, I will use a single shorthand for the articulatory and perceptual outputs (put between brackets), and often write the perceptual input between equally traditional slashes, with the understanding that it is a shorthand for a collection of specified perceptual features.

disharmony associated with its violation is greater than that of B, and its violation becomes the ***crucial violation*** for candidate 2, as shown by the exclamation symbol, which is put after the ***crucial mark***. Thus, candidate 1 is more ***harmonic*** (less offensive) than 2, so it becomes the ***winner***, as shown by the pointing finger. Some cells are grey because any violations that might occur in these cells cannot contribute to determining the winner.

Our example of nasal place assimilation is written as

|an+pa|	*GESTURE (tip)	*REPLACE (cor)
[anpa] /anpa/	*!	
☞ [ampa] /ampa/		*

(6.4)

The candidate [ampa] (shorthand for "pharyngeal narrowing plus lip closure and opening plus velum raising..."), which is perceived as /ampa/ (shorthand for "high F_1 plus labial place plus nasality..."), is the winner.

6.4 The perception grammar

We will likewise assume that the listener's categorization system can be described by an Optimality-Theoretic ***perception grammar***.

We can thus represent a typical categorization process with the following tableau:

[*ac*]	A	B
☞ $/cat_1/$		*
$/cat_2/$	*!	

(6.5)

This tableau shows the following representations (visible twice in figure 6.1):

(1) An acoustic input *ac*.
(2) Several candidate perceptual categories cat_i.

For instance, on the perceptual tier F_1 (first formant), the listener may have three categories of 300, 500, and 700 Hz (for high, mid, and low vowels, respectively). If the acoustic input is 440 Hz, a relevant constraint (ch. 8) is:

*WARP (F_1: [440], /300/):
"do not initially classify an acoustic input of 440 Hz as a high vowel"

The decision of the categorization system can now be described with the following tableau, if the system is trying to initially classify any acoustic input into the "nearest" category:

[440]	*WARP ([440], /700/)	*WARP ([440], /300/)	*WARP ([440], /500/)
/300/		*!	
☞ /500/			*
/700/	*!		

(6.6)

The winner is the category /500/, i.e., the input of 440 Hz is initially perceived as a mid vowel (the recognition system may correct this initial categorization on the basis of other information).

6.5 Conclusion

The hypothesis of Functional Phonology is that the production and categorization systems can be described with Optimality-theoretic constraint-ranking grammars that contain direct translations of principles of minimization of articulatory effort and perceptual confusion. This is an empirical hypothesis, since the Optimality-theoretic maxim of strict ranking predicts a restricted typology of possible languages; our functional version will add to this restrictivity, by proposing a principle that accounts for universal rankings of some constraint pairs (chapter 11).

This hypothesis will be worked out in the remaining chapters of part II, and its descriptive adequacy will be established in part III.

7 Articulatory constraints

Abstract. This chapter formalizes the principle of minimization of articulatory effort into gestural constraints and their universal local rankings.

In his *Dictionary of Phonetics and Phonology*, Trask (1996) calls the principle of *maximum ease of articulation* “A somewhat ill-defined principle sometimes invoked to account for phonological change”. In this chapter, I will formalize effort, and turn it into a well-defined principle that will be seen to work for phonetic implementation (ch. 10), segment inventories (ch. 16), and autosegmental processes (ch. 18, 19).

As we will see below, previous attempts to formalize articulatory effort run short of several generalizations, because they try to express articulatory effort into one variable. The relevant constraint in such an approach would be:

Def. *EFFORT (*effort*)

“We are too lazy to spend any positive amount of *effort*.” (7.1)

The constraint-ranking version of minimization of effort would be stated as:

Minimization of effort:

“An articulation which requires more effort is disfavoured.” (7.2)

This would be formalized into a universally expected constraint ranking:

$$\text{*EFFORT}\ (x) \gg \text{*EFFORT}\ (y) \Leftrightarrow x > y \qquad (7.3)$$

where “>>” stands for “dominates”, “⇔” expresses logical equivalence, and “>” means “greater than”. However, articulatory effort depends on at least six primitives: energy, the presence of articulatory gestures, synchronization of gestures, precision, systemic effort, and coordination, and languages seem to be able to rank these separate measures individually to a certain extent. All of these will prove to be crucial in phonology.

7.1 Energy

A formula for the physiological effort needed by a muscle is at least as involved as

$$\int \left(ma + F_{el}\right) v \, dt + \int F_{el} v_0 \, dt \qquad (7.4)$$

where

t = time. Ceteris paribus, the longer the utterance, the more energy.

x = displacement of the muscle.

$v = dx/dt$ = the velocity of the moving muscle. For a constant force, the power spent is higher for higher velocity.

m = mass to move.

$a = d^2x/dt^2$ = the acceleration. The heavier the moving structures, the more energy is spent in accelerating them.

F_{el} = elastic forces and forces exerted by other muscles (gravitational forces can be included here). Stretching other muscles costs energy.

v_0 = some constant expressing the energy needed for an isometric contraction. Applying a force costs energy, even in the absence of motion.

Negative integrands should be ignored in (7.4), because no energy can be regained by the muscle. If summed over all muscles, formula (7.4) defines a global effort measure, analogously to the global contrast measure of (4.24).

The energy constraint against a positon change, i.e., a slow movement of an articulator from one position to the other, is associated with the *work* done by the muscle, i.e., the term $\int F_{el} v\, dt$ in (7.4). It can be expressed as:

Def. *DISTANCE (*articulator*: *a* | *b*)

> "An *articulator* does not move from location *a* to *b*, away from the neutral position." (7.5)

This constraint is not really a single constraint, but a ***family*** of constraints, parametrized by the articulator and the locations. Within the *DISTANCE family, a universal ranking is given by the following principle:

Minimization of covered distance:

> "An articulator moving away from the neutral position prefers to travel by the shortest distance possible." (7.6)

This is expressed in a constraint-ranking formula as:

$$\text{*DISTANCE } (\textit{articulator}: x_1 \mid x_2) >> \text{*DISTANCE } (\textit{articulator}: y_1 \mid y_2) \Leftrightarrow |x_1 - x_2| > |y_1 - y_2| \qquad (7.7)$$

where the "|" stands for an articulatory contour, i.e., a change in position or tension of the articulator. This is expected to hold within each articulator in every language.

The energy constraint against maintaining a non-neutral position of an articulator is associated with the energy spent in holding an isometric contraction, i.e., the term $\int F_{el} v_0\, dt$ in (7.4). It can be expressed as:

Def. *HOLD (*articulator*: *position*, *duration*)

> "An *articulator* stays at its neutral position, i.e., it is not held in any non-neutral *position* for any positive *duration*." (7.8)

The universal ranking of these constraints are given by the following principles:

Minimization of extension:

> "An articulator likes to stay as near as possible to the neutral position." (7.9)

Minimization of duration:

> "A non-neutral position should be maintained as briefly as possible." (7.10)

In formulas, where the position *x* is measured relative to the neutral position:

$$\text{*H}\textsc{old}\ (\textit{articulator}\text{: } x, \Delta t) >> \text{*H}\textsc{old}\ (\textit{articulator}\text{: } y, \Delta t) \Leftrightarrow |x| > |y| \quad (7.11)$$

$$\text{*H}\textsc{old}\ (\textit{articulator}\text{: } x, \Delta t) >> \text{*H}\textsc{old}\ (\textit{articulator}\text{: } x, \Delta u) \Leftrightarrow \Delta t > \Delta u \quad (7.12)$$

In a model for vowel inventories, Ten Bosch (1991) constrained the articulatory space with a boundary of equal effort, which he defined as the distance to the neutral (straight-tube, [ə]-like) position. In terms of the ranking (7.11), this would mean having all *HOLD constraints undominated above a certain displacement *x*, and all constraints maximally low for smaller displacements.

Finally, equation (7.4) contains the term $\int mav\,dt$, which expresses the fact that a displacement costs more energy if it has to be completed in a short time, at least if no energy is regained in the slowing down of the movement. The related constraint is:

Def. *FAST (*articulator*: *a* | *b*, *duration*)
“An *articulator* does not complete its displacement from *a* to *b* in any finite *duration*.” (7.13)

The universal ranking within this family is given by:

Minimization of speed:
“Faster gestures are disfavoured.” (7.14)

This can be formalized as

$$\text{*F}\textsc{ast}\ (\textit{articulator}\text{: } a \mid b, \Delta t) >> \text{*F}\textsc{ast}\ (\textit{articulator}\text{: } a \mid b, \Delta u) \Leftrightarrow \Delta t < \Delta u \quad (7.15)$$

The *DISTANCE, *HOLD, and *FAST constraint families associated with a certain articulator, can probably not be freely ranked with respect to one another, because there are no signs that the production system, let alone phonology, treats them individually. Rather, we could regard them as aspects of a general articulator-specific *ENERGY (*articulator*: *x*(*t*)) constraint, to whose ranking they contribute additively. This *ENERGY constraint is ranked by its energy value (7.4). The *ENERGY constraint clan is active in the case of phonetic implementation (ch. 10), but will be seen to show surprisingly little *organizational* power, especially seen in the light of the prominent role played by the principle of energy minimization in the literature on the phonetic simulation of sound inventories (for a discussion on this subject, see chapter 16).

7.2 Number of gestures

The ***number of articulatory contours*** on the gestural tiers is a first rough measure of the organizational effort of an utterance. The constraints that favour a reduction of the number of articulatory contours, express the qualitative difference between making and not making a gesture: the loss of a gesture implies a discrete organizational articulatory gain.

In this coarse measure, therefore, the *amount* of movement does not matter (by definition). Compare the simplest implementations of /apa/ and /awa/:

	a	p	a
lips	wide	closed	wide
pharynx	narrow		

	a	w	a
lips	wide	narrow	wide
pharynx	narrow		

(7.16)

Both contain two contours, so they are equally difficult in that respect.

The *number* of movements does matter. Compare /tɛnt/ with /tɛns/:

	ɛ	n	t
velum	closed	open	closed
blade	wide	closed	

	ɛ	n	s
velum	closed	open	closed
blade	wide	closed	crit

(7.17)

The utterance /tɛns/ ends with two contours, and is therefore more difficult organizationally than /tɛnt/.

The constraint family associated with the minimization of the number of contours can be called *GESTURE:

Def. *GESTURE (*gesture*)
"A *gesture* is not made." (7.18)

For instance, the constraint *GESTURE (blade: closure) can be held responsible for the deletion of the coronal gesture in Dutch /n+p/ sequences. Since *GESTURE has no continuous parameters, there is no universal ranking within this family. A universal *tendency* within the *GESTURE family, however, is expected to be

*GESTURE (*gesture*$_1$) >> *GESTURE (*gesture*$_2$) ⇔
⇔ effort (*gesture*$_1$) > effort (*gesture*$_2$) (7.19)

Such a ranking expresses an articulatory markedness relation across articulators. As with implicational markedness statements, these rankings can probably only be determined or predicted for "neigbouring" gestures. For instance, the larger rate of occurrence of coronal plosives with respect to labial plosives in most languages, may be attributed to the universal ranking *GESTURE (lips) >> *GESTURE (blade). However, the ranking of these constraints with respect to, say, *GESTURE (lowered velum) is not only difficult to determine; it is plausible that languages have a free choice in this ranking. For instance, there are a few languages without labial plosives, and a few other languages without nasal stops; this can be interpreted as the typology expected from a free ranking of *GESTURE (lips) with respect to *GESTURE (lowered velum).

Although (7.19) may express cross-linguistic and intralinguistic markedness relations, it is not valid in the realm of articulatory detail within a language. Rather, the finiteness of available articulatory tricks in every language forces us to admit that

*GESTURE (*gesture*) is undominated with probability 1 (7.20) where *gesture* spans the infinite number of thinkable articulations in the human speech apparatus. This effect is due to ***motor learning***: only those few gestures that the child has managed to master during the acquisition of her speech, are associated with a violable *GESTURE constraint. For instance, speakers of English apparently have a low *GESTURE (corono-alveolar closure) constraint, because they obviously know how to make alveolar plosives; the *GESTURE (corono-dental closure) constraint, on the other hand, is ranked high (or better: it is a virtual constraint not yet visible in the production grammar; see ch. 14). Speakers of French have the reverse ranking. Considerations of minimization of energy, therefore, seem not to be involved.

The emergence of motor skills is reflected in the reranking that takes place during speech development. Children start out with very few usably low-ranked *GESTURE constraints. While learning, the acquisition of coordinative skills causes the emergence of more low *GESTURE constraints, giving the *ENERGY constraints a chance to play a role.

Now that we have two constraint families, we can study an interaction. Below (1.15), I discussed the conflict between an active maintenance of lip spreading and the organizational problem of issuing a command to move the lips back to their rest position. In terms of tension control, the conflict is between *HOLD (risorius: 20% active, 100 ms) and *GESTURE (risorius: relax from 20% active); in terms of length control, the conflict is between *HOLD (risorius: 40% spread, 100 ms) and *GESTURE (risorius: from 40% spread to neutral); and in terms of the control of articulator position, the conflict is between *HOLD (lips: 40% spread, 100ms) and *GESTURE (lips: from 40% spread to neutral). The un-English implementation (1.15) would be the result of the ranking *GESTURE (relax lips) >> *HOLD (lips: spread, 100ms):

/tɛns/	*GESTURE (relax lips)	*HOLD (lips: spread)
☞ thɛɛ̃n̜t̜s̜		*
thɛɛ̃nts	*!	

(7.21)

It should be noted that a candidate without any lip spreading (i.e., satisfying *GESTURE (lips: spread)) is ruled out by the specification of maximum F_2.

Now that we have constraint interaction, we can predict a typology. Languages that have the ranking *GESTURE (relax lips) >> *HOLD (lips) are expected to maintain any non-neutral lip shape as long as possible, because that would minimize the number of articulatory contours, since there is always a chance that a following strong perceptual rounding specification requires the same lip shape. A typical phonologization of this effect would be the restriction of its domain to the morphological word: this would give a rightward rounding harmony, spreading from the strongly specified root onto weakly specified suffixes, like in many Turkic languages. Languages that have the ranking *HOLD (lips) >> *GESTURE (relax lips) will return to a neutral lip shape as soon as possible; their weakly specified suffixes typically contain central vowels, as in many Germanic languages. See further §19.1.

7.3 Synchronization

It is difficult to synchronize two articulatory contours exactly. All the intricate timing relations of adult speech have to be learned: the child starts out with simple single gestures (Koopmans - Van Beinum & Van der Stelt 1986, to appear; Vihman 1996: ch.5). If /tɛns/ is produced maximally faithfully as [[tʰɛns]] (the aspiration is considered part of the specification), we have a perfect synchronization of the nasal opening gesture with the dorsal closing gesture, and a synchronization of the nasal closing gesture with the dorsal opening gesture. This is depicted in the gestural score as the synchronization of the relevant contours:

Articulate:

velum	closed	open	closed
blade	wide	closed	crit

(7.22)

The resulting perceptual features and microscopic transcription are:

Perceive:

nasal		+	
coronal		+	
voiced	sonorant		
friction			sib
	ɛ	n	s

(7.23)

This output [[ɛns]] is perfectly faithful to the input. However, the required articulatory implementation apparently involves the violation of two contour-synchronization constraints:

Def. *SYNC (*articulator*$_1$: *from*$_1$ | *to*$_1$; *articulator*$_2$: *from*$_2$ | *to*$_2$[; Δ*t*])

“The movement of *articulator*$_1$ from *from*$_1$ to *to*$_1$ is not synchronous with the movement of *articulator*$_2$ from *from*$_2$ to *to*$_2$ [within any finite time span Δ*t*].” (7.24)

For a discrete version of *SYNC, the temporal distance parameter Δ*t* can be left out; it is then assumed to be “zero” for practical (perhaps perceptual) purposes. The universal ranking within the *SYNC family is given by:

Minimization of synchronization:

“Two articulatory contours on different gestural tiers like to be far apart.”

(7.25)

This can be formalized as

$$
\begin{gathered}
{}^*\text{S}\textsc{ync}\ (\textit{articulator}_1\text{: } \textit{from}_1 \mid \textit{to}_1\text{; } \textit{articulator}_2\text{: } \textit{from}_2 \mid \textit{to}_2\text{; } \Delta t) >> \\
>> {}^*\text{S}\textsc{ync}\ (\textit{articulator}_1\text{: } \textit{from}_1 \mid \textit{to}_1\text{; } \textit{articulator}_2\text{: } \textit{from}_2 \mid \textit{to}_2\text{; } \Delta u) \Leftrightarrow \\
\Leftrightarrow |\Delta t| < |\Delta u|
\end{gathered}
\tag{7.26}
$$

The two *SYNC constraints violated in [[ɛns]] would be:

*SYNC (velum: closed | open; apex: open | closed)
*SYNC (velum: open | closed; apex: closed | critical)

Both of these constraints can be satisfied by a different timing:

Articulate:

<table>
<tr><td>velum</td><td>closed</td><td colspan="2">open</td><td colspan="2">closed</td></tr>
<tr><td>blade</td><td colspan="2">wide</td><td colspan="2">closed</td><td>crit</td></tr>
</table>

Perceive:

<table>
<tr><td>nasal</td><td></td><td colspan="2">+</td><td colspan="2"></td></tr>
<tr><td>coronal</td><td colspan="2"></td><td>side</td><td></td><td>cont</td></tr>
<tr><td>voiced</td><td colspan="3">son</td><td colspan="2"></td></tr>
<tr><td>noise</td><td colspan="4"></td><td>sib</td></tr>
<tr><td></td><td>ɛ</td><td>ɛ̃</td><td>n</td><td>_ t</td><td>s</td></tr>
</table>

(7.27)

The resulting sound in that case is [[ɛɛ̃n_ts]]. Of course, this is different from the input /ɛns/ (it violates some *INSERT constraints, §9.9), but this is no reason to feel uncomfortable, because we have Optimality Theory to handle constraint interactions.

7.4 Precision

In his “quantal theory of speech production”, Stevens (1989) states that languages prefer those articulations whose acoustic result is not very sensitive to the accuracy of the articulation. For instance, an [i] is characterized by the proximity of its third and fourth formants; this closeness is preserved for a large range of tongue positions around the optimal palatal position. Thus, Stevens’ account can be translated into the principle of the minimization of the articulatory precision needed to reach a reproducible percept (Stevens 1990); this contrasts with Keating’s (1990) window model of coarticulation, which expresses the freedom of articulation as a window of solely articulatory specifications.

Another working of precision is the cross-linguistic predominance of plosives over fricatives. After all, it is easier to run into a wall than to stop one inch in front of it.[1] Thus, *controlled* movements, as found in fricatives and trills, involve more precision than *ballistic* movements, as found in stops (Hardcastle 1976).

The relevant constraint family can be written as

[1] This is not my metaphor, but I don’t remember whose it is.

Def. *PRECISION (*articulator*: *position* / *environment*)
“In a certain *environment*, a certain *articulator* does not work up the precision to put itself in a certain *position*.” (7.28)

The environment will often be something like *left* _ *right*, which stands for “between *left* and *right*”, where *left* and *right* are preceding and following articulatory specifications, often on the same tier. For instance, the constraint acting against the precision (constant equilibrium position of the lungs) needed to hold your breath between the inspiratory and expiratory phase is expressed as (when your upper respiratory pathways are open):

*PRECISION (lungs: hold / in _ out)

Quite probably, it is much more difficult to temporarily hold your breath during the course of an exhalation. This means that the constraint just mentioned is universally ranked below *PRECISION (lungs: hold / out _ out).

7.5 Coordination

There is no principled difference between assuming that the number of vowels in a language is finite, and assuming that vowel systems are structured within themselves, i.e. that they can be expressed in smaller units. Having a finite number of vowels means having a finite number of tricks, and there is no principled reason why these tricks could not be perceptual features and articulatory gestures, instead of whole segments. So: [e] and [o] form a natural class because of equal F_1 (perceptual feature), while [t] and [n] also form a natural class because of the use of the same tongue-tip gesture.

A first rough measure of the systemic effort of a language would be the number of articulatory and perceptual tricks needed to speak and understand that language, plus the number of *combinations* of these tricks that the language uses. E.g., if we find the sound change /k/ > /k^h/ in a language, chances are that *all* voiceless plosives get aspirated at the same time, since that would keep the number of trick combinations at a more manageable level: the trick combination “plosive + voiceless” is replaced by “plosive + aspiration”, whereas if the other voiceless plosives did not become aspirated, the language would end up with having the *two* combinations “plosive + voiceless” and “plosive + aspiration”. Alternatively, if the sound change /k/ > /k^h/ renders the sound system asymmetric, this principle may work later on in simplifying the now unbalanced system by causing the aspiration of /p/ and /t/, too.

The principle examined here is very important in building sound systems, and is usually called *maximum use of available features*, though, as we saw in our example, this term should be extended with: *and their combinations*.

Because every combination of articulatory tricks has to be learned, we have the following constraints:

Def. *COORD (*gesture*$_1$, *gesture*$_2$)
“The two gestures *gesture*$_1$ and *gesture*$_2$ are not coordinated.” (7.29)

As with *GESTURE, most of these constraints are undominated.

These negative relations between gestures are the common situation in speech development. Skilled speakers, on the other hand, have many *positive* relations between gestures, resulting from the acquired coordinations that implement the perceptual specifications of the utterances of the language.

For instance, Dutch has two perceptually contrasting degrees of voicing for plosives: fully voiced and fully voiceless. Both require an active laryngeal adjustment in their articulatory implementations. Now, a lax voiceless stop, as the English or South-German word-initial *b*, which requires no actions of the laryngeal muscles, can hardly be pronounced consciously by native speakers of Dutch; instead, it must be elicited by an extralinguistic experiment, for instance, the simulation of a repetitive mandibular gesture like the one found with babbling infants.

Another example is the extreme difficulty displayed by Dutch students when learning to produce unrounded back vowels: they typically produce either an unrounded front vowel modified with a backing gesture of the tongue body, or a rounded back vowel modified with a spreading gesture of the lips. No-one, by contrast, has any trouble producing the extralinguistic sound that expresses disgust, which combines voicing, lip spreading, and dorsal approximation. That sound, again, can hardly be produced without pulling the facial muscles that are associated with disgust but are superfluous for producing unrounded back vowels.

Thus, while plosives and rounded back vowels require complex coordinations not mastered by beginners, adults have several constraints that are the results of the plasticity of the human motor system:

Def. IMPLY ($gesture_1$, $gesture_2$) $\equiv \exists\, gesture_1 \Rightarrow \exists\, gesture_2$
“The presence of $gesture_1$ implies the presence of $gesture_2$.” (7.30)

This is an example of language-specific effort. Several muscles can only be pulled as a group (at least when speaking). These coordinations are language-specific and reflect the organizational shortcuts that characterize experienced speakers. The cross-linguistic pervasiveness of some of them have led some phonologists to ascribe to them the status of universal principles. For instance, numerous underspecificationists want us to believe that the implication [+back] → [+round] is a universal (innate) default rule, whereas, of course, the tendency for back vowels to be round is related to their maximal perceptual contrast with front vowels. If we stay by the functions of language, we can unequivocally assign the roles of cause and consequence.

Still, we have to ask to what extent (7.30) plays a role in the phonology of the language. It is quite probable that we have to invoke it for explaining the phenomena found in second-language acquisition: the trouble for speakers of English in producing unaspirated French plosives is not due to a perceptual failure or low faithfulness constraint, but must be attributed directly to the need to bypass a soft-wired (i.e., built-in but not innate) coordinative structure. Thus, the language-specific constraint (7.30) must play a role in articulatory implementation, i.e., the speaker uses it to her advantage in minimizing the number of higher neural commands, delegating some of the more automatic work to the more peripheral levels; in this way, [+back], with its automatic implication of [+round], is a simpler command than [+back; –round]. On the other hand, in explaining sound inventories, the combination [+back; +round] must be considered more complex than [+back; –round], because it involves one more active gesture (under the interpretation that [–round] involves no active lip spreading); the requirements of perceptual contrast then force the implementation of the more complex combination. From the functional standpoint, we would like to postpone the assumption of innateness until positive evidence arrives.

7.6 Global or local rankings of effort?

It is probable that the first steps of learning to move or speak are chiefly controlled by the principle of the minimzation of the number of gestures, and that later on, the development of coordination makes the minimization of energy a more important criterion. In general, however, it is hard to determine how to rank the various effort principles with respect to one another; not only for the linguist, but also, I would like to propose, for the speaker.

In discussing the relation between motor activity and effort in sports, it is impossible, for instance, to give a universal answer to the question whether skating or skiing is the more difficult of the two: it depends on the learning history of the person who performs these activities; but it is a universal fact that skiing becomes more difficult for very steep slopes, and that skating requires more effort on poor ice or if the rider is making a contest out of it.

Likewise, a speaker cannot assign numerical values to the various principles of effort, but she can locally rank different kinds of efforts within the separate families, along the lines of (7.7, 7.11, 7.12, 7.15, 7.26). The rankings across the families are determined by the learning history, i.e., by the language environment in which the speaker has grown up.

If languages differ as to what kinds of effort they consider important, a global measure of effort is not feasible. So I hypothesize that the holistic ranking (7.3) is not valid, and that only the rankings within the separate families are universal:

Local-ranking hypothesis for articulatory constraints:

> "A constraint cannot be ranked universally with respect to a constraint in a different family; and constraints within a family can only be ranked universally if only a single parameter is varied." (7.31)

Apart from being a negative condition on possible rankings, this is also a positive condition on the freedom assigned to every language: all ranking of constraints across families or of constraints with two different parameters, is free. An example of the single-parameter condition in (7.31) is: a language can freely rank its *HOLD constraints as long as the rankings (7.11) and (7.12) are honoured.

If this hypothesis is true, speech researchers will not have to try to assign numerical values to articulatory effort: we can get along with simple local rankings, and these can be predicted from known relations of monotonicity between effort on one side, and extension, duration, speed, number of contours, synchronization, precision, and coordination on the other.

7.7 Ranking by specificity

Another intrinsic ranking applies to the articulatory constraints. The gesture [bilabial closure] is, on the average, more difficult to make than the gesture [labial closure], because the underspecification of the latter would allow a labiodental implementation if the phonotactics of the situation favoured that:

Minimization of specificity of articulatory constraints:

> "For articulatory constraints, more specific constraints are ranked above less specific constraints." (7.32)

It can be formalized as

$$(\text{A} \Rightarrow \text{B}) \Rightarrow {}^*\text{G}\textsc{esture}\ (\text{A}) \gg {}^*\text{G}\textsc{esture}\ (\text{B}) \qquad (7.33)$$

Ranking (7.33) can be used as a universal ranking condition for *PRECISION constraints: the larger the window, the lower its *PRECISION constraint.

Ranking (7.33) is the reverse of an analogous ranking for perceptual constraints (see §9.10).

7.8 A restriction on functional rankings of articulatory constraints

Articulatory constraints cannot be ranked by considerations of perceptual importance. For instance, an alleged ranking *GESTURE (labial / stem) >> *GESTURE (labial / affix) or *GESTURE (labial / –stress) >> *GESTURE (labial / +stress), where the "/" means "in the domain of", would confuse articulatory constraints with faithfulness constraints: the ranking of *GESTURE (labial) can only depend on its articulatory environment. In chapters 10 and 11, I will show that asymmetries between the surfacing of gestures in environments of varying degrees of perceptual importance, arise from dependencies in the rankings of faithfulness constraints.

7.9 A comprehensive gestural constraint and additive ranking

The Optimality-theoretic device of ***strict ranking*** works most strongly if constraints are simple and separate. For instance, a geminate fricative violates both, say, *HOLD (long) and *PRECISION (high). Now, one of these must be ranked higher than the other. If the higher-ranked constraint is *HOLD (long), the facts of the geminate system are mainly explained by the ranking of this single constraint. In some cases, however, we might prefer to express the idea that a geminate fricative is so articulatorily demanding because it is a geminate *and* a fricative, and we would do so by means of a single high-ranked composite constraint like *GESTURE (long duration & high precision).

So if we allow such ***additive ranking*** of the various articulatory-effort factors, we can combine the *ENERGY, *GESTURE, and *PRECISION constraints into a single family:

Def. *GESTURE (*art*: *gesture* / *distance*, *duration*, *speed*, *precision* / *env*)

> "A *gesture* (or combination of gestures) is not performed along a certain *distance,* with a certain *speed,* held for a certain *duration*, with a certain *precision*, in a certain *environment.*" (7.34)

The ranking of this generalized constraint is conditioned by the four parameters: it is higher if the distance, duration (of holding), speed, or precision is greater, and everything else stays equal. We will see this multidimensional constraint family many times in later chapters. Instead of as a family, though, it could be interpreted as a single constraint whose ranking depends on the four parameters.

7.10 Conclusion

Organizational constraints like *GESTURE and *COORD and phonotactic constraints like *SYNC can be thought of as motivated by the principle of minimization of articulatory effort. These constraints are violable and can therefore be stated in general terms, so that they can be thought to be language-independent and phonetically motivated. Their rankings with respect to heterogenous constraints must be language-specific.

8 Perceptual categorization and the emergence of finiteness

Abstract. This chapter formalizes the functional principles that play their roles in the perception grammar. Every language uses a finite number of phonological feature values, because speakers learn a finite number of perceptual categories and articulatory gestures.

The most salient aspect of sound inventories is their finite size: each language uses a finite number of underlying lexical phonological segments or feature values. The functional explanation for this fact contains two sides: the finiteness of the number of articulatory features, and the finiteness of the number of perceptual features.

Prince & Smolensky (1993) maintain that any theory of phonology can only be called 'serious' if it is "committed to Universal Grammar" (p. 1). The learning algorithm of Tesar & Smolensky (1995) explicitly assumes "innate knowledge of the universal constraints" (p. 1). They also have to assume that there are a finite number of constraints. However, we have seen for articulatory constraints (ch. 7), as we will see for perceptually motivated constraints (ch. 9), that there are an infinite number of them. In this section, I will show that, though the constraints themselves are universal, separate languages warp the continuous articulatory and perceptual spaces in such a way that each language ends up with a unique set of allowed gestures and specificational elements (features): the articulatory space is warped by motor learning, which lowers a few articulatory constraints, and the perceptual space is warped by categorization, which lowers some constraints of speech perception.

8.1 Feature values are not innate

If we talk about certain linguistic phenomena as being 'universal', we can mean either of two things: first, in the sense of Universal Grammar, that these phenomena exemplify ***innate*** properties of the human language faculty; secondly, that languages tend to have these phenomena because the functions of communication are similar in most languages, and because our speech-production organs and our ears are built in similar ways. Though these two views need not be conflicting as they stand, I will take the stronger functional position: that humans are capable of learning to speak without the necessity of innate phonological feature values, i.e., that languages can make their own choices from the perceptual and articulatory possibilities identified in §1.2.

As we see from the success of sign languages for the deaf (Brentari 1995), a phonology can be based on the capabilities of any motor system (talking, signing) and any sensory system (audition, vision) considered suitable for expressing intentions, wishes, and thoughts. We must conclude that nature did not force any specific motor system upon us for communication. This supports the view that we are not confined to using a universally fixed set of features if we choose to use the speech apparatus for our communication.

As an example, consider the division of the vowel height continuum. All too often, vowels are put into categories on the basis of a dogmatic “principle” that states that all languages use the same feature set (Kenstowicz 1994, Clements & Hume 1995). The International Phonetic Alphabet, for instance, seems to have been developed for languages with four vowel heights, having [ɛ] and [e] to represent front unrounded mid vowels. However, in most languages with three vowel heights (e.g., Spanish, Russian, Japanese), the height of this vowel is in between [ɛ] and [e] (Akamatsu 1997). This means that vowels are distributed along the height dimension in a way that enhances the perceptual contrast between them, and not according to a universal set of binary features, not even, I would like to conjecture, “underlyingly”.

The illusion of a universal set of features probably originated in the fact that the speech systems of most humans are very much alike, so that many languages do use the same features. Generalizing this to assuming a universal innate set of features is unwarranted.

Though there is no such thing as cross-linguistic sameness, much work in contemporary phonology is done to find the allegedly universal features, and put them into larger classes and hierarchies (manner versus place features, or major class features versus the rest). For instance (emphasis added):

> “*since* features are universal, feature theory *explains* the fact that all languages draw on a similar, small set of speech properties in constructing their phonological systems. *Since* features are typically binary or one-valued, it also *explains* the fact that speech sounds are perceived and stored in memory in a predominantly categorial fashion.” (Clements & Hume 1995, p. 245)

My position on this subject is that the causal relationships in these assertions should be reversed: because of the content of the constraints on human speech production and perception, different languages may sometimes show up with similar feature sets, and the functional interpretation of categorization predicts into how many values a perceptual feature space can be divided (§14.2.5). An analysis of the emergence of language-specific features from an infinite universal pool of possible articulations and perceptual categories, is advanced in the remaining part of this chapter.

8.2 Constraints in speech production

Most articulatory gestures have to be learned. Before this is accomplished, all *GESTURE constraints are ranked quite high, but once a gesture has been learned because it occurs in a mastered word, the relevant *GESTURE constraint must have descended below the relevant faithfulness constraint. But this will facilitate the surfacing of the gesture in other words, too. For instance, a language with a click consonant will probably have more than one click consonant, because some of the coordinations required for those other clicks have been mastered already for the first consonant. Likewise, speakers of a language with corono-dentals stops will have trouble with the corono-alveolar stops of other languages, and vice versa; there is no universal preference for either of these implementations of coronal stops.

Thus, in the end, though most *GESTURE constraints are still undominated (see (7.20)), some of them are so low as to allow the gestures to be made. This means that gestures and coordinations are the articulatory building blocks of sound inventories:

Articulatory inventory constraints:

> "Low-ranked *GESTURE and *COORD constraints determine the finite set of allowed articulatory features and feature combinations." (8.1)

This explains not only the finiteness of the segment inventory, but also (partly) the symmetries that we find inside inventories.

8.3 Functional constraints in speech perception: categorization

Because of the overwhelming variation in the world they live in, human beings organize their view of the world with the help of *categories*. Besides reducing cognitive load, categorization leads to fewer mistakes in identifying groups of things that we had better treat in the same way.

Like the production, the perception of speech has to be learned, too. The process of speech recognition entails that an acoustic representation is ultimately mapped to an underlying lexical form. A part of this process is the categorization of the acoustic input (fig. 1.1). This section will describe the relation between the acoustic input and the perceptual result in terms of the ***faithfulness and categorizarion constraints of speech perception.*** They are part of the ***perception grammar*** (fig. 6.1).

First, it is desirable that an acoustic feature is recognized at all by the listener. The following constraint requires a corresponding perceived feature value for every acoustic feature value (the subscript *i* denotes correspondence):

Def. PERCEIVE (f) $\equiv \exists x_i \in f_{ac} \Rightarrow \exists y_i \in f_{perc}$

> "A value x on a tier f in the acoustic input is recognized as any corresponding value y on the same tier." (8.2)

As always in Optimality Theory, the constraint has to be interpreted as gradiently violable: each unrecognized feature incurs one violation mark; this differs from the purely logical interpretation of "$\exists x_i \Rightarrow \exists y_i$" (if there is an x, there must also be a corresponding y) or its equivalent alternative "$\forall x_i \exists y_i$" (for every x, there must be a corresponding y).

An analogous constraint DONTPERCEIVE requires that a recognized feature should have a correspondent in the acoustic input: it militates against perceiving features

Secondly, it is undesirable that an acoustic feature value is recognized as something which is normally associated with a very *different* acoustic feature value. For instance, a vowel with a F_1 of 600 Hz is most properly perceived as a lower mid vowel, and a recognition as a high vowel is disfavoured. The following faithfulness constraint militates against distortions in perception (the asterisk can be read as "don't"):

Def. *WARP (f: d) $\equiv \exists x_i \in f_{ac} \wedge \exists y_i \in f_{perc} \Rightarrow \left| x_i - y_i \right| < d$

"The perceived value y of a feature f is not different from the acoustic value x of that feature by any positive amount of distortion d." (8.3)

Note that if a feature is not perceived, *WARP is not violated because the acoustic input feature has no correspondent: it is then ***vacuously satisfied***. In other words, this constraint can be subject to ***satisfaction by deletion***, which is also suggested by its negative formulation.

Because it is worse to perceive [ɛ] as /i/ than it is to perceive [ɛ] as /e/ (as will be proved in §9.2), *WARP has the following universal internal ranking:

Minimization of distortion:

"A less distorted recognition is preferred over a more distorted recognition." (8.4)

This can be formalized as

$$\text{*WARP } (\textit{feature}: d_1) \gg \text{*WARP } (\textit{feature}: d_2) \Leftrightarrow d_1 > d_2 \qquad (8.5)$$

Together, (8.3) and (8.5) assert that if a higher *WARP constraint is violated, all lower *WARP constraints are also violated.

Besides the above faithfulness constraints, and analogously to the *GESTURE family (7.18), which is an inviolable constraint for most of the universally possible gestures, we have a family of constraints that express the learnability of categorization:

Def. *CATEG (f: v) $\equiv \exists x_i \in f_{perc} \Rightarrow x_i \neq v$

"The value v is not a category of feature f, i.e., a perceptual feature f cannot be recognized as the value v." (8.6)

Analogously to the situation with *GESTURE, as stated in (7.20), we have

*CATEG (*feature*: *value*) is undominated with probability 1 (8.7)

where *value* spans the whole range of values that *feature* can attain along its continuous auditory dimension. This expresses the finiteness of available perceptual categories within a language: *CATEG is high-ranked for almost all values, and low-ranked only for a small number of discrete values, which correspond to the centres of the language-specific categories.

The interaction of the *CATEG, PERCEIVE, and *WARP constraints in recognition is the subject of the following section.

8.4 Categorization along a single perceptual dimension

As an example, we will look at the interaction of the constraints for the recognition of an auditory feature *f* that can have any value between 0 and 1000 along a continuous scale: the first formant, with a scale in Hz[1]. If PERCEIVE is undominated (i.e., every acoustic input will be categorized), and *WARP is ranked internally in the universal way, and *CATEG is ranked high except for the values *f* = 260, *f* = 470, and *f* = 740, then a partial hierarchy may look like (the parameter *f* is suppressed from now on):

PERCEIVE
*WARP (400)
*WARP (300)
*CATEG (280), *CATEG (510), *CATEG (600) etc. etc. etc.
*WARP (240)
*WARP (140)
*WARP (100)
*CATEG (260)
*CATEG (740), *CATEG (470)
*WARP (50)
*WARP (20) (8.8)

Note that all the *WARP constraints not mentioned here do belong somewhere in this ranking, according to (8.5), and that all the *CATEG constraints not mentioned in (8.8) take fourth place in ranking, together with *CATEG (280). We will now see how this constraint system controls the recognition of any input value f_{ac} between 0 and 1000.

First, consider the input [260], which is a phonetic realization of *f* with a value of 260 (e.g., a vowel pronounced with a first formant of 260 Hz). We see that this auditory input is recognized as /260/ (in this tableau, the constraints have been abbreviated):

[260]	PERC	*W(400)	*C(280)
			*C(510)
			*C(590)
				*W(240)	*W(100)	*C(260)	*C(470)
							*C(740)
								*W(30)

☞ /260/						*		
/470/					*!		*	*
/740/		*!		*	*		*	*
nothing	*!							

(8.9)

[1] For (8.5) to be valid, we should use the perceptually calibrated Bark scale instead, but since the current case is meant as an example only, we use the more familiar physical frequency scale.

The candidates /470/ and /740/, though chosen in (8.8) to be stronger categories than /260/, lose because honouring them would violate some stronger *Warp constraints. The winning candidate violates only the *Categ(260) constraint, which cannot be helped: satisfying all *Categ and *Warp constraints would require violating Perceive.

The case of an input that is quite close to one of the preferred categories, yields an analogous result, as shown in the following tableau for the realization [510], which will be recognized as /470/:

[510]	Perc	*W(400)	*C(280)
			*C(510)
			*C(590)
				*W(240)	*W(100)	*C(260)	*C(470)
							*C(740)
								*W(30)

/260/				*!	*	*		*
☞ /470/							*	*
/510/			*!					
/740/					*!		*	*
nothing	*!							

(8.10)

In this case, we must consider the candidate /510/, which satisfies all *Warp constraints, but violates the strong *Categ(510) constraint. Thus, because it is worse to map the input into the non-existing category /510/ than to distort the input by 40 Hz, the input [510] maps to the output /470/.

Another case is the recognition of an input that is not close to any of the good categories. The following tableau shows the recognition of [600]:

Input: [600]	Perc	*W(400)	*C(280)
			*C(510)
			*C(600)
				*W(134)	*W(100)	*C(260)	*C(470)
							*C(740)
								*W(30)

/260/				*!	*	*		*
☞ /470/					*		*	*
/600/			*!					
/740/				*!	*		*	*
nothing	*!							

(8.11)

The output candidate /470/, being 130 Hz off from the input, violates *Warp (129) but not *Warp (131). Thus, it is slightly better than the candidate /740/, which violates *Warp (139). So we see that stray inputs like [600] are put into the "nearest" category.

Generalizing from these three examples, we can draw a picture of the recognition of all possible inputs between [0] and [1000]. Figure 8.1 shows the relevant Perceive and

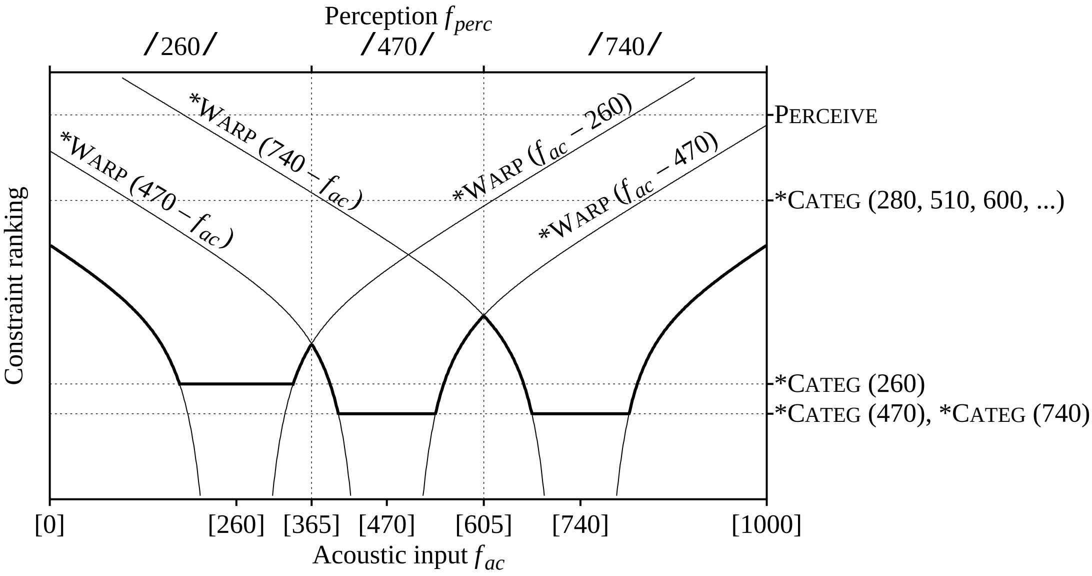


Fig. 8.1 Categorization of the input along a continuous auditory parameter. The curves represent the heights of the *Warp constraints in the cases that the auditory input is recognized as /260/, /470/, or /740/. The thick curve represents the height of the highest violated constraint if the categorization divides the domain into the three parts shown at the top.

*Categ constraints as horizontal dotted lines, and the three *Warp constraints *Warp $\left(\left|f_{ac}-260\right|\right)$, *Warp $\left(\left|f_{ac}-470\right|\right)$, and *Warp $\left(\left|f_{ac}-740\right|\right)$ as functions of the auditory input parameter f_{ac}.

The picture shows that Perceive is ranked as high as *Warp (550): the curve *Warp (740 – f_{ac}) crosses the Perceive line at f_{ac} = 190; also, *Categ (280 etc.) are as high as *Warp (350): the same curve crosses that *Categ line at f_{ac} = 390. Two *criteria* (category boundaries) emerge exactly half-way between the categories, at 365 and 605. Note that though /260/ is a weaker category than /470/ (its *Categ constraint is higher), the location of the boundary between the /260/ and /470/ equivalence classes is not influenced by this height difference: the height of the horizontal thick line above ‘260’ in the figure does not influence the location of the cutting point of the two *Warp curves at [365], unless this line would actually be higher than the cutting point. This is an example of *strict ranking*: the two struggling *Warp constraints determine the outcome, without being influenced by any lower-ranked third constraint (§9.5 will show that the height of *Categ correlates with the width of the *Warp curve, so that the criterion does shift).

In a more realistic model of speech recognition, the thick curve in figure 8.1 does not represent the ultimately recognized category. In the phase of recognition proper (seen here as occurring “after” categorization), which involves lexical access and information on context and syntax, we must assign a probabilistic interpretation to the curve (§9.2, §9.5): it only shows the *best* candidate, i.e., the candidate with highest probability of being correct; other, lower-ranked, candidates have lower probabilities, and a global optimization algorithm will find the best time path through the candidates.

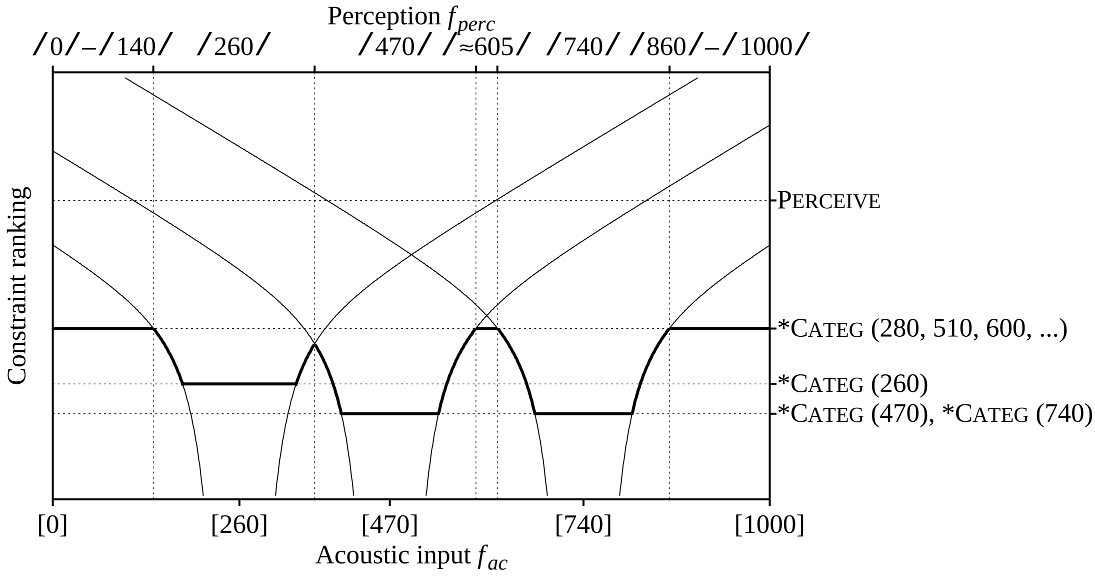


Fig. 8.2 Categorization along a one-dimensional continuum, if the *CATEG constraints for the weak categories are ranked rather low.

8.5 Special case: weak categories

If the *CATEG constraints of the weak categories are ranked low enough, they can interact with *WARP constraints. In this case, highly distorted categorizations will not take place. Instead, inputs that are far away from the centre of the equivalence class of a strong category, will be recognized into one of the weak categories:

[600]	PERC	*W(400)	*W(125)	*C(280)
				*C(510)
				*C(600)
					*W(100)	*C(260)	*C(470)
							*C(740)
								*W(30)

/260/			*!		*	*		*
/470/			*!		*		*	*
☞ /600/				*				
/740/			*!		*		*	*
nothing	*!							

(8.12)

Figure 8.2 shows the classification of any input between [0] and [1000] in the case of low weak-category constraints.

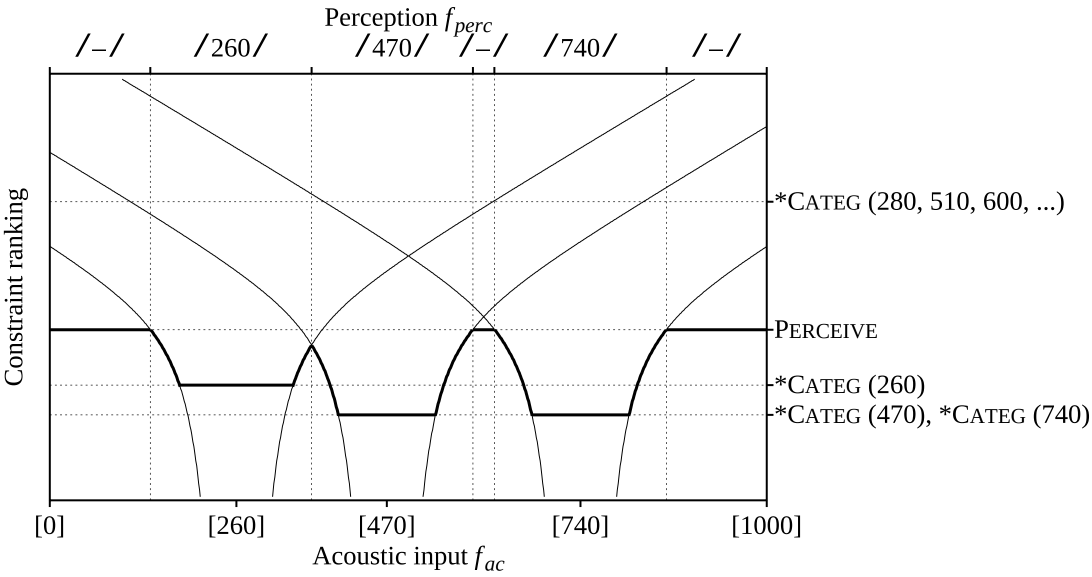


Fig. 8.3 Categorization along a one-dimensional continuum, if the PERCEIVE constraint is ranked low. Non-recognition is denoted as “/-/”.

8.6 Special case: unparsed features

If the PERCEIVE constraint is ranked low, it is allowed to interact with the *WARP constraints. In this case, highly distorted categorizations will not take place; instead, inputs that are far away from the centre of the equivalence class will not be recognized (“/-/” stands for “not recognized”):

[590]	*W(400)	*C(280)
		*C(510)
		*C(590)
			*W(110)	PERC	*W(100)	*C(260)	*C(470)
							*C(740)
								*W(30)

/260/			*!		*	*		*
/470/			*!		*		*	*
/590/		*!						
/740/			*!		*		*	*
☞ /-/				*				

(8.13)

Figure 8.3 shows the classification of any input between [0] and [1000] in the case of a low PERCEIVE constraint.

8.7 Dependence on environment

The ranking of the constraints of speech perception depends on several external and internal phenomena:

- A higher frequency of occurrence of a certain category in the vocabulary of a language means that that category is recognized more often, and, therefore, that categorization into this category is easier. Thus, frequently visited categories have low *CATEG constraints. This is formalized and proved in §9.5.
- A higher frequency of occurrence also lowers the distinctive power of a feature value and, with it, the height of the PERCEIVE constraint for this feature.
- The presence of background noise, too, reduces the importance of the classification of the individual features; thus, it lowers the ranking of PERCEIVE.
- More variation in the acoustics of a feature value gives more latitude in the admission to the corresponding category, and this leads to relatively low *WARP constraints for high distortions ("wide" *WARP functions).

8.8 Merger

We can now predict what happens when two categories come to overlap. The source of the overlap is usually an increase in the variation in the production, often caused by the merger of a migrating group of people with another population that speaks a related but slighly different dialect.

Because of the large variation, the *WARP functions will be wider, as shown in figure 8.4. The more common (stronger) category (550) will have the lower *CATEG constraint; figure 8.4 shows us that this will lead to a shift of the criterion in the direction of the weaker category (to "442"). As every input greater than 442 will be classified as belonging to the stronger category, this criterion shift will again increase the rate of recognition into the stronger category, and decrease the rate of recognition into the weaker category. As a result of this, the *CATEG constraint of the stronger category will become lower, and that of the weaker category will become higher. This will cause a further criterion shift. Apparently, the larger class is eating away at its peer, and this positive-feedback mechanism will ultimately send the weaker class into oblivion (unless checked by the requirements of information content, see §9.6): an irreversible process of lexical diffusion ends up as a blind law of sound change. The resulting merger of the categories may well result at first in an asymmetry between production and perception: the speaker may still know that she produces a contrast, but the listener may be indifferent to it, because not considering the information found in a poorly reproducible contrast may decrease the error rate of the recognition. In chapter 17, we will see some evidence for the idea that in the perdition of phonological categories, the development of production lags behind the development of perception.

The problem in figure 8.4 can also be solved by the weaker category moving away from the encroaching stronger one (push chain).

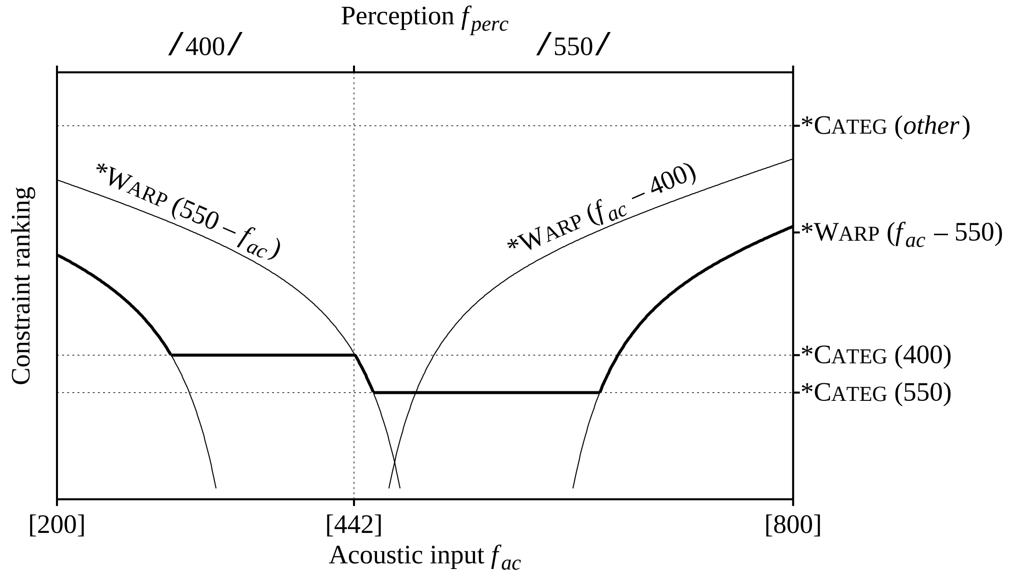


Fig. 8.4 The recognition into two overlapping categories of unequal strength.

8.9 Conclusion

The finiteness of sound inventories is explained by the articulatory inventory constraints (§8.2) and their perceptual counterpart:

Perceptual inventory constraints:

> "Low-ranked *CATEG constraints determine the finite set of allowed perceptual feature values." (8.14)

The term "features" here is used in a broad sense: it may refer to values on a continuous auditory scale (e.g., F_1 or F_2), or to combinations of those (e.g., a location in a vowel triangle). Functionally, there is no reason why features should be one-dimensional; some latitude in the dimensionality of primitive perceptual spaces would explain why besides languages with highly symmetric vowel systems, we also find languages with asymmetric vowel systems; in the former case the language has several vowel-height and vowel-place categories, in the latter case it has vowel-quality categories. It is only natural that the languages with two distinct dimensions of categorization have more vowels than those with direct categorization of the two-dimensional quality space.

We can also draw an important conclusion from our functional standpoint: though all constraint families may be universal, the features that build inventories are language-specific. For instance, all languages have the same constraints against the categorization of all vowel heights, i.e., they all have *CATEG (F_1: x) for all possible values of x. In every language, almost all of these constraints are undominated (see (8.7)). But though all languages have the *CATEG (F_1: 320 Hz) and *CATEG (F_1: 620 Hz) constraints, only a language with two categorizable vowel heights has them at a low rank, so that this language shows vowel heights at 320 Hz ("i") and 620 Hz ("a"). Its sister language, with three vowel heights, has the same constraints, but has three different *CATEG constraints at a low rank, giving recognizable heights at 260 Hz (“i”), 470 Hz (“e”), and 740 Hz (“a”). Finally, a typical language with four vowel heights will have them around 240 Hz (“i”), 380 Hz (“e”), 560 Hz (“ɛ”), and 780 Hz (“a”). The interaction of *ENERGY and faithfulness constraints dictates the dependence of the peripheral heights (“a” and “i”) on the number of vowel heights (see §10.8), and the interaction of *WARP constraints determines the positions of the categories. The use of the label “a” with all three languages should not mean that we pose a universal category /a/, and the label “e” (which is especially arbitrary for the three-height language) does not mean the same for languages with three and four vowel heights: there is no universal vowel /e/. Thus, from the universal *CATEG family emerges a language-specific division of the vowel-height dimension, which is, moreover, partially determined by the functional principle of maximal minimal contrast. This leads to an important conclusion:

The functional view: no universal phonological feature values

> “The continuous articulatory and perceptual phonetic spaces are universal, and so are the constraints that are defined on them; the discrete phonological feature values, however, are language-specific, and follow from the selective constraint lowering that is characteristic of the acquisition of coordination and categorization.” (8.15)

9 Perceptual contrast and faithfulness

Abstract. This chapter formalizes the functional principle of minimization of perceptual confusion into faithfulness onstraints and their univrsal local rankings.

As with the maximization of articulatory ease, Trask (1996) calls the principle of *maximum perceptual separation* “a somewhat ill-defined principle sometimes invoked to account for phonological change”. But, again, we will see that it can be expressed in a linguistically meaningful way.

9.1 How to implement minimization of confusion

We will first compare a few strategies for implementing the functional principle of minimization of confusion, which is the more fundamental principle that often underlies the maximization of perceptual contrast.

9.1.1 Global optimization

A ***global*** utilitarian optimization strategy would minimize the total number of confusions that would occur in a long series of utterances. A global egalitarian optimization strategy, by contrast, would minimize the maximum confusion probability. The latter option is more in line with the idea behind Optimality Theory, where the highest-ranked constraint, i.e., the constraint against the largest problem, outranks all others. Interestingly, Ten Bosch (1991) showed that in a model of vowel inventories, the optimization strategy of maximizing the minimum distance between pairs of vowels, performed better than maximizing a global contrast measure along the lines of Liljencrants & Lindblom (1972) or Vallée (1994). An output-oriented contrast constraint would be

Def. *CONFUSION (*confusion*)

> “We are too petty to allow any positive amount of ***confusion***.”

The constraint-ranking version of minimization of confusion would then read:

Minimization of confusion:

> “Within the set of all pairs of utterances with distinctive meanings, the pairs with higher confusion probabilities are disfavoured.”

This rote functionalism is obviously not supported by the facts. It would predict, for instance, that sound changes would change only those words that are most easily confused with others, or that otherwise homogeneous sound changes would have exceptions where they would create homonyms. Such phenomena are very unusual, especially for gradual processes such as vowel shifts. This is explained by the facts of categorization: if categories are important, they move as a whole, dragging along all the words in which they occur. If the movement is gradual, there is no way for isolated lexical items to stay behind; only for sound changes that involve category jumps, like processes of lexical diffusion, could it be functionally advantageous not to jump if that would increase homonymy.

9.1.2 The local-ranking principle

In §7.6, I argued for restricting the measurability of the ranking of articulatory effort to minimally different pairs of situations. The same holds for perceptual contrast.

In discussing similarity, it is impossible to give a universal answer to the question which pair is more alike: a horse and a cow, or an apple and a peach. But most people would agree that a horse is more similar to a cow than it is to a duck, and that an apple is closer to a pear than to a peach. Likewise, the listener cannot assign numerical values to the various degrees of contrast, but she can rank locally different contrasts. Thus, the main thing we will have to know about contrasts is the monotonicity of the relation between distance and contrast: the higher the distance between two sounds along a single acoustic/perceptual scale, the lower their probability of confusion.

9.1.3 Local implementation: faithfulness

Instead of by the global method of §9.1.1, which is also problematic because of its required processing load, the functional principle that utterances with different meanings should be sufficiently different, can be implemented non-globally by a pair of requirements: the underlying forms should be sufficiently different, and every underlying form (***specification***) is close to the corresponding surface form (***perceptual result***).

Each candidate articulation in the specification-articulation-perception triad (§1.3.3) may produce a different perceptual result. The differences between the input specification and the perceptual output are caused by articulatory constraints, which tend to decrease the perceptual contrast between utterances. For instance, if the constraint against the laryngeal gestures that implement the voicing contrast for obstruents is ranked high, underlying /ba/ and /pa/ will fall together; and honouring the constraint against the synchronization of the velar and coronal gestures in /tɛns/ 'tense' will make it sound like the output of /tɛnts/ 'tents'. Thus, the principle of maximization of perceptual contrast can be translated:

- indirectly: into families of faithfulness constraints that state that aspects of the specification should appear unaltered in the output;
- directly: into the contrast-dependent rankings of these constraints.

A global formulation would be:

Def. FAITH (*d*)

"The perceptual output should not be different from the specification by any positive difference *d*." (9.1)

The constraint-ranking version of maximization of contrast would then be stated as:

Maximization of faithfulness:

"A less faithful perceptual output is disfavoured." (9.2)

This would be formalized into a universally expected constraint ranking:

$$\text{FAITH}\ (d_1) >> \text{FAITH}\ (d_2) \Leftrightarrow d_1 > d_2 \qquad (9.3)$$

Just as with the constraints of articulatory effort, the faithfulness constraints branch into several families, which cannot be universally ranked with respect to each other along the lines of (9.3), which uses a global measure of contrast like equation (4.24). The various aspects of the underlying specification will be identified in the following sections.

9.1.4 Faithfulness in phonetic implementation

The first thing that is apparent from the specification (1.14) is the presence of features. For instance, the morpheme /tɛns/ contains specifications for [coronal], [+nasal], and [lower mid]. Because the speaker will try to accomodate the listener, it is desirable that the acoustic output contains something (anything) that corresponds to them. Analogously to the PERCEIVE constraint of perception, the speaker would adhere to the following imperative of correspondence:

Def. PRODUCE (f) $\equiv \exists x_i \in f_{spec} \Rightarrow \exists y_i \in f_{ac}$

"A value x on a tier f in the specification has *some* corresponding value y on the same tier in the acoustic output." (9.4)

An analogous constraint DONTPRODUCE, which can be formalized by reversing the implication in the definition of PRODUCE, requires that anything in the acoustic output has a correspondent in the specification (cf. DONTPERCEIVE in §8.3).

Mostly, the speaker is also intent on maximizing the probability of correct recognition of her utterance. So, analogously to *WARP, we would have a constraint that penalizes the variation of production, as far as this leads to deviant acoustic results:

Def. *VARY (f: d) $\equiv \exists x_i \in f_{spec} \wedge \exists y_i \in f_{ac} \Rightarrow \left| x_i - y_i \right| \le d$

"The produced value y of a perceptual feature f is not different from the specified value x by any positive amount of variation d." (9.5)

The wording of this constraint is deliberately symmetric between input and output. Like *WARP, *VARY is satisfied vacuously if the underlying feature has no correspondent in the acoustic signal: this may occur in situations where it is better not to produce a feature than to produce the wrong value. The universal ranking within the *VARY family is

Minimization of variation:

"A less deviant production is preferred over a more deviant production". (9.6)

This can be formalized as

$$\text{*VARY}\ (\textit{feature}: d_1) >> \text{*VARY}\ (\textit{feature}: d_2) \Leftrightarrow d_1 > d_2 \qquad (9.7)$$

The picture presented here of the listener is that she will hear every acoustic output as it is. As we have seen, however, the effects of categorization discretize the perceptual output and, therefore, the perceptual specification. Discretized versions of PRODUCE and *VARY will be presented below.

9.2 Faithfulness in phonology

The listener will not rank the acoustically realized feature values directly along continuous scales. Rather, she will categorize the acoustic input into perceptual feature values along one-dimensional scales ("before" recognition of the utterance). The standpoint of Functional Phonology, inspired by the presence of an auditory-feedback loop (§1.2.1, figure 1.1), is that the version of faithfulness that plays a role in the organization of spoken language, evaluates the difference between the perceptual specification and the perceptual features *as categorized* by the listener.

We can view the medium of information transfer between speaker and listener as a system of parallel communication channels, each of which represents one perceptual tier. Each tier tries to transmit serially events associated with a particular perceptual feature. The presence of a message on each tier is transmitted successfully if the PRODUCE and PERCEIVE constraints are both satisfied (also in the unlikely case that PRODUCE and DONTPERCEIVE are both violated):

Def. TRANSMIT $(f \,/\, x) \equiv \exists x_i \in f_{spec} \Rightarrow \exists y_i \in f_{perc}$

"The value (category) x on a tier f in the specification corresponds to any category y on the same tier in the perceptual output." (9.8)

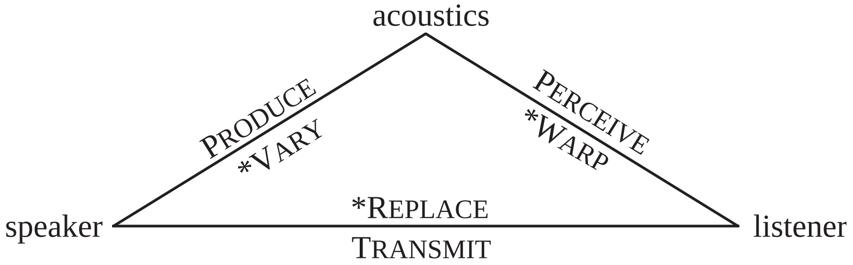
And, again, we have DONTTRANSMIT, which is satisfied if both DONTPRODUCE and DONTPERCEIVE are satisfied (or if DONTPRODUCE and PERCEIVE are both violated).

Analogously to *WARP and *VARY, we have a constraint that penalizes the difference between the specified and the perceived feature value:

Def. *REPLACE $(f\colon x, y) \equiv \exists x_i \in f_{spec} \wedge \exists y_i \in f_{perc} \Rightarrow \left| x_i - y_i \right| \le d$

"The perceived category y on a tier f is not different from the specified value x by any positive distance d." (9.9)

Thus, the effect of TRANSMIT is the product of the effects of PRODUCE and PERCEIVE, and the effect of *REPLACE constraint is the convolution of the effects of *VARY and *WARP. The communication process can thus be summarized as



(9.10)

The phonology handles TRANSMIT and *REPLACE constraints, because language users are speakers and listeners at the same time, and do not know about the acoustic medium.

In contrast with *VARY, which worked, by definition, along a perceptually homogeneous scale, *REPLACE has to be parametrized with the feature values *x* and *y*, because its ranking depends on the distances and strengths of the categories, as will be seen below and in §9.5. The universal ranking within the *REPLACE family, based on the principle that the listener will compensate for near categorical errors more easily than for distant errors (by adapting the recognition probabilities, see §9.5), is:

Minimization of categorization error:

> "A production that gives rise to a less distant categorization error is preferred over one that leads to a more distant error". (9.11)

This can be formalized as (if y_1 and y_2 are on the same side of x):

$$\text{*REPLACE } (\textit{feature}: x, y_1) \gg \text{*REPLACE } (\textit{feature}: x, y_2) \Leftrightarrow |y_1 - x| > |y_2 - x| \qquad (9.12)$$

Because of the discreteness of categorization (if only a finite number of *CATEG constraints are dominated), it now becomes sensible to talk about a homogeneous version like *REPLACE (*feature*: *x*): "do not replace the feature value *x* by *any* different value". With *VARY, this would have made no sense because *VARY (*f*: 0) would always be at the extreme lower end of the constraint system: it is utterly unimportant to have a F_1 which is within 0.01 Hz from the desired value, whereas recognizing, say, /tɛns/ as the neighbouring /tæns/ (in a language that has both of them) could already constitute a noticeable problem. The constraint family associated with generalizing this over all values of *x*, could be called *REPLACE (*feature*); if featural correspondence is forced by segmental correspondence (ch. 12), such a family can be identified with the homogeneous segment-based IDENTIO (*feature*) constraint proposed for hybrid features by McCarthy & Prince (1995). However, we will see in §9.5 that the ranking of *REPLACE generally depends on its arguments *x* and *y*.

The ranking effects of (9.12) will be seen only for features that have been divided into many categories, like vowel height. Thus, for |tɛns| in a language with four vowel heights, the outputs [tæns] and [tens] will be less offensive than the output [tins]. We can see how this works if we assign numeric values to the variation. For instance, figure 9.1 shows the distributions of the acoustic and perceptual results of a large number of replications of four vowel heights with specifications of 260, 430, 580, and 810 Hz, assuming a Gaussian model with equal standard deviations of 100 Hz (which could be caused by variations within and between speakers and by background noise).

With the help of figure 9.1, we can make a probabilistic version of what was presented in figure 8.1 as strict ranking. The shaded area in figure 9.1 represents the events in which /ɛ/ was intended, but /e/ was recognized. Its area is 0.218 (relative to the area under the Gaussian curve). The following table shows the conditional probabilities $P\left(f_{perc} = y_j \,\middle|\, f_{prod} = x_i\right)$ (the "|" reads as "given that") of classifying the four intended categories x_i into each of the four categories y_j available for perception:

f_{prod}↓ f_{perc}→	/i/	/e/	/ɛ/	/æ/	$P(f_{prod} = x)$
/i/	0.802	0.191	0.007	$8{\cdot}10^{-6}$	0.25
/e/	0.198	0.575	0.223	0.004	0.25
/ɛ/	0.009	0.218	0.648	0.125	0.25
/æ/	$2{\cdot}10^{-6}$	0.001	0.124	0.875	0.25
$P(f_{perc} = y)$	0.252	0.246	0.251	0.251	

(9.13)

The right column contains the marginal probabilities $P\left(f_{prod} = x_i\right)$ of the four intended classes x_i, and the bottom row contains the total probability of finding each of the four initial recognitions y_j:, i.e. $P\left(f_{perc} = y_j\right) = \sum_i P\left(f_{perc} = y_j \mid f_{prod} = x_i\right) P\left(f_{prod} = x_i\right)$.

Under the assumption of complete categorical perception, the best global strategy for the recognition of the categories is for the listener to assume the following Bayesian probabilities for the intended sounds, as functions of the initial categorization:

$$P\left(f_{prod} = x \mid f_{perc} = y\right) = \frac{P\left(f_{perc} = y \mid f_{prod} = x\right) P\left(f_{prod} = x\right)}{P\left(f_{perc} = y\right)} \tag{9.14}$$

This results in the following table for these probabilities (the sum of each row is 1):

f_{perc}↓ f_{prod}→	/i/	/e/	/ɛ/	/æ/
/i/	0.795	0.196	0.009	$2{\cdot}10^{-6}$
/e/	0.194	0.584	0.221	0.001
/ɛ/	0.007	0.223	0.646	0.124
/æ/	$8{\cdot}10^{-6}$	0.004	0.125	0.871

(9.15)

We can now see that a more distant *REPLACE violation is worse than an "adjacent" *REPLACE violation: if the speaker produces [t^{h}ens], the listener hears /t^{h}ens/ but the candidate /t^{h}ɛns/ has still a probability of 22.1% of being the correct candidate; if the speakers produces [t^{h}ins], the listener hears /t^{h}ins/ and the candidate /t^{h}ɛns/ has only a probability of 0.9% of being the correct candidate. Thus, during the process of recognition, which, apart from the initial phonological classification, involves the lexicon, the syntax, and the semantics, the candidate /t^{h}ɛns/ has a much larger chance of emerging on top if the production was [t^{h}ens] than if the production was [t^{h}ins].

The conclusion of this is that even in the idealized case of complete categorical perception before recognition, the *REPLACE constraints can be universally ranked. The reader would probably have believed this without all the above machinery, but we will need it again for a more complicated case below.

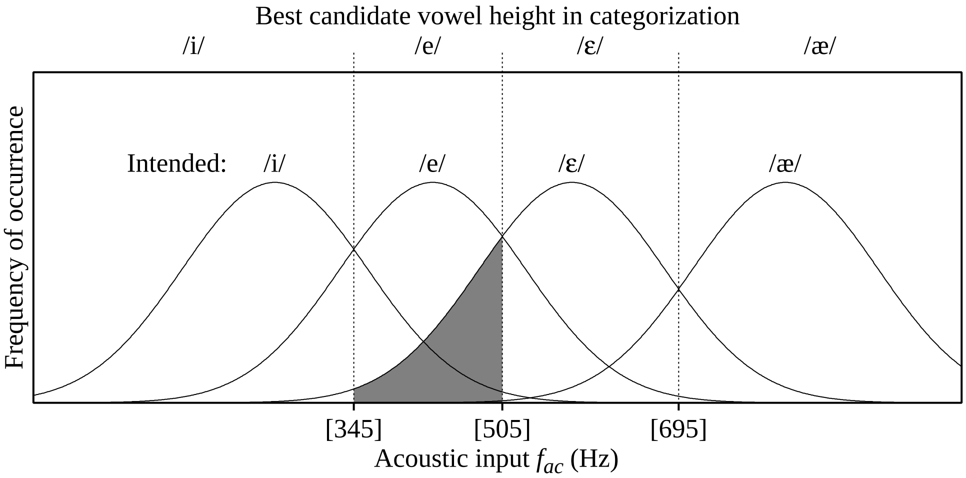


Fig. 9.1 The curves represent the variation in the production of four equally strong categories. The horizontal axis is the realized acoustic result. Along the top, the speaker's optimal classification is shown.

9.3 The emergence of equally spaced categories

In figure 9.1, we see that the centres of the production distributions are not necessarily equal to the centres of the perceptual categories. For /ɛ/, the centre of the production was 580 Hz, whereas the centre of the perceptual category came out as 600 Hz, which is the midpoint between the two criteria that separate /ɛ/ from /e/ and /æ/. This seems an unstable situation. The speaker will cause fewer confusions in the listener if she produces an /ɛ/ right into the middle of the perceptual category, namely at 600 Hz. Thus, the slight asymmetry that arises in figure 9.1 as a result of the different distances from /ɛ/ to /e/ and /æ/, may cause a category shift from 580 to 600 Hz. This shift causes the criteria to move to the right, which induces a new shift. The equilibrium will be reached when the centre of the /ɛ/ category will be in the middle between the centres of /e/ and /æ/, i.e., at 620 Hz. Thus, the drive to equalize the category centres of production and perception favours the emergence of equal spacings between the categories, if they are equally strong.

Another prediction of this model is that languages tend to have their back vowels at the same heights as their front vowels, because they use the same F_1 categorization. If the number of back vowels is different from the number of front vowels, there is a tension between minimization of the number of height categories that have to be recognized, and equalization of the height distinctions among the front and back vowels separately, so our model predicts that in this case, the language will either still have its front and back vowels on the same height, or have them spaced at locally optimized distances.

9.4 Extreme feature values

In figure 9.1, the extreme categories /i/ and /æ/ behave differently from /ɛ/. If we assume an undominated PERCEIVE constraint, all feature values above 695 Hz will be perceived as /æ/. There is no centre, then, of the perceptual category /æ/; rather, its value is specified as "max" (maximal). The associated production constraint is

Def. MAXIMUM (f: v) $\equiv \exists x_i \in f_{spec} \wedge \exists y_i \in f_{ac} \Rightarrow \left(x_i = \text{“max”} \Rightarrow y_i > v \right)$
"If the value x on a tier f in the input is specified as "max", its acoustic correspondent y, if any, should be greater than any finite value v." (9.16)

For the non-categorizing listener, this constraint ensures the lowest probabilities of recognition into the adjacent category /ɛ/. The universal ranking is:

Maximization of the maximum:
"For "max" specifications, lower produced values are worse than higher values." (9.17)

This can be formalized as

$$\text{MAXIMUM } (\textit{feature}: v_1) >> \text{MAXIMUM } (\textit{feature}: v_2) \Leftrightarrow v_1 < v_2 \qquad (9.18)$$

Of course, analogous MINIMUM constraints should also be assumed.

The name of MAXIMUM is deliberately ambiguous. On the one hand, it can be seen as a universal constraint, because its logical formulation asserts that it only actively applies to features specified as "max". On the other hand, it can be seen as a language-specific output-oriented constraint: "the value of *feature* is maximal".

Since it is impossible for the produced value to reach infinity, the actually realized value will depend on the interaction of the MAXIMUM constraints with the articulatory constraints, which tend to disfavour extreme perceptual results (see §10.4).

9.5 Category strength: ranking *REPLACE by markedness

This section describes a strategy for determining universal rankings of *REPLACE constraints.

Of the labial and coronal gestures, the coronal seems to be the 'easier', since it is this articulator that is used most in many languages (the three stops most common in Dutch utterances are /n/, /d/, and /t/), and the coronal gesture can often occur in places where the labial (or velar) gesture cannot. If the inference of easiness from frequency is correct (the asymmetry could also be due to coronals making better voicing contrasts etc.), we hereby identify the universal tendency *GESTURE (lip) >> *GESTURE (blade). But if there are more coronal than labial gestures in an average utterance, the distinctivity of the acoustic correlate of the labial gesture is larger than that of the coronal gesture. In this section, we will see how the listener reacts to this bias.

Imagine that we have two gestures, [lip] and [blade], and that the lip gesture is more difficult (or slower) than the blade gesture. Thus, *GESTURE (lip) >> *GESTURE (blade).

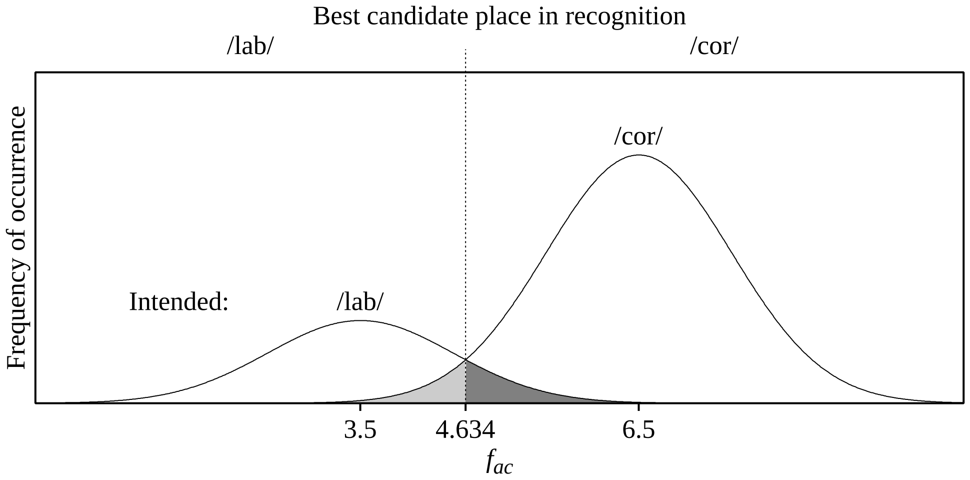


Fig. 9.2 Variation in production and acoustics causes an overlap of acoustic regions, leading to probabilistic recognition strategies in the listener.

The result of this is that in a certain language, the blade gesture is used three times as much for plosive consonants as the lip gesture. Imagine further that the perceptual categories that correspond with these gestures are [labial] and [coronal], both measured along a perceptual dimension of place. What is the best categorization strategy for the listener, i.e., where along the place dimension does she have to put her criterion for distinguishing the two feature values in order to make the fewest mistakes?

Suppose that the auditory inputs from both gestures show variations (perhaps from imperfections in the production or from background noise) whose distributions can be described by Gaussian curves with equal standard deviations σ. Figure 9.2 shows, then, the distributions of the auditory input of a large number of replications of lip and tip gestures, produced with a ratio of 1 to 3, where the distance between the averages μ_1 and μ_2 is 3σ. The curve for [coronal] is three times as high as the curve for [labial].

The best criterion for discriminating the two categories is the point along the place dimension where the two curves cross, which is to the left of the mid-point between the averages, or, to be precise, at

$$\frac{\mu_1+\mu_2}{2}-\frac{\sigma^2 \ln 3}{\mu_2-\mu_1} \tag{9.19}$$

With this criterion, the total number of confusions (the shaded area) is minimal: if you shift the criterion to the left or to the right, the shaded area will still contain everything that is shaded in figure 9.2, and a little more.

We can now derive a bias for confusion probabilities. We see from the figure that the shapes of the shaded areas to the left and to the right of the criterion are very similar, which tells us that the expected absolute number of incorrect [labial] categorizations is about equal to the number of incorrect [coronal] categorizations. However, the ***probability*** that a lip gesture is recognized as [coronal] equals the shaded area to the right of the criterion, *divided by* the total area under the [labial] curve, and the probability that a blade gesture is recognized as [labial] equals the shaded area to the left of the criterion, divided by the total area under the [coronal] curve. So we must expect from the ratio of the areas of the Gaussians that the probability that a lip gesture is recognized as [coronal] is approximately three times as high as the probability that a blade gesture is recognized as [labial]. The exact ratio, as a function of the distance between the averages, is

$$\left(\tfrac{1}{2}-\tfrac{1}{2}erf\left(\tfrac{1}{2}\sqrt{2}\left(\frac{d}{2}-\frac{\ln 3}{d}\right)\right)\right)\Big/\left(\tfrac{1}{2}-\tfrac{1}{2}erf\left(\tfrac{1}{2}\sqrt{2}\left(\frac{d}{2}+\frac{\ln 3}{d}\right)\right)\right) \tag{9.20}$$

where d is the distance between the averages, expressed in standard deviations (in figure 9.2, d is 6.5 – 3.5 = 3). For strongly overlapping distributions, which can occur if the background noise is very strong, the ratio increases dramatically. Thus, we predict that relatively uncommon feature values will be mistaken for their relatively common neighbours more often than the reverse, and that this bias is stronger for higher levels of background noise. This prediction is corroborated by some data:

- Pols (1983) for Dutch: initial /m/ is recognized as /n/ 26.1% of the time, the reverse confusion occurs 10.4% of the time; the plosives show a slight reverse bias: 5.4% versus 7.1%.
- Gupta, Agrawal & Ahmed (1968) for Hindi: initial /m/ becomes /n/ 67 times, the reverse occurs 27 times; /p/ → /t/ 66 times, the reverse 7 times (all sounds were offered 360 times).
- English /θ/ is more often taken for /f/ than the reverse.

This asymmetry will inform us about the ranking of *REPLACE (place: lab, cor) versus *REPLACE (place: cor, lab). The example of figure 9.2 gives the following confusion probabilities, obtained by dividing the shaded areas by the areas of the Gaussians:
$P\left(place_{perc}=\text{cor}\mid place_{prod}=\text{lab}\right)=12.8\%$,
$P\left(place_{perc}=\text{lab}\mid place_{prod}=\text{cor}\right)=3.1\%$.
Thus, from every 100 replications of a [place] specification, we expect the following numbers of occurrences of produced and perceived values:

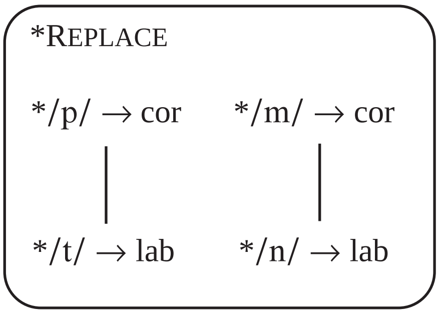
prod↓ perc→	lab	cor	total produced
lab	21.8	3.2	25
cor	2.3	72.7	75
total perceived	24.1	75.9	100

(9.21)

Doing the Bayesian inversion (9.14) (for our pre-categorizing listener) from the columns in this table, we can see that the probability that a perceived [labial] should be recognized as a produced [coronal], is 2.3 / 24.1 = 9.6%. In figure 9.2, this is the ratio of the lightly shaded area and the sum of the two areas at the left of the criterion. Likewise, the probability that a perceived [coronal] should be recognized as [labial], is 3.2 / 75.9 = 4.2%. In other words, perceived labials are far less reliable than perceived coronals.

Now consider a language in which underlying NC clusters arise from the concatenation of two morphemes. If coronals are three times as common as labials, 9/16 of those clusters will be /anta/, 1/16 will be /ampa/, and both /amta/ and /anpa/ will occur 3/16 of the time. We will now determine which of the two, /amta/ or /anpa/, will be more likely to show place assimilation.

If /amta/ is produced as [anta] (because the speaker deletes the labial gesture), the listener assigns the candidate /amta/ a probability of 4.2% · 95.8% = 4.1% (at least if she makes the [coronal] feature of [n] correspond to the [labial] feature of /m/; see §12 for a discussion of this ***segmental hypothesis***). If, on the other hand, /anpa/ is produced as [ampa], the candidate /anpa/ still has a probability of 9.6% · 90.4% = 8.7%. Comparing these figures, we see that for a successful recognition of NC clusters, it is much more detrimental to replace a [labial] specification with a [coronal] output than the reverse. This means that a faithful surfacing of the labial place feature is more important than a faithful surfacing of the coronal place feature. Thus, because the speaker is also a listener, the constraint *REPLACE (place: lab, cor) must be ranked higher than *REPLACE (place: cor, lab). This gives the following partial universal ranking tendency of *REPLACE, written as segmental filters:



(9.22)

We thus see that the weaker specification (which may, at the surface, look like ***underspecification***, see §13) of coronals is the ultimate result of an asymmetry in articulatory ease (or any other cause that leads to a frequency bias). This unmarkedness conspiracy can be summarized as follows:

$$\begin{gathered}
\text{*GESTURE (lower lip)} \gg \text{*GESTURE (tongue tip)} \\
\rightarrow \\
\text{frequency (tongue tip)} > \text{frequency (lower lip)} \\
\rightarrow \\
\text{frequency (place = coronal)} > \text{frequency (place = labial)} \\
\rightarrow \\
P\left(place_{perc} = \text{cor} \,\middle|\, place_{prod} = \text{lab}\right) > P\left(place_{perc} = \text{lab} \,\middle|\, place_{prod} = \text{cor}\right) \\
\rightarrow \\
P\left(place_{prod} = \text{lab} \,\middle|\, place_{perc} = \text{cor}\right) < P\left(place_{prod} = \text{cor} \,\middle|\, place_{perc} = \text{lab}\right) \\
\rightarrow \\
P\ (prod = \text{/amta/} \mid perc = \text{[anta]}) < P\ (prod = \text{/anpa/} \mid perc = \text{[ampa]}) \\
\rightarrow \\
\text{*REPLACE (place: labial, coronal)} \gg \text{*REPLACE (place: coronal, labial)}
\end{gathered} \qquad (9.23)$$

Since, like labials, *dorsal* stops are also less common than coronals in most languages, the same ranking is expected for dorsals versus coronals (see §11.7 for Tagalog). Ranking (9.23) predicts that there are languages that show assimilation of coronals but not of labials and dorsals, namely, those languages where an articulatory constraint like *GESTURE is ranked between the two *REPLACE constraints (ch. 11):

/anpa/	*REPLACE (place / _V)	*REPLACE (place: lab, cor / _C)	*GESTURE	*REPLACE (place: cor, lab / _C)
[anpa]			*!	
☞ [ampa]				*
/amta/				
☞ [amta]			*	
[anta]		*!		

(9.24)

Note that the ranking difference between *GESTURE (lips) and *GESTURE (blade) must be small for this to work; they are represented here as a single homogeneous constraint. The deletion of the coronal gesture in [ampa] is accompanied by a lengthening of the labial gesture; thus, the candidate [aãpa] must lose because a constraint for the preservation of the link between nasality and non-orality outranks *HOLD (labial). We fix the direction of assimilation by noting that perceptual place contrasts are larger before a vowel than in other positions because of the presence of an audible release, so that the environmentally conditioned universal ranking *REPLACE (place: *x*, *y* / _V) >> *REPLACE (place: *x*, *y* / _C) appears to be valid. The environments "_V" and "_C" refer to material present in the *output*, because that is the place where perceptual contrast between utterances must be evaluated.

We thus derived a picture that is radically different from Prince & Smolensky (1993), who confound articulatory and perceptual principles by stating that "the constraint hierarchy [*PL/Lab >> *PL/Cor] literally says that it is a more serious violation to parse labial than to parse coronal" (p. 181). Moreover, they attribute this to "Coronal Unmarkedness", an alleged principle of Universal Grammar. We can replace the claim of built-in references to phonetic content with a functional explanation: the ranking (9.23) follows from a general principle of perception: the adaptation of the listener's expectations to variations in the environment.

9.6 Information

Following the reasonings from §8.8 and §9.5, you could think that the [coronal] category would eat away at the [labial] category until there were no labials left. In general, there are no classification errors if there is only a single category. However, this process is checked by another principle of communication: "maximize the information content of the average utterance" (§1.1). The information (measured in bits) that can be stored in every instance of a feature is

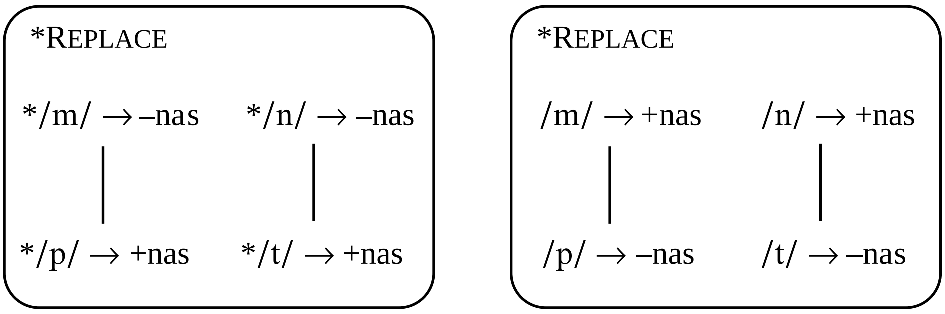
$$-\sum_i P\left(f = x_i\right) \log P\left(f = x_i\right) \tag{9.25}$$

where the sum is over all categories. For instance, if a binary feature has two equally common categories, the information content is 1 bit per instance. If a binary feature has a category that occurs three times as much as the other category, the information content is $-0.75 \cdot \log_2 0.75 - 0.25 \cdot \log_2 0.25 \approx 0.8$ bits per instance. This means that for transferring 4 bits of information, an utterance should have a length of five instead of four instances of such a feature, which is not a world-shattering problem. However, if the frequency ratio of the two categories is 1000, a 100 times greater length would be required. Somewhere, an optimum exists, and it may be found by a technique analogous to the one that will be developed for the interaction between articulatory effort and perceptual contrast in §10.4.

9.7 Binary features

Several features are categorized with only two values in most languages. A typical example is [nasal], which can be seen as having the possible values “max” and “min”, which we can write as [+nasal] and [–nasal] because our notation does not have to heed any other values. Somewhat more symmetrically, we have [H] and [L] on the tone tier in some languages.

For binary features, the *REPLACE constraints are simplified to having a single argument: *REPLACE (nasal: +, –) is not any different from *REPLACE (nasal: +), because [+nasal] cannot be replaced with anything but [–nasal]. So we write *REPLACE (+nasal), *REPLACE (H) etcetera. Analogously to the argument of §9.5, we can posit universal rankings for binary features as functions of the commonness of their values. For the feature [nasal] (§1.2.6), this would give the following universal ranking:



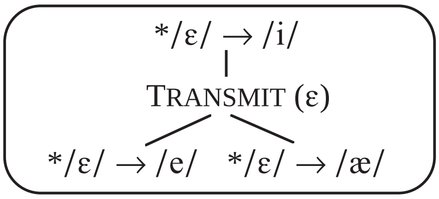
(9.26)

Next to the usual filter notation on the left, we see an equivalent positive notation on the right: *REPLACE constraints expressed directly as specifications. This is possible only for binary features. A word of caution is appropriate here: the positive formulation of the specification /m/ → [+nas] obscures the fact that the constraint is vacuously satisfied if correspondence fails, e.g., if no segment corresponding to /m/ appears in the output; the correct interpretation is more straightforward with a negative formulation.

9.8 Correspondence strategy for binary features

Correspondence is a part of the input-output relationship, and as such it is evaluated by the faithfulness constraints; no separate theory of correspondence is needed.

We will consider an interesting interaction between the correspondence constraint TRANSMIT and the identity constraint *REPLACE for features with few values. In the case of the four-valued height feature discussed in §9.2, the listener could have followed the strategy of finding out the vowel height by guessing. On the average, this would give a result that is 1.25 categories away from the intended category (1.25 is the average of 0, 1, 2, 3; 1, 0, 1, 2; 2, 1, 0, 1; 3, 2, 1, 0). Such a strategy would, therefore, be only slightly less advantageous than recognizing an intended category into an adjacent category, but more advantageous than a recognition that is off by two categories. This gives the following ranking:

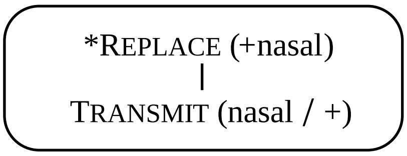


(9.28)

A similar ranking would be derived for PERCEIVE and *WARP.

A probabilistic argument will also work. In the case of the vowel in /tɛns/, it would not have been worse for the speaker not to produce any value for F_1 at all, than to produce an /e/. If the listener has to find out the vowel height by guessing, the /ɛ/ will have a probability of 25%, which is not worse than the probability that a perceived /e/ should be recognized as /ɛ/, which was 22.1% in our example. The probability that a perceived /æ/ should be recognized as /ɛ/ is even smaller: 12.5%. So, with the locations and widths of §9.2, all *REPLACE constraints would be ranked higher than TRANSMIT.

This situation is even stronger for features with two categories: it will always be better not to produce any value (50% correct from guessing), than to produce the wrong value (always less than 50%); or, by guessing, you will be half a category off, on the average, and by choosing an adjacent category you will be one category off, which is worse. Thus, binary features will always have a stronger *REPLACE than TRANSMIT constraint:



(9.29)

Now, because a violation of TRANSMIT will automaticaly cause the satisfaction of the higher *REPLACE, the best strategy for the listener will be not to make the output feature correspond to the specification at all, if no other constraints interact with TRANSMIT:

tεns │ +nas$_i$	*REPLACE (+nas)	TRANSMIT (nas / +)
tεts │ –nas$_i$	*!	
☞ tεts │ –nas$_j$		*

(9.30)

If the listener follows the strategy described here, the *REPLACE constraint will be invisible in her grammar, and a single combined TRANSMIT-*REPLACE constraint, equally highly ranked as the original TRANSMIT, will do the job. It combines a negative with a positive attitude:

Def. *DELETE $(f: x) \equiv \exists x_i \in f_{spec} \Rightarrow \left(\exists y_i \in f_{perc}: y_i = x_i \right)$
"An underlyingly specified value x of a perceptual feature f appears (is heard) in the surface form." (9.31)

For instance, we have *DELETE (tone: H) and *DELETE (nasal: +), which can easily be abbreviated as *DELETE (H) and *DELETE (+nasal). Note that *DELETE (*feature*) cannot be satisfied by deletion of its bearing segment, in other words: *DELETE (*feature*) can actually force the parsing of whole segments, if ranked above *DELETE (timing: X).

Because of the impossibility of vacuous satisfaction of *DELETE, a positive name would be appropriate. In line with current usage, which refers to the surfacing of underlying material with the term "parsing", we will sometimes use the name PARSE, which originally comes from Prince & Smolensky (1993), who restricted it to the parsing of a prosodic constituent, like a segment, into a higher constituent, like a syllable. McCarthy & Prince (1995) coined a similar constraint MAX-IO, as an analogy with MAX-BR, which stated that a Reduplicant should take the maximum number of segments from the Base. For the faithfulness of hybrid features, some names based on the slightly inappropriate PARSE and MAX are: PARSE$^{\text{FEAT}}$ (Prince & Smolensky 1993), PARSEFEAT (Itô, Mester & Padgett 1995), MAXF (Lombardi 1995), MAX(FEATURE) (Zoll 1996). Also, in a declarative wave, we may decide to give this constraint no name at all, taking the specification "/+nasal/" or "∃[+nasal]" to mean: "there should be a [+nasal] in the output". In any case, a universal ranking for [nasal] is given by

*DELETE (+nasal)
│
*DELETE (–nasal)

PARSE (+nasal)
│
PARSE (–nasal)

(9.32)

which expresses the cross-linguistic preference for the assimilation of [+nasal] as in /akma/ → [aŋma], over the assimilation of [–nasal] as in /aŋpa/ → [akpa]. Besides promoting the presence of specified material in the output, a specification also implicitly states that unspecified material does *not* surface. If *REPLACE dominates DONTTRANSMIT, we have

Def. *INSERT (f: y) $\equiv \exists y_i \in f_{perc} \Rightarrow \left(\exists x_i \in f_{spec}: x_i = y_i\right)$
“A value y of a perceptual feature f, that is heard in the surface form, corresponds to the same underlying feature value.” (9.33)

For instance, a replacement of /H/ by /L/ now violates both *DELETE (H) and *INSERT (L), if the listener manages not to make the two values correspond:

/H_i/	*REPLACE (H)	TRANSMIT (tone / H) *DELETE (tone: H)	DONTTRANSMIT (tone / L) *INSERT (tone: L)
/L_i/	*!		
☞ /L_j/		*	*

(9.34)

Again, because of its combined negative/positive interpretation, a positive name like FILL (Prince & Smolensky 1993) or DEPF (McCarthy & Prince 1995) could be used instead of *INSERT. For the feature [nasal], we could superficially translate (9.32) into the fixed ranking (still restricted to assimilation):

*INSERT (–nasal)
|
*INSERT (+nasal)

FILL (–nasal)
|
FILL (+nasal)

(9.35)

but this would only be valid under a linear OCP-less interpretation of perceptual correspondence (ch. 12).

The overlapping functions of *DELETE and *INSERT for binary features will be collapsed in §9.9 for those features which can be considered monovalent.

As an example of how a feature reversal as in (9.34) may come about, consider the floating H prefix found in Mixteco, as analysed by Zoll (1996: ch. 2). An underlying LM sequence as in /kìku/ ‘child’, enriched with the H affix, gives a HM sequence (/kíku/):

/kìku/ + H	*DELETE (tone: H / H-affix)	*DELETE (tone: L / base)
kìku	*!	
☞ kíku		*

(9.36)

Zoll notes that the constraint MAX(FEATURE) (i.e., a homogeneous *DELETE (tone)) does not do the job, not even if helped by IDENT(F), which is roughly a segment-based homogeneous *REPLACE (tone). This situation is reason for Zoll to propose a constraint MAX(SUBSEG), which we could translate as a homogeneous *DELETE (tone / floating). However, I can think of no functional explanation as to why the ranking of a constraint should depend on whether a feature is linked or not. Rather, two alternative approaches (also touched upon by Zoll), combined in the formulation of (9.36), follow from the theory of Functional Phonology developed so far.

First, we note that §9.5 proved that *DELETE constraints should be parametrized with feature values, because their ranking depends on the commonness of the feature values. For Mixteco, we could have *DELETE (tone: H) >> *DELETE (tone: L), or Zoll's MAX(H) >> MAX(L). With such a ranking, a floating L-affix would only be able to affect one of the eight possible tone sequences of Mixteco (namely, MM), whereas the floating H-affix affects four of them (MM, LH, LM, and ML); this would explain why Mixteco does not have any L-affixes.

The second possibility is conditioning the ranking by the base/affix opposition: *DELETE (tone / H-affix) >> *DELETE (tone / base), or Zoll's MAX (affix) >> MAX (base). This would be the approach when H and L values are equally common in the language, so that neither of them can be considered unmarked. Morphological conditioning of faithfulness is quite common: the cross-linguistic tendency *DELETE (*feature* / base) >> *DELETE (*feature* / affix) has an obvious functional explanation (it is more important to keep all the information in content morphemes than to keep all the information in function morphemes), and manifests itself in the preference for base-to-affix spreading above affix-to-base spreading in vowel-harmony systems. The reversal of this ranking in the Mixteco case, where a failure to parse the H tone would obscure the entire affix, can be attributed to the idea that it is more important to keep *some* information about the affix than to keep *all* the information about the base. I would like to contend that functional arguments like these are the real explanations for facts of ranking (this is not the sole role of function in the grammar: even if most of the rankings are given, function is needed in describing the competence of the speaker, at least at the postlexical level, as shown in ch. 10 and §11.5).

9.9 Privative features

Unary features are a special kind of binary features (§1.2.6).

For nasality, the probability of correct categorization depends on the quality of the nasality cues (heights of spectral peaks and depths of valleys) in the acoustic signal. It is probable that the categorization of this feature for almost all existing languages has resulted in two perceptually distinct values for the feature [nasal]: *present* and *absent*. With many aspects of perception, there is an asymmetry, a qualitative difference, between presence and absence. Also in this case, non-nasality is the default: perceptually, nasality is associated with some extra peaks and valleys in the auditory frequency spectrum, as compared to the more common spectra of vowels. Thus, we can posit the existence of a single-valued perceptual feature of nasality, and (1.11) contains the specification [nasal]. The following constraint ensures that it is present in the output:

Def. PARSE $(f) \equiv \exists x_i \in f_{spec} \Rightarrow \exists y_i \in f_{perc}$

"A specified feature *f* appears (is heard) in the surface form." (9.37)

This constraints plays the parts of both TRANSMIT and *REPLACE, because you cannot replace a value of a unary feature with any other value, and it is equivalent to *DELETE. Thus, if PARSE (nasal) is violated, /tɛns/ will surface as [t^hɛts].

Not all features are privative. The feature [sibilant], for instance, is not a clear candidate for a privative feature: failure to satisfy an alleged PARSE (sibilant) would result in the output [tɛnt] or [tɛnθ]; but the latter is better because TRANSMIT (noise) probably dominates *REPLACE (noise: sibilant, strident) (§1.2.5), in contrast with requirement (9.29) for the existence of PARSE.

Also, we may have PARSE (coronal) and PARSE (labial), if the separate place features have their own tiers instead of being values of a perceptual feature [place]. But this is doubtful. For instance, the fact that it is less offensive to replace [θ] with [f] than to replace it with [χ], suggests a single perceptual feature [place], with *REPLACE constraints ranked by the perceptual contrasts of their argument pairs.

The global ranking of PARSE for unary features could be thought to depend on:

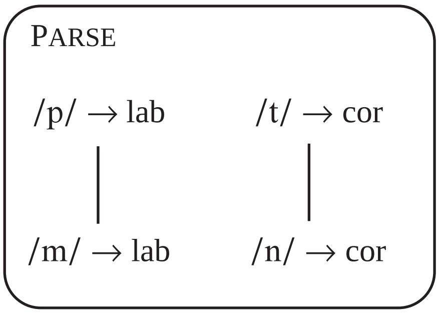
Maximization of conservation of salience:

“The greater the distinctive power of a feature (value), the higher the ranking of its specification.” (9.38)

The parenthesized “value” in (9.38) suggests that multivalued features may also show a presence/absence asymmetry. On the noise scale, for instance, we have [aspirated], [fricative], and [sibilant], next to the absence of noise. For instance, if [sibilant] is a salient feature value, the contrast between [asa] and [ata] is large, so that the probability of the candidate /asa/ if the listener hears [ata], is low; if [aspiration] is a less salient feature value, the contrast between [aka] and [ak^{h}a] is small, so that the probability of the candidate /ak^{h}a/ is reasonably high, even if the listener hears [aka]. This would imply that TRANSMIT (noise / sibilant) >> TRANSMIT (noise / aspiration): it is less bad for the speaker to leave out underlying aspiration than it is for her to leave out sibilancy.

However, it will be strongly language-dependent what features are considered salient and what are not. After all, it is a common property of human perception that it is difficult to compare unlike entities along scales like “conspicuity”, “salience”, or “notability”. For instance, people would disagree about whether a *duck* or a *lamp post* were the more conspicuous of the two. Thus, the conjecture (9.38), which, by the way, expresses the same idea as the ***production hypothesis*** of Jun (1995) (though that referred to acoustic cues, not perceptual features, see §11.8), is subject to language-specific variation and can, at best, be used to explain cross-linguistic ***tendencies***, or the workings of very large salience/conspicuity contrasts, such like that between an *airplane* and a *tulip* (though even that depends on the environment).

For practical purposes, the ranking (9.38) is valid only for comparisons of a feature *with itself* in different environments. A clear example (for non-unary features) is the confusion probability of [m] and [n], as compared with the confusion probability of [p] and [t]. Measurements of the spectra of these sounds agree with confusion experiments (for Dutch: Pols 1983), and with everyday experience, on the fact that [m] and [n] are acoustically very similar, and [p] and [t] are farther apart. Thus, place information is less distinctive for nasals than it is for plosives. This means that for the understanding of the utterance, the emergence of the underlying place information in the actual phonetic output is less important for nasals than for plosives. In constraint terminology, we can express this as a general ranking of two parsing constraints, namely that PARSE (*place* / plosive) dominates PARSE (*place* / nasal). An alternative terminology would represent these constraints directly as *specifications*, e.g., /m/ → [labial]. A partial representation of the PARSE family will then look like (cf. 9.26):



(9.39)

A more accurate account would use *REPLACE instead of PARSE, as in ch. 11.

The unary-feature version of both DONTTRANSMIT and *INSERT is:

Def. FILL (*f*:) ≡ $\exists y_i \in f_{perc} \Rightarrow \exists x_i \in f_{spec}$

"A feature *f* that is heard in the surface form, also occurs in the specification." (9.40)

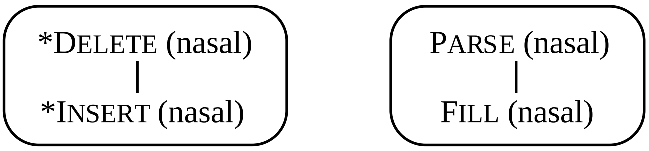
The implementation of /ɛns/ as [[ɛɛ̃n_ts]], chosen in order to satisfy two synchronization constraints, involved the epenthesis of a silence plus stop burst, in other words, a violation of FILL (plosive) (or FILL (silence) and FILL (burst) if we are talking autonomous-cue faithfulness, but according to §11.8, we should not). The alternative ordering of the two contours between [n] and [s] would give [[ɛɛ̃nəs]] or so, epenthesizing a syllable and thus violating FILL (timing: σ). Depending on the language, one or the other is worse. However, the epenthesis of [t] in this environment of a coronal nasal and a coronal obstruent is not as bad as the epenthesis of [t] between the elements of the sequence [ia]; a [j] would be more appropriate there. This means that the ranking of the FILL constraints depends strongly on the environment, and that the ranking is especially low if the syntagmatic perceptual salience of the utterance is hardly increased, as is the case in [[ɛɛ̃n_ts]] (continuous place information) and in [ija]:

Minimization of intrusive salience:

"The greater the distinctive power of a feature, the higher the problem of its insertion into the output." (9.41)

We can note that the implementation of /ɛn/ as [[ɛɛ̃n]] does not violate any FILL constraint, because all output features are already present in the input. Reversing the order of the two contours involved would give [ɛt^{n}], with epenthesis of a nasal plosive, which would be worse than the other candidate no matter how the language ranks the FILL constraints. This explains the universal preference for the implementation of the vowel-nasal transition with the nasal contour first; the few languages that do implement /ɛn/ as [ɛt^{n}], may heed a "path" (§9.11) constraint against the insertion of simultaneous nasality and vocalicness.

An interesting property of faithfulness constraints is that they do not distinguish between unary and binary features. If we conjecture, analogously to the universal ranking for place features (9.22), that the relative uncommonness of nasals in the average utterance causes the universal rankings (9.26), (9.32), and (9.35), we could equally well phrase this in privative terminology as the following near-universal ranking for the unary perceptual feature [nasal]:



(9.42)

So, whether or not we specify a value for [–nasal] in (1.11), makes little difference, if any.

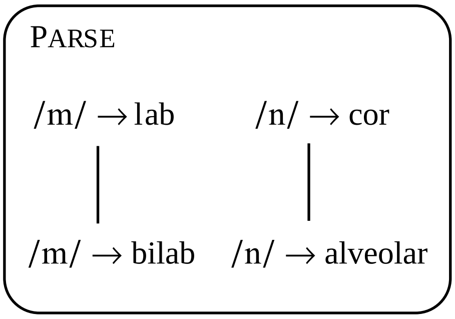
9.10 Ranking by specificity

Besides considerations of contrast, PARSE can be universally ranked by the specificity (perceptual precision) of its arguments. Like in the case of MAX, where a less specific constraint like F1 > 500 Hz was ranked higher than a more specific constraint like F1 > 700 Hz, we have analogous constraints for place of articulation. For instance, an /m/ is specified for [bilabial], but its [labial] specification must be stronger, because all bilabial consonants must necessarily be labial. For instance, Dutch /m/ may assimilate to a following labiodental consonant, but not to anything else; its labiality, therefore, seems more important than its bilabiality. Likewise, an /n/ is specified for [alveolar], but its [coronal] specification is stronger. These are instances of a more general principle:

Minimization of specificity:

> "More specific perceptual features are less urgent than less specific features." (9.43)

This is in line with the functional principle "if you cannot have it all, settle for something less", and completely in accord with that maxim of Optimality Theory, ***minimal violation***. After (9.22) and (9.39), we have our third partial universal hierarchy for place faithfulness:



(9.44)

The general principle (9.43) can be formalized as

$$(A \Rightarrow B) \Rightarrow \text{PARSE}\ (B) \gg \text{PARSE}\ (A) \tag{9.45}$$

or, as a generalization of PARSE (bilabial ∨ labiodental) >> PARSE (bilabial) ("∨" = "or), which, like the universal ranking of MAXIMUM, expresses the lower importance of narrow perceptual windows:

$$\text{PARSE}\ (A \vee B) \gg \text{PARSE}\ (A) \tag{9.46}$$

Note the asymmetry between articulation and perception, and between markedness and specificity:

$$\begin{array}{lcl} \text{*GESTURE (lab)} \gg \text{*GESTURE (cor)} & ; & \text{PARSE (lab)} \gg \text{PARSE (cor)} \\ \text{*GESTURE (bilab)} \gg \text{*GESTURE (lab)} & ; & \text{PARSE (lab)} \gg \text{PARSE (bilab)} \end{array} \tag{9.47}$$

Because of this asymmetry, the PARSE and *GESTURE hierarchies generally interact in such a way that there is a working-point where the perceptual problems arising from imperfect contrastivity equal the problems associated with articulatory effort and precision; an example of this will be shown in §10.

There can be counterexamples to hypothesis (9.43), forced by other constraints. In §12.7, we will see an example of the somewhat perverse principle "if I cannot have it all, I'd rather have nothing".

The above example was somewhat unrealistic, because it hinges on a hierarchical place feature, divided into several (perceptual!) articulator features. If we accept the continuity of the perceptual place feature, so that the cross-articulator contrast between [θ] and [f] is smaller than the within-articulator contrast between [θ] and [ʃ], the ranking in (9.44) reduces to the less spectacular rankings of *REPLACE (bilabial, alveolar) >> *REPLACE (bilabial, labiodental) etc., which can be immediately related to confusion probabilities.

The asymmetry in (9.47) can be formulated in terms of precision: precise articulations are disfavoured, and precise productions are not needed.

9.11 Simultaneity constraints

Besides separate feature values, the specification (1.11) contains information about simultaneity of features. For instance, the /n/ of /tɛns/ is specified as simultaneously nasal and coronal. Simultaneous feature values on the perceptual tiers f and g can combine to new feature values on a combined tier $f \times g$. For instance, the combination [coronal nasal] may be a member of a higher-level perceptual feature, say, [spectrum], and have its own correspondence and faithfulness constraints, which I will call *path constraints* as a tribute to Archangeli & Pulleyblank (1994), who use the term "path" to refer to simultaneously occurring features or nodes[1]:

[1] Within a Containment version of OT with hybrid features, Itô, Mester & Padgett 1995 suggest PARSELINK and FILLLINK as constraints for faithfulness of association lines. They use it as part of a homogeneous FAITH constraint.

Def. TRANSMITPATH $(f \times g) \equiv \exists x_i \in f_{spec} \times g_{spec} \Rightarrow \exists y_i \in f_{perc} \times g_{perc}$
“Every value x on the tiers f and g in the specification corresponds to any category y on the same tiers in the perceptual output.” (9.48)

Def. *REPLACEPATH $(f \times g\colon x, y) \equiv \exists x_i \in f_{spec} \times g_{spec} \wedge \exists y_i \in f_{perc} \times g_{perc} \Rightarrow \left|x_i - y_i\right| \le d$
“The perceived category y on the tiers f and g is not different from the specified value x by any positive distance d.” (9.49)

Def. *DELETEPATH $(f \times g) \equiv \exists x_i \in f_{spec} \times g_{spec} \Rightarrow \exists y_i \in f_{perc} \times g_{perc}$
“A specified combined unary feature on the tiers f and g appears (is heard) in the surface form.” (9.50)

Def. *INSERTPATH $(f\colon \times g) \equiv \exists y_i \in f_{perc} \times g_{perc} \Rightarrow \exists x_i \in f_{spec} \times g_{spec}$
“A combined unary feature on the tiers f and g that is heard in the surface form, also occurs in the specification.” (9.51)

For our example /tɛns/, the output [t^{h}ɛs] would violate TRANSMITPATH (place × nasal), and the output [t^{h}ɛms] would violate *REPLACEPATH (place × nasal: +nas cor, +nas lab), which is a more precise formulation than *REPLACE (place: cor, lab / +nas), because the latter wording is not explicit about whether the environment “+nas” should refer to a feature in the input or in the output or in both (but it must be the output, because that is where contrast is evaluated), and whether the input and output [+nas] should have to stand in correspondence; according to (9.49), they do not have to (and often, they do not, see ch. 12), because links are autonomous.

Normally, we will write the constraint PARSEPATH (nas & cor) simply as PARSE (nas & cor) or PARSE (coronal nasal), expressing the unity of composite features. This constraint might be expected to be ranked below the less specific PARSE (nas) and PARSE (cor) (§9.10), so that it would be redundantly violated in [t^{h}ɛms], [t^{h}ɛts], and [t^{h}ɛs], and visibly violated in [t^{h}ɛ̃ts], which satisfies both PARSE (cor) and PARSE (nas). A recalcitrant ranking of PARSEPATH (nas & cor) above PARSE (cor) and PARSE (nas) may yield an all-or-none behaviour of the surfacing of /n/; a possible case of this is shown in §12.7.

The inevitable companion of a complex PARSE is a complex FILL. For instance, [t^{h}ɛms] would violate FILLPATH (nas & lab) (which can simply be written as FILL (labial nasal)) as well as FILL (lab). Possible cases of crucial high rankings of this constraint are presented throughout §12.7. The usual output of /tɛns/, [[t^{h}ɛɛ̃n_ts]], violates FILL (nasal mid vowel) and FILL (coronal plosive).

9.12 Precedence constraints

In /tɛns/, the feature value [sibilant] should occur after the vowel (this is satisfied) and after [nasal] (also satisfied), and [nasal] should occur after the vowel (partly violated). The candidate [snɛt] would violate both of these ordering relations, except the basic CVC ordering. For segments, McCarthy & Prince (1995) proposed a constraint LINEARITY to handle this. The featural version is:

Def. PRECEDENCE $(f{:}\ t;\ g{:}\ u) \equiv \exists t_i, u_j \in f_{spec} \wedge \exists v_i, w_j \in f_{perc} \Rightarrow \left(t_i < u_j \Rightarrow v_i < w_j \right)$

“A pair of contours at times *t* and *u*, defined on two perceptual tiers *f* and *g* and ordered in the specification, have the same ordering in the output, *if they occur there*.” (9.52)

This constraint can be satisfied by deletion, because the relevant TRANSMIT constraints independently control the presence of perceptual features in the output. This constraint expresses the difficulty of making reversely ordered feature values correspond to each other. For instance, does the underlying sequence /H_iL_j/, if surfacing as LH, correspond to L_iH_j or to L_jH_i? The answer depends on the relative ranking of PRECEDENCE (tone) and *REPLACE (tone).

To clarify this, consider the relation between the input /bɛrk/ and the output /brɛk/ on the root tier (in a language that disallows branching codas, for instance). If we subscript the input as /bɛ$_i$r$_j$k/, the output candidates /br$_i$ɛ$_j$k/ and /br$_j$ɛ$_i$k/ must be evaluated separately. Because an output /r/ is made to correspond with an input /ɛ/, the first of these candidates violates *REPLACE (ɛ, r). The second candidate violates a precedence constraint on the root tier. If we call the process metathesis, the second analysis must win:

/bɛ$_i$r$_j$k/	*CC]$_\sigma$	*REPLACE (ɛ, r)	*REPLACE (r, ɛ)	PRECEDENCE (root: ɛ, r)
bɛ$_i$r$_j$k	*!			
br$_i$ɛ$_j$k		*!	*!	
☞ br$_j$ɛ$_i$k				*

(9.53)

A violation of PRECEDENCE brings about metathesis. While this phenomenon can be seen as advocating segmental integrity, this typically segmental behaviour can also arise as a consequence of the dominance of combinatory feature constraints, not necessarily at the root level. For instance, PARSE (lower mid front vowel) and PARSE (vibrant sonorant), expressing the perceptual unity of some feature paths, would have sufficed in this case, but would, admittedly, have been less simple and generalizing. On the other hand, metathesis also exists on the featural level. Consider, for instance, the complicated correspondence relations in /hˈufnit/ → [snˈuftit] ‘I don’t want to eat that’, spoken by Jildou (aged 1;10, learning Dutch): it involves hopping of the feature [nasal] to a position where it is better licensed (in her speech), leaving behind a coronal stop.

9.13 Alignment constraints

Coincidence relations exist between the beginnings and ends of the feature values in the specification. These often occur at equal times in a simple representation like (1.11): in /tɛns/, the nasal should start where the coronal starts, the vowel should end where the nasal starts, and [sibilant] should start where [nasal] ends. We can formulate a constraint that requires approximate simultaneity of the contour pairs in the output:

Def. *SHIFT (f: t; g: u; d) $\equiv \exists t_i, u_j \in f_{spec} \wedge \exists v_i, w_j \in f_{perc} \Rightarrow \left(t_i = u_j \Rightarrow v_i - w_j < d\right)$
"A pair of contours (edges) at times t and u, defined on two perceptual tiers f and g and simultaneous in the specification, are not further apart in the output (if they occur there) than by any positive distance d." (9.54)

*SHIFT expresses the difficulty for the listener to reconstruct the simultaneity of contours, and the triple implication can be logically reversed:

Correspondence-strategy interpretation of *SHIFT:
"If two contours in the output do not coincide by anything less than d, they do not correspond to simultaneous contours in the input." (9.55)

A universal ranking of *SHIFT is

Minimization of shift:
"A less shifted contour pair is preferred over a more shifted pair." (9.56)

This can be formalized as

*SHIFT (f: t; g: u; d_1) >> *SHIFT (f: t; g: u; d_2) $\Leftrightarrow d_1 > d_2$ (9.57)

The formulation (9.54) is sensitive to the direction of the shift, and, therefore, to the order of the two arguments: we do not take the absolute value of $v_i - w_j$. Thus, [[t$^{\text{h}}$ɛɛ̃n_ts]] violates *SHIFT (coronal: –|+; nasal: –|+; 50 ms), because the coronal closure lags the lowering of the velum by 50 ms; likewise, it violates *SHIFT (vowel: +|–; nasal: –|+; 50 ms), and *SHIFT (sibilant: –|+; nasal: +|–; 30 ms). In a phonologized situation, time will be measured in moras (or some such measure), instead of seconds. With unary features, we cannot refer to minus values, so we will have to refer to edges: we have *SHIFT (cor: Left; nas: Left), which does some of the work of PARSE (nas cor); *SHIFT (voc: Right, nas: Left), which does some of the work of FILL (voc nas); and *SHIFT (sib: Left; nas: Right), which expresses adjacency. In general, *SHIFT (*a*: Left; *b*: Left) expresses left alignment of *a* and *b*, *SHIFT (*a*: Right; *b*: Right) expresses right alignment, *SHIFT (*a*: Left; *b*: Right) militates against material intervening between *a* and *b*, and *SHIFT (*a*: Right; *b*: Left) militates against overlap.

If we get rid of the confusing edges, we can rephrase the four *SHIFT constraints as LEFT (*a*, *b*, *d*), RIGHT (*a*, *b*, *d*), *INTERVENE (*b*, *a*, *d*) (note the order of the arguments), and *OVERLAP (*a*, *b*, *d*).

Other general alignment constraints have been proposed. The best known is ALIGN (McCarthy & Prince 1993b):

Def. ALIGN (*cat*$_1$, *edge*$_1$, *cat*$_2$, *edge*$_2$)

"for every morphological, prosodic, or syntactic category *cat*$_1$, there is a category *cat*$_2$ so that *edge*$_1$ of *cat*$_1$ and *edge*$_2$ of *cat*$_2$ coincide." (9.57)

There are several differences between *SHIFT and ALIGN:

(a) ALIGN is homogeneous, i.e., it is not ranked by the amount of misalignment or intervening or overlapping material. It does incur a number of marks which is proportional to the extent of the violation, but this only allows ALIGN to interact with itself in the grammar. If this is realistic behaviour, the more restricted ALIGN should be preferred over *SHIFT in this respect.

(b) ALIGN is asymmetric with respect to its arguments: it is vacuously satisfied if *cat*$_1$ is missing, but not if *cat*$_2$ is missing (except under the assumption of Containment). No motivation for this asymmetry has ever been given. The alternative constraint ANCHOR, proposed by McCarthy & Prince (1995), does not show this asymmetry.

(c) ALIGN is symmetric with respect to overlap versus intervention, whereas *SHIFT allows to be ranked differently for these functionally very different situations.

(d) ALIGN is partly a positive constraint: deletion of *cat*$_2$ typically causes it to be violated. However, surfacing of *cat*$_2$ is independently controlled by its transmission constraint, so vacuous satisfaction should be allowed.

(e) ALIGN is formulated as a binary constraint; it needs a separate clause for assessing the number of violation marks. *SHIFT solves this problem with its distance parameter.

(f) ALIGN is morpheme-specific: it states the preferred positions of its arguments as constraints, whereas other (e.g., featural) specifications are part of the underlying form. *SHIFT is more consistent: if morphology is taken care of representationally, i.e., by time-aligning two contours in the input specification, the *SHIFT constraints automatically evaluate the deviations from this representational alignment. Thus, *SHIFT is language-independent, though its ranking (not its arguments) can be morphologically conditioned.

(g) ALIGN is not a faithfulness constraint. Instead of relating input and output, it evaluates the output in a declarative manner. Its implicational formulation allows it to be used for controlling *licensing*, if that happens to involve the edge of a domain. As Zoll (1996) shows, licensing does not always refer to edges, so a separate licensing constraint is needed anyway, like Zoll's COINCIDE (*a*, *b*) "if (the marked structure) *a* occurs in the output, it must be within a domain (strong constituent) *b*".

The binarity problem was noted by Zoll (1996), and she proposes an alternative:

Def. NO-INTERVENING (ρ; *E*; *D*)

"there is no material intervening between ρ and edge *E* in domain *D*."(9.58)

For concatenative affixation, Zoll rewords this as "if there is an element *x* in the base, and an affix *y*, *x* does not intervene between any part of *y* and the edge of the word"; the usual interpretation of gradient violation incurs one mark for every *x* that violates this. Besides solving the binarity problem (e), the negative formulation of this constraint fixes the problems of asymmetry (b), and vacuous satisfaction (d). Despite the existence of a

COINCIDE constraint, however, NO-INTERVENING can still be misused for licensing purposes, because it still evaluates the output only. Moreover, the *empirical* (rather than technical) differences between ALIGN and NO-INTERVENING are few (Zoll does not provide any).

The largest empirical difference between *SHIFT and ALIGN/NO-INTERVENING is the distance parameter. While both ALIGN and NO-INTERVENING must be considered gradient constraints (in their workings), *SHIFT is a family of binary constraints with fixed internal ranking based on the distance between the realized edges.

First, we will see that *SHIFT can do the work of ALIGN. I will take the cherished example of Tagalog um-infixation, but analyse it very differently from Prince & Smolensky 1993; McCarthy & Prince 1993a, 1993b et seq. The prefixation of the root /basa/ with the actor-trigger morpheme /um/ (Schachter & Otanes 1972) gives /bumasa/ 'read', and /um/ + /ʔaral/ gives /ʔumaral/ 'teach' (that's the difference: not /um/ + /aral/ → /umaral/, because prefixation of another actor trigger gives /mag/ + /ʔaral/ → /magʔaral/ 'study', not */magaral/, showing that the glottal stop can be considered underlyingly present). The undominated licensing constraint ONSET "every syllable has an onset" (rather than the very violable NOCODA, which we may only need for cluster-initial loans like /gr(um)adwet/) forces violation of the lowest possible *SHIFT constraint:

/u$_i$m$_j$ | ʔ$_k$aral/	$*_{\sigma}$[V	FILL (ʔ)	*OVERLAP (um, base, σσ)	*OVERLAP (um, base, σ)
u$_i$m$_j$ʔ$_k$aral	*!			
ʔ$_l$u$_i$m$_j$ʔ$_k$aral		*!		
ʔ$_k$u$_i$m$_j$ʔ$_l$aral		*!		*
☞ ʔ$_k$u$_i$m$_j$aral				*
ʔ$_k$aru$_i$m$_j$al			*!	*

(9.59)

Some differences with ALIGN and NO-INTERVENING appear. Because *OVERLAP refers to an alignment difference between the right side of /um/ and the left side of /ʔaral/, the amount by which it is violated in /ʔumaral/ is actually /ʔum/. The output-oriented left-alignment constraint ALIGN (um, Left, Stem, Left) measures the distance between the left edge of the substring /um/ and the left edge of the entire string (stem) /ʔumaral/, which is /ʔ/. The non-directional constraint NO-INTERVENING measures the distance between the substring /um/ and the left edge of the entire string /ʔumaral/, which is also /ʔ/ (the constraint is non-directional, i.e., able to flip between right and left according to which side is closest to the specified edge of the word).

Intuitively, describing the violation as /ʔ/ seems preferable, and we could get this result with a faithfulness constraint that honours the left-aligned specification of /um/ instead of its adjacency to the base: the idea is that the "stem" already occurs in the input specification: it is the entire string /um | ʔaral/ as specified in the input. The violated constraint would then be Left (um, “stem”, C), and *Overlap (um, base, σσ) would be replaced with Left (um, “stem”, CVC), giving a tableau completely analogous to (9.59).

However, there is some very scant (probably dubious) evidence that the *Overlap constraints as stated in (9.59) are appropriate for Tagalog: if *Overlap (um, base, σσ) dominates Fill (C) (the two are not crucially ranked for /ʔumaral/), we can explain the fact that Tagalog has no bisyllabic infixes. For instance, the instrument-trigger morpheme /ʔipaŋ/, which Prince et al. would analyse as /ipaŋ/, is a regular prefix (/ʔipaŋ-hiwa/ ‘cut with’, not */h-ipaŋ-iwa/), and Prince et al. provide no explanation for the fact that bisyllabic “vowel-initial” prefixes are exceptions to the generalization that all and only the vowel-initial consonant-final prefixes show infixation.

Positing *Shift as a family predicts that its members can interact with other constraints separately, i.e., that it shows *inhomogeneity* effects. Now, Align has always been considered a homogeneous constraint, so it would be interesting to find inhomogeneous alignment effects. Such an effect can be found in Yowlumne[2] glottalization (Newman 1944; Archangeli 1984; Archangeli & Pulleyblank 1994; Zoll 1994, 1996), in its interaction with vowel shortening.

The Yowlumne durative morpheme can be represented as the suffix /ˀaː/, where /ˀ/ represents a floating [glottal plosive][3] feature (Archangeli 1984). This feature prefers to dock on the rightmost post-vocalic sonorant, with which it combines to give a single glottalized segment: /tˢaːw-/ ‘shout’ + /ˀaː/ gives [tˢaːwˀaː]. We see that [wˀ] (the glottal constriction is centred around the middle of [w]) acts as a single segment: an utterance like *[tˢaːwʔaː] would be ill-formed in Yowlumne, because this language only allows CV, CVC, CVV syllables, so that CVVCCVV is not syllabifiable, and CVVCVV is. These syllabification requirements often lead to shortening of vowels: /ʔiːlk-/ ‘sing’ + /ˀaː/ gives [ʔelˀkaː], where we see the expected glottalization and shortening of an ill-formed VVCCV to VCCV. If there are no glottalizable sonorants, as in /maːx-/ ‘procure’ (the /m/ is not post-vocalic), the result is a full glottal stop: it appears in /maxʔaː/, with shortening of the long vowel, which proves that /xʔ/ must be analysed as a consonant cluster, not as a single glottalized obstruent. Finally, the glottal stop does not surface if there is no glottalizable sonorant and no licit syllabification: /hogn-/ ‘float’ + /ˀaː/ gives [hognaː], not *[hognʔaː]; syllabification requirements could be satisfied by an otherwise well-attested epenthesis procedure, which could give a well-syllabified *[hoginʔaː], but glottalization does not appear to be able to enforce this.

Zoll (1994) notes that the output [tˢaːwˀaː] violates a base-affix alignment constraint by one segment, because the left edge of the suffix coincides with the left edge of the segment [wˀ], and the right edge of the base [tˢaːw] coincides with the right edge of that segment. In order to satisfy Align, the result should have been [tˢawʔaː], with a separate glottal-stop segment; but this would force shortening of the long vowel /aː/ in the base to [a]. Apparently, the constraint Transmit (timing), or, more precisely, Parse (μ), dominates Align. In the following tableau, I have translated this idea into the current framework (with some undominated syllable-structure licensing constraints):

[2] Also known as Yawelmani, which is a plural form denoting members of the tribe (Newman 1944:19; Zoll 1996: ch. 1: fn. 13).

[3] I use the perceptual formulation of this feature instead of the usual hybrid [constricted glottis].

/t^{s}aːw | ^{ʔ}aː/	*VVC]$_{\sigma}$	*$_{\sigma}$[CC	*DELETE (μ)	*OVERLAP (base, suffix, C)
☞ t^{s}aː.w^{ʔ}aː				*
t^{s}aːw.ʔaː	*!			
t^{s}aː.wʔaː		*!		
t^{s}aw.ʔaː			*!	

(9.60)

The above account works for all suffixes that start with a floating glottal stop. However, Yowlumne has more suffixes with latent segments, and Zoll (1994, 1996) argues that these should be treated in the same way: like /^{ʔ}aa/, the suffix /^{h}nel/ '(passive adjunctive)' does not trigger epenthesis: when suffixed to /hogon/ 'xx', it gives [hogonnel], not *[hogonihnel] or so. However, it does induce vowel shortening, suggesting the ranking of ALIGN above *DELETE (μ):

/maxaː | ^{h}nel/	*VVC]$_{\sigma}$	*$_{\sigma}$[CC	*OVERLAP (base, suffix, σ)	*DELETE (μ)
ma.xaːh.nel	*!			
ma.xaː.hnel		*!		
☞ ma.xah.nel				*
mah.xaː.nel			*!	

(9.61)

Thus, Yowlumne would be a case for *OVERLAP (base, suffix, σ) >> *DELETE (μ) >> *OVERLAP (base, suffix, C), showing that alignment can work as an intrinsically ranked family of independently interacting constraints.

For the Yowlumne facts, other analyses may be possible. Zoll (1994) did not notice the discrepancy described above, but still, her 1996 version takes care of it. The non-directionality of NO-INTERVENING solves the problem of [t^{s}aːw^{ʔ}aː]: the *right* edge of the [glottal stop] feature perfectly aligns with the right edge of the base, so NO-INTERVENING is not violated. Therefore, the homogeneity of Zoll's alignment constraint is preserved.

Some of these problems relate to the idea that the real problem with infixation is not its lack of alignment, but its violation of the integrity of the base or the affix or both. This cannot be handled by general contiguity constraints, like those proposed by McCarthy & Prince (1995), because these also militate against epenthesis of new material. Rather, a constraint like *MIX (base, affix) could rule out outputs that correspond to the underlying morphemes in affix-base-affix or base-affix-base order (or, as in [ʔel^{ʔ}kaː], base-affix-base-affix). That would be a morphological constraint, whereas *SHIFT only refers to phonological material, though its ranking could be morphologically conditioned.

9.14 Global or local ranking of faithfulness constraints?

As was the case with effort constraints, and follows from the arguments in §9.1.2 and §9.9, the perceptually motivated constraints of speech production cannot be ranked in a universal way, except for local variations. Phonology translates system-wide contrast into a system of local, manageable universal rankings and language-specific rankings of non-neighbouring constraints. In chapter 11, we will see the role of this principle in the phonologization of phonetic principles.

9.15 Conclusion

The ***faithfulness*** constraints favour the correspondence and similarity between the perceptual specification of the input to the speech-production mechanism and the perceptual result of each candidate articulatory implementation. Functionally, these constraints can be attributed to the principle of maximizing perceptual contrast: they try to bring all (often contrasting) feature specifications to the surface of the utterance. These constraints are thus perceptually based, although some of them are cast in terms that look deceptively articulatory in nature.

If underlying autosegments are freely floating objects, PARSE and FILL would be the only faithfulness constraints we need, but in reality we will also have to deal with constraints that favour the surfacing of any underlying simultaneity, precedence, and alignment.

The question of the correspondence between input and output features and their combinations is deferred to chapter 12.

10 Acoustical faithfulness

Abstract. This chapter shows how articulatory and perceptual constraints interact, centring on phonetic implementation. All the phenomena that occur with vowel reduction can be handled within our framework of strictly ranked functional constraints.

In the previous chapter, we met with some interactions between various kinds of faithfulness constraints. In this and all following chapters, we will see how faithfulness constraints interact with the articulatory constraints identified in chapter 7. After redeeming my promise of §1.3.3 in §10.1, I will show the interaction between specification, articulation, and perception in phonetic implementation, using as an example the phenomenon of vowel reduction.

10.1 Interaction between articulation and perception

In §1.3.3, I stated that the perceptual output should look *more or less* like the specification. Constraint ranking determines what is more and what is less. In the /tɛns/ example, the following interactions between articulatory and faithfulness constraints occur:

- In the output [[t^{h}ɛɛ̃n_ts]], all forward faithfulness constraints (TRANSMIT and *REPLACE) are satisfied, i.e., all specified feature values emerge in the output: /t/ → [aspirated], /ɛ/ → [voiced], /ɛ/ → [max F2], /s/ → [sibilant], etc.
- The articulatory implementation shows the minimum number of *GESTURE violations given complete forward faithfulness. The constantly spread lips involve an appreciable violation of *HOLD.
- There are no simultaneous articulatory contours, so there are no violations of *SYNC.
- The complete satisfaction of *SYNC must sometimes lead to epenthesis. The chosen order of the nasal opening gesture and coronal closing gesture gives no epenthesis, because the resulting [ɛ̃] contains no perceptual features that are not present in [ɛ] or [n] as well. The chosen order of the nasal closing gesture and the coronal medial release gesture, however, leads to epenthesis of silence and a coronal release burst. Thus, *INSERT (plosive) is violated.
- The backward path constraints *INSERT (nasal vowel) and *INSERT (aspirated mid front vowel) are violated.

The following constraint tableau evaluates some candidate implementations for /tɛns/. The candidates are shown with a microscopic transcription, which should suggest the articulatory as well as the acoustic result, and with the probable early-categorized perceptual results, which determine the faithfulness:

/tɛns/	PARSE	*SYNC	*GESTURE	*INSERT (plosive)
(a) [[thɛns]] /tɛns/		*!*	*******	
(b) [[thɛs]] /tɛs/	*!*******		****	
☞ (c) [[thɛɛ̃nts]] /tɛnts/			*******	*
(d) [[thɛɛ̃ns]] /tɛns/		*!	*******	
(e) [[thɛ͠ıs]] /tɛ͠ıs/	*!		******	

(10.1)

The candidate /tɛ͠ıs/, which violates PARSE (consonantal) and FILL (oral / nasal), is not a well-formed utterance in English, but it is the result of feature-level categorization, as assumed in this chapter. This is midway between gestalt recognition of the utterance (or segments) and grammaticization of separate acoustic cues (§11.8).

A concise justification of the specification (1.14) can now be given:

- Perceptual vowel features (as opposed to articulatory gestures) do not matter for the non-vowels (though, of course, the perception of nasality requires its own spectral features), so vowel features are not shown for /t/, /n/, and /s/. In constraint language: the perceptual distinctivity between rounded and unrounded /s/ is so small, that the relevant PARSE constraints are very low, so low that we cannot determine the underlying value, because it will always be overridden by an articulatory constraint[1]. The only way to construe a rounding value for /s/ is by noticing that an isolated /s/ is pronounced without rounding; so there may be a specification after all, but a very weak one. However, the *GESTURE (lips) constraint may be strong enough to override any rounding specification for /s/; suddenly, we cannot determine the underlying rounding value of /s/ any longer, because it would always be overridden.
- In the same way, no coronal specification is needed for /ɛ/.
- Some values can be predicted from the values of other features. For instance, the coronal burst of /t/ forces the minus value for the [nasal] feature. But this is only true if the burst is parsed. For instance, if the specified /t/ is pronounced (and heard) as [n] (probably forced by an articulatory constraint), we may not only have a violation of PARSE (plosive), but also a violation of FILL (nasal & coronal).
- The vowel /ɛ/ is specified for [+voiced], because a voiceless vowel would be unacceptable in English. This specification is redundant in the sense that all English

[1] This does not mean that the rounding of [s] cannot play a role in the recognition of /su/ versus /si/.

vowels are voiced. To capture this generalization, the lexicon might just contain the specification “vowel”, and some rules filling in the values of [sonorant] and [voiced]. However, for the determination of constraint satisfaction, we need the [+voiced] value, because a voiceless implementation of /ɛ/ is obviously unfaithful to the specification, and must, therefore, violate PARSE (voice). Our specification, therefore, is more phonetic than the minimal lexical specification. See also §13.2.

- We included a specification of [–nasal] for /ɛ/, because English vowels show up as oral, especially in isolation.

Whether /n/ and /s/ share a single coronal specification can be doubted, because of the different cues involved in /n/ and /s/, but I represented them that way in (1.14) so as not to give the impression that specifications are ‘linear’ rather than autosegmental. The question is whether there is anything against adjacent identical autosegments on specificational tiers, for instance, whether we should collapse the two [voiced] specifications for /ɛ/ and /n/. See chapter 12.

In /tɛns/, the output correspondents of the [coronal] specifications of /t/ and /n/ must be separate: although the output [ɛns] satisfies one [coronal] specification, it does violate PARSE (coronal), because the listener will not be able to link the single recognized /coronal/ to the corresponding features of both /n/ and /t/ (because of the precedence constraints of §8.12, she will probably link it with /n/). Whether the [coronal] specifications of /n/ and /s/ should also have separate correspondents is another matter: they may be part of a homorganic NC cluster singly specified for [coronal] (ch. 12).

10.1.1 Inherent conflicts

If we think of combining the articulatory and perceptual drives that build sound systems and determine phonological processes, we must conclude that not all functional principles can be honoured simultaneously.

For instance, the principles of maximizing perceptual salience and minimizing articulatory effort seem to be on especially bad terms. However, there are some utterances that combine a minimal number of articulatory contours with a maximal number of perceptual contours: the utterance [b̥ab̥a] only involves one opening and one closing gesture for each syllable without any laryngeal activity; the very young even manage to produce this utterance without lip gestures, only using their jaw muscles. Thus, this utterance involves only a minimal *GESTURE violation and no *SYNC violation at all; moreover, the labial closing gesture is ballistic, so no precision constraints are violated. The perceptual contours of [b̥ab̥a], on the other hand, are many: silence vs. loudness, voiceless vs. voiced, low vs. high first formant. This explains the preference of languages for the alternation of consonants and vowels.

As another example, the combination of maximization of salience and minimization of physical effort favours small movements that yield swift variations in perceptual parameters. This interaction predicts exactly the reverse effects from Stevens’ holistic precision criterion (§7.4), which favours large movements with slowly varying perceptions. A comprehensive theory of functional phonology will show more interesting conflicts between the various articulatory and perceptual needs.

10.1.2 No interaction constraints

Our standpoint assumes a rigorous division of labour between articulatory and faithfulness constraints, so it does not allow the use of surface-true constraints that can be reanalysed as an interaction between articulatory and perceptual needs. For instance, a constraint like “nasals assimilate in place to any following consonant” (§11.6) is not allowed in the grammar, because it should be seen as the interaction of a constraint that minimizes articulatory gestures, and a constraint that tries to preserve place contrasts (§6.2.2). The relative weakness of the latter constraint for nasals as compared with plosives, causes the surface-truth of the hybrid constraint in some languages.

10.2 Constraints for vowel height

We will now turn to the example of the reduction of the vowel /a/ in various contexts, in a language with the front vowels /a/, /e/, and /i/.

10.2.1 Specification: perceptual constraints

The perceptual specification of the vowel /a/ includes directions to make its height contrastive with that of its neighbours.

In our example, its nearest neighbour will be an /e/ with an F_1 of 500 Hz. The probability of confusing /a/ with /e/ as a function of the first formant of the realization of /a/, is roughly as shown in figure 10.1 (on the left): if the realized F_1 is 500 Hz, confusion with /e/ is complete, and confusion is much less for larger distances. Ideally, we should use a frequency scale calibrated in difference-limen units (§4.2), but if we crudely assume that we can use a linear Hz scale and that formula (4.22) for the relation between distance and confusion probability holds, the logarithm of the confusion probability for an F_1 higher than 500 Hz is a parabolic function of the distance in Hz between the F_1 of the realization of /a/ and the F_1 of the neighbouring /e/. We then specify the vowel /e/ on the F_1 tier as [500 Hz], the vowel /i/ as [“min”], and the vowel /a/ as [“max”] (i.e., minimum vowel height).

In the phonetic implementation, actual values will have to be assigned to the first formant of /a/. Because of the continuous range of F_1, the [“max”] specification will branch into an infinite number of constraints, ranked logically according to the principle that a less restrictive specification is ranked lower than a more restrictive specification (9.17): thus, for the maintainance of the height contrast it is more important for /a/ to have its F_1 at least 100 Hz away from that of its neighbour, than it is to have its F_1 at least 200 Hz away. The constraint “F_1 is maximal” will therefore be divided up into a continuously parametrized constraint family MAXIMUM (F_1, *f*), or just ($F_1 > x$), where *f* is a frequency, and a partial ranking within this family is:

$$(F_1 > 600\text{ Hz}) \gg (F_1 > 700\text{ Hz}) \gg (F_1 > 800\text{ Hz}) \qquad (10.2)$$

Instead of ranking these three arbitrary members only, we can express the logical ranking of the complete family as

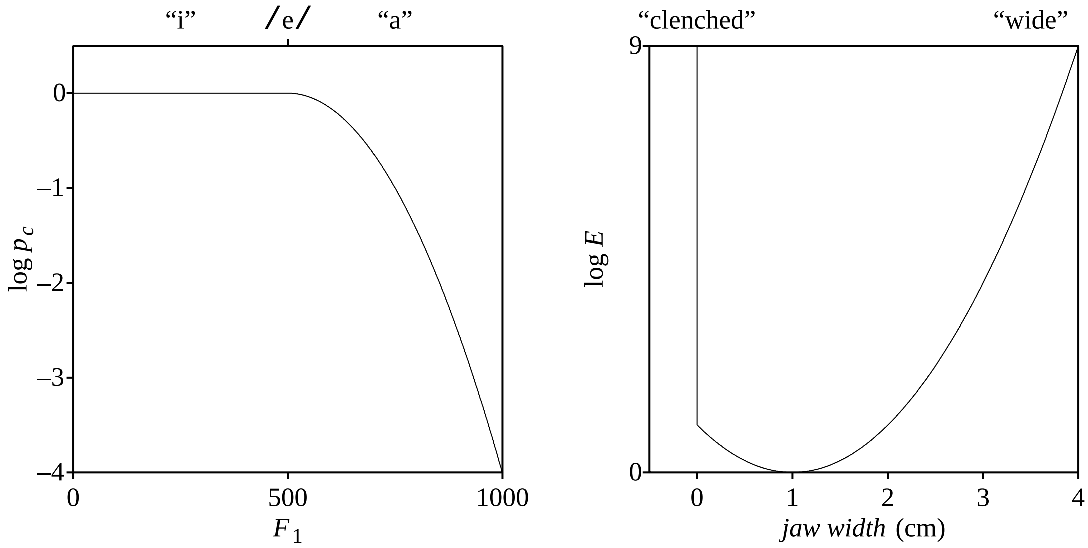


Fig. 10.1 Confusion probability of an intended /a/ as a function of first formant (left), and energy expenditure as a function of jaw width (right).

$$(F_1 > x_1 \,/\, env) >> (F_1 > x_2 \,/\, env) \Leftrightarrow x_1 < x_2 \qquad (10.3)$$

where *env* is any environment (the everything else that is kept equal). Hence, the falling slope between 500 and 1000 Hz in figure 10.1 (left-hand side) can be interpreted as the rankings of these specificational constraints along an arbitrary scale of importance.

10.2.2 Articulatory constraints

To find the actual resulting F_1 value, the MAXIMUM constraints have to be matched by articulatory constraints. A very high F_1 is difficult to produce, because of the strong jaw and tongue depression needed.

Consider first the /a/ spoken in isolation. The jaw opening, which is much wider for a typical /a/ than if all the muscles are relaxed, must be maintained by an isometric contraction of the mylohyoid and other jaw depressors (for simplicity, the tongue is ignored). According to formula (7.4), this involves more energy as the opening gets wider, because the elastic restoration forces increase. Figure 10.1 (right-hand side) shows the effort as a function of the jaw width, measured at the teeth: the resting width is 1 cm and all other widths take some amount of continued muscle activity, shown by the parabolic curve; widths below 0 cm are impossible to achieve, so the curve shoots off into space there. According to (7.11), we can translate this energy hierarchy into a *HOLD constraint hierarchy, analogously to the MAXIMUM constraint family of the previous section. This is reflected in the following formula (for openings wider than neutral), where I use the more general term *ENERGY (§5.1):

$$\text{*ENERGY (jaw opening} = x_1) >> \text{*ENERGY (jaw opening} = x_2) \Leftrightarrow x_1 > x_2 \qquad (10.4)$$

Hence, the curve in the right-hand side of figure 10.1 can be interpreted as the rankings of these articulatory constraints along an arbitrary scale of importance.

10.3 Articulation-to-perception transformation

If we know the relative heights of all the MAXIMUM and *ENERGY constraints, we can compute the resulting F_1 value if we know the relation between jaw opening and F_1. Let's assume that this relation is (figure 10.2, left-hand side):

$$F_1 = 500\ \text{Hz} \cdot \sqrt{\frac{jaw\ width}{1\ \text{cm}}} \tag{10.5}$$

Thus, with a neutral jaw width of 1 cm, the first formant is 500 Hz, and a width of 4 cm is needed to increase it to 1000 Hz. Of course, this is a gross simplification of all the factors contributing to F_1, but it expresses the idea that the more peripheral a vowel must be, the more energy must be spent to achieve the necessary vocal-tract shape.

10.4 Interaction of articulatory and perceptual constraints

The right-hand side of figure 10.2 shows the energy needed to reach a given F_1. It is computed from

$$\log E = \left(\frac{width}{1\ \text{cm}} - 1\right)^2 = \left(\left(\frac{F_1}{500\ \text{Hz}}\right)^2 - 1\right)^2 \tag{10.6}$$

where the "– 1" ensures that the effort constraint has a minimum at a width of 1 cm (all muscles relaxed), i.e. at an F_1 of 500 Hz. Now that we know both the confusion probability and the needed energy as functions of F_1, we are in a position to compare the rankings of the two constraint families. The following tableau shows four candidates for the expression of the underlying feature value [max F_1], for a certain choice for the interleaving of the constraint families ("*ENERGY (jaw opening = *x*)" abbreviated to "*E(*x*)"):

[max F_1]	*E(4cm)	F_1>600	*E(3cm)	F_1>700	*E(2cm)	F_1>800	*E(1cm)
550 Hz		*!		*		*	*
650 Hz				*!		*	*
☞ 750 Hz					*	*	*
880 Hz			*!		*		*

(10.7)

From these four candidates, the winner is 750 Hz. The first two candidates have a too low F_1, and the fourth candidate involves a too difficult gesture (jaw more than 3 cm wide).

We can represent the same intertwining of the constraint families with the two curves in figure 10.3a. As a measure of the "importance" of the specificational constraints, we take $10 + 5 \log p_c$; as a measure of the importance of the articulatory constraints we take $3 + \log E$. The two curves cross at about 760 Hz. To the left of this point, the perceptual

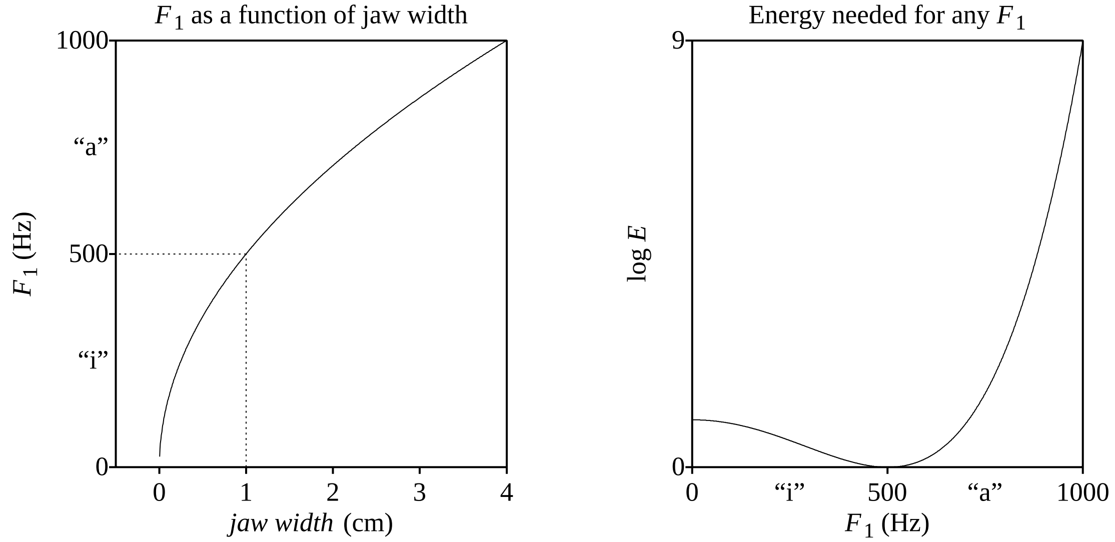


Fig. 10.2 The realized first formant as a function of the jaw width (left), and the energy needed to realize any F_1 (right).

constraint is the stronger, so that it forbids candidates with low F_1; to the right of the crossing, the articulatory constraint forbids candidates with a large jaw opening; at the crossing, both constraints are equally strong, and there must be a stable equilibrium here because we cannot optimize two interdependent quantities at a time. Thus, the OT optimization criterion is:

Minimize the maximum problematic phenomenon:

"The working point of a system of continuous constraints is located where the two strongest optimization principles pose equal problems." (10.8)

We should compare this strategy with the strategy most commonly found in the literature: that of minimizing a weighted sum over the various factors. Figure 10.3b shows the resulting curves of adding $\log E$ to $\frac{1}{2} \log p_c$, $\log p_c$, $2 \log p_c$, and $5 \log p_c$. The gross features of these functions vary wildly, and only the third function has a minimum between 500 Hz and 1000 Hz. This should be compared with figure 10.3c, where $\log E$ is *subtracted* from the four functions $5 + \log p_c$, $1 + \log p_c$, $5 + 5 \log p_c$, $1 + 5 \log p_c$, after which the absolute value is taken. Though the four zeroes appear at somewhat varying locations (600, 660, 720, and 840 Hz), they all lie well within the region of interest, i.e. between 500 and 1000 Hz.

The cause of the trouble is the fact that it is a poor optimization strategy to add a monotonically increasing function to a monotonically decreasing function; whether it shows a minimum at all strongly depends on the precise shape of these functions, as well as on the weighting factor. By contrast, the presence of a cutting point in figure 10.3a does not depend on the exact shapes of the functions, as long as these are monotonic. A comparable strategy of minimizing the maximum problem (in his case, vowel contrast) was shown by Ten Bosch (1991) to outrank Liljencrants & Lindblom's (1972) global optimization criterion for simulating vowel systems with phonetic principles; yet, Vallée (1994), in the same kind of simulations, returns to additive global optimization criteria, meticulously adapting her distance functions to the needs of stability. We must conclude, however, that the OT-compatible strategy of minimizing the largest problem is a more robust way of showing the *presence* of an equilibrium point (which is a prerequisite for finding its *position*).

10.5 Shifting the working point

We shall now turn to the environmental conditioning of the interleaving of the perceptual and articulatory constraint family, and prove that phonetic explanations can be adapted very well to an optimality-theoretic framework.

10.5.1 Dependence on stress

As usual, the ranking of the MAXIMUM constraints depends on the environment if the environment influences the distinctivity. Now, all distinctions are fainter in an unstressed than in a stressed environment (the average background noise masks more of the spectrum). This gives the functional ranking

$$(F_1 > x \; / \; +\text{stress}) \gg (F_1 > x \; / \; -\text{stress}) \qquad (10.9)$$

Thus, in unstressed position, the MAXIMUM constraints are ranked lower, and if the stressed position has its constraints ranked as in the previous tableau, the ranking in unstressed position may be as in the following tableau:

[max F_1]	*E(4cm)	*E(3cm)	F_1>600	*E(2cm)	F_1>700	*E(1cm)	F_1>800
550 Hz			*!		*	*	*
☞ 650 Hz					*	*	*
750 Hz				*!		*	*
880 Hz		*!		*		*	

(10.10)

Suddenly, the optimal candidate is only 650 Hz. The previous winner (750 Hz) now involves a jaw width (more than 2 cm) that costs too much in relation to the importance of very high F_1.

Figures 10.4a and 10.4b show curves of the constraint families in stressed and unstressed positions. Figure 10.4a is the same as 10.3a, i.e., the isolated /a/ is thought of as stressed. In the unstressed situation of figure 10.4b, the lowering of the MAXIMUM family with respect to the stressed environment causes the working point to move down to 650 Hz. In the ultimate unstressed case, the MAXIMUM curve falls entirely below the *ENERGY curve, so that the *ENERGY constraints determine the working-point all by themselves: the resulting working-point is the minimum of the *ENERGY curve, i.e. the neutral position of the jaw, and the only vowel left in the system is a vowel with an F_1 of

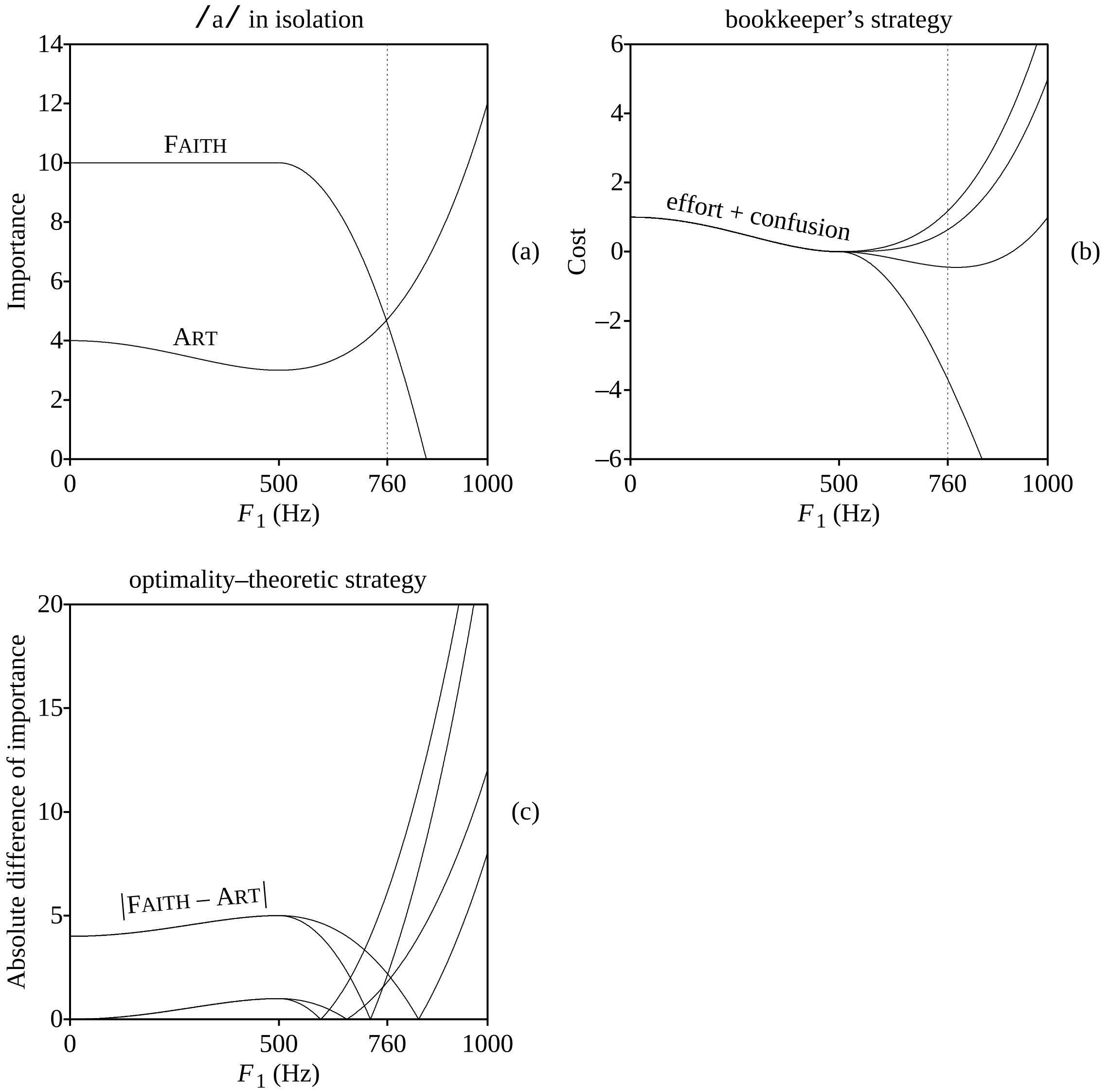


Fig. 10.3 Construction of the working point (the realized F_1) for the interacting perceptual and articulatory constraints in the phonetic implementation of /a/.

500 Hz. Here we see an example of how the weakness of a faithfulness constraint can cause a change in the language's inventory of sounds; in chapter 16, I will defend the hypothesis that the interaction between articulatory and perceptual constraints indeed determines the exact shape of every sound inventory, *including its size* (which is different from what all other phonetically-based models have done so far).

Other explanations. Another explanation for vowel reduction in unstressed syllables is that the relative shortness of these syllables (which itself may be a result of enhancing the perceptual prominence of stressed syllabels by lengthening) causes an increased effort needed to implement a given high F_1 (the *FAST constraint). This would shift the ART curve to the left, as in fig. 10.4c, also giving a lowere working point. We see that several phonetic explanations can be expressed formally in a constraint grammar, but also that the formalism cannot automatically decide which explanation is correct.

Two cutting points? In figure 10.4b, we can see that if the ranking of the faithfulness constraint decreases by another 2 points along the scale, there will be *two* cutting points instead of one, and one of them will be below 500 Hz. The *working point*, however, will still be the cutting point above 500 Hz, because it is lower and, therefore, satisfies the imperative of minimzing the maximum problem.

10.5.2 Dependence on surrounding consonants

Very probably, the energy, and thereby the ranking of the separate constraints of this family, does not depend on stress. The energy does depend, however, on the position of the articulators before and after the vowel. A given jaw opening is easier to achieve before the isolated [a] than in the utterance [pap], which involves two lip closures that can only be brought about with the help of a closing jaw. According to equation (5.4), the movement costs more energy as the distance to travel is larger, either because of the extra duration of the gesture, or because of the higher velocity and acceleration.

*ENERGY (jaw=*x* / [pap]) >> *ENERGY (jaw=*x* / [pa]) >>
>> *ENERGY (jaw=*x* / [a]) (10.11)

The constraint *ENERGY (jaw=*x* / [ap]) also belongs between the highest and lowest constraints in this formula, but can be ranked a priori with *ENERGY (jaw=*x* / [pa]) only if we can find a way of locally comparing [ap] and [pa] (i.e., seeing them as differing in one respect only), presumably by an argument involving time asymmetry.

If we want to know the resulting F_1, we can make a tableau like the previous one. Instead of weakening MAXIMUM constraints, we now see strengthening *ENERGY constraints, but this produces the same kind of shift of these families with respect to each other. Again, therefore, the resulting F_1 will be lower in [pap] than in the ideal isolated [a]. This can also be seen in figure 10.4c: the zero-energy position of the jaw is more closed than in the isolated “environment”, so the *ENERGY constraint curve moves to the left with respect to figure 10.4a, which results in a lower working-point.

10.5.3 Dependence on duration

A fast movement takes more energy than a slow movement. According to equation (7.4), if a given trajectory in space must be walked twice as fast, the double velocity combines with the double acceleration to give a fourfold increased power expenditure. Because the gesture is finished in half time, this leaves us with a doubled energy cost:

*ENERGY (jaw opening = *x* / –long) >> *ENERGY (jaw opening = *x* / +long)
(10.12)

Along the lines of the previous sections, this will mean that the resulting F_1 is lower for short vowels than for long vowels. If we assume that the isolated /a/ was long, figure 10.4c shows the displacement of the *ENERGY curve with respect to the curve of figure 10.4a, which again results in a lower working-point.

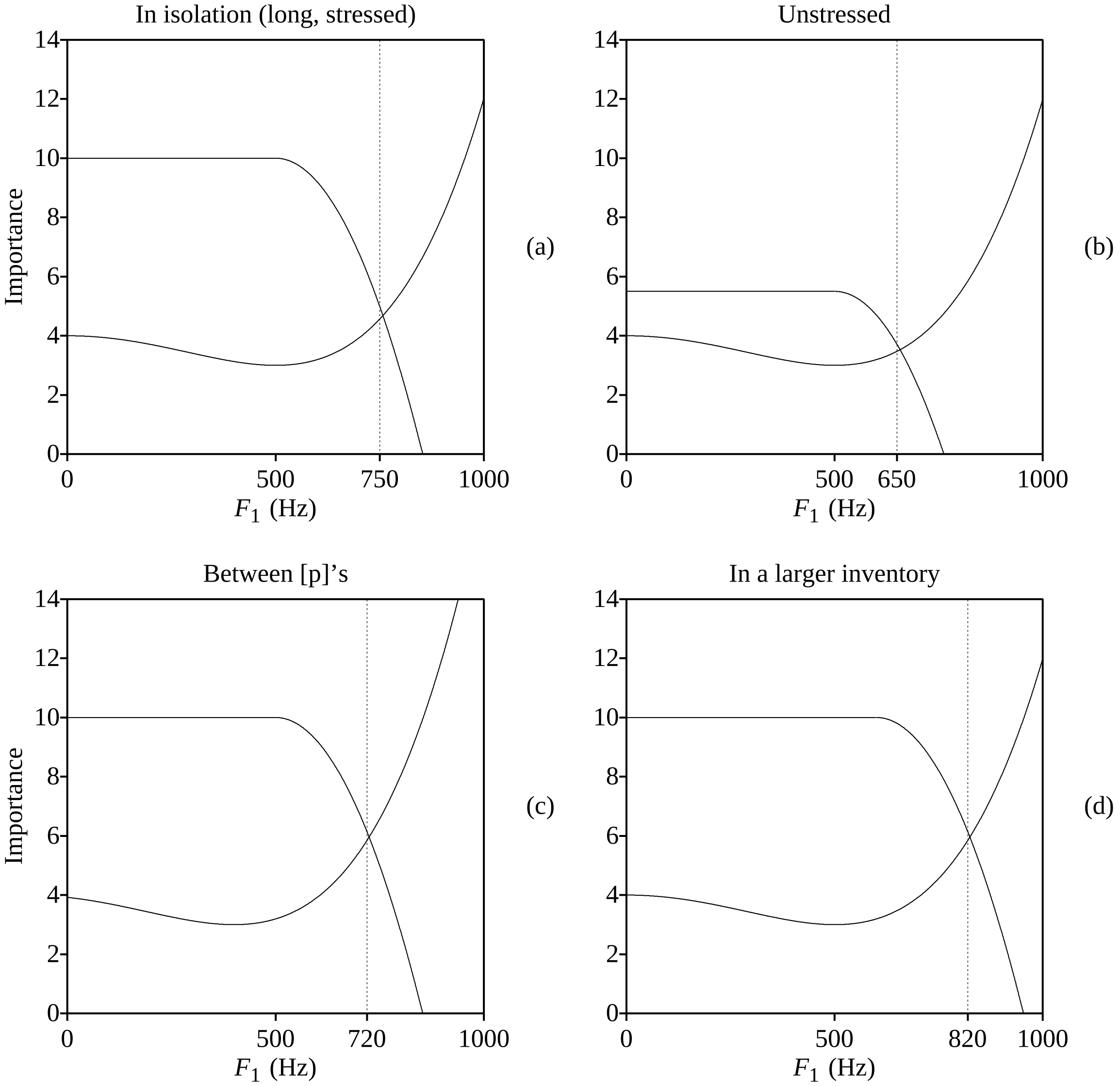


Fig. 10.4 The influence of various environments on the working-point in the interaction between a perceptual and an articulatory constraint.

10.5.4 Dependence on inventory size

Above, we considered a front-unrounded vowel system consisting of /a/, /i/, and only one mid vowel with an F_1 of about 500 Hz. Now imagine that we have two mid vowels instead of one. Their F_1 values are likely to be around 400 and 600 Hz. The nearest neighbour to /a/ is now the higher mid vowel with an F_1 of 600 Hz. This means that the MAXIMUM curve of figure 10.4a should now by centred around 600 Hz. This is shown in figure 10.4d. The 100-Hz change in the formant of the nearest neighbour causes the working point to move up by 60 Hz. The working-point does not move by 100 Hz, because the *ENERGY curve is not horizontal; thus, though the preferred F_1 of /a/ rises, the distance to its nearest neighbour decreases by 40 Hz.

10.5.5 Comparison to other models

Because of the comprehensive nature of our formalism, the account of vowel reduction presented here is in accordance with almost every theory about it. From the presentation above, we can conclude that the shorter an open vowel is, the lower its F_1 will be; this is in line with Lindblom's (1963, 1990b) target undershoot model. Note that what happens here is not ***centralization***, but ***coarticulation***: the vowel triangle gets smaller because the low vowels *rise* in the direction of their neighbouring consonants; for low vowels, this is the same as centralization, but there is no articulatory or perceptual gain in centralizing *high* vowels in unstressed or shortened environments. This is in accord with the findings of Van Bergem (1995), who showed that Dutch high vowels do not centralize in these positions.

But we must also conclude that vowel reduction in unstressed syllables is caused by two phenomena: first, because of the lower intensity the contrasts are smaller, so that it becomes less important to maintain them (§9.9); secondly, because of their unimportance, unstressed syllables will be shorter than stressed syllables, and this will reduce the vowels further because of the extra energy that would be needed to bring them to their 'long' position. As usual, a comprehensive optimality-theoretic account proves capable of reconciling the articulatory and perceptual viewpoints.

Two other vowel-reduction ideas should be noted here. Van Son (1993) showed that in rapid speech, a professional radio announcer was able to compensate for the shorter vowel durations by raising the velocity of the articulators in such a way that the same formant values were reached as in slow speech. A tentative explanation is that there are few situations where faithfulness is ranked higher than in the case of a speaker whose living depends on being clearly understood by a million people at the same time.

The other idea is that not the isolated long stressed vowel, but a short vowel in an unstressed environment might be the 'target' defined in our lexicon for that vowel, and that the clear stressed varieties are actually perceptual and/or articulatory enhancements over these moderately contrastive vowel targets (Koopmans-van Beinum 1980). Now, in the account presented here, the question is whether we need 'targets' at all: none of the four situations depicted in figure 10.4 was granted the status of a 'target', and, in fact, the concept is meaningless in the context of interacting continuous constraint families.

Finally, we must note that listeners can compensate for the variation that results from the constraint interactions in various environments. For instance, so-called 'target undershoot' can be compensated for by a mechanism of 'perceptual overshoot' (Lindblom & Studdert-Kennedy 1967). For understanding the structure of sound systems, the existence of these mechanisms helps explain why listeners are so resilient that speakers may let their faithfulness constraints be dominated by so many articulatory constraints that phonology stays such an interesting subject.

10.5.6 Lexical vowel reduction

Van Bergem (1995) distinguishes between ***acoustic*** vowel reduction, which is the phenomenon described in the previous sections, and ***lexical*** vowel reduction, which is the replacement of a full vowel by a schwa in the lexicon.

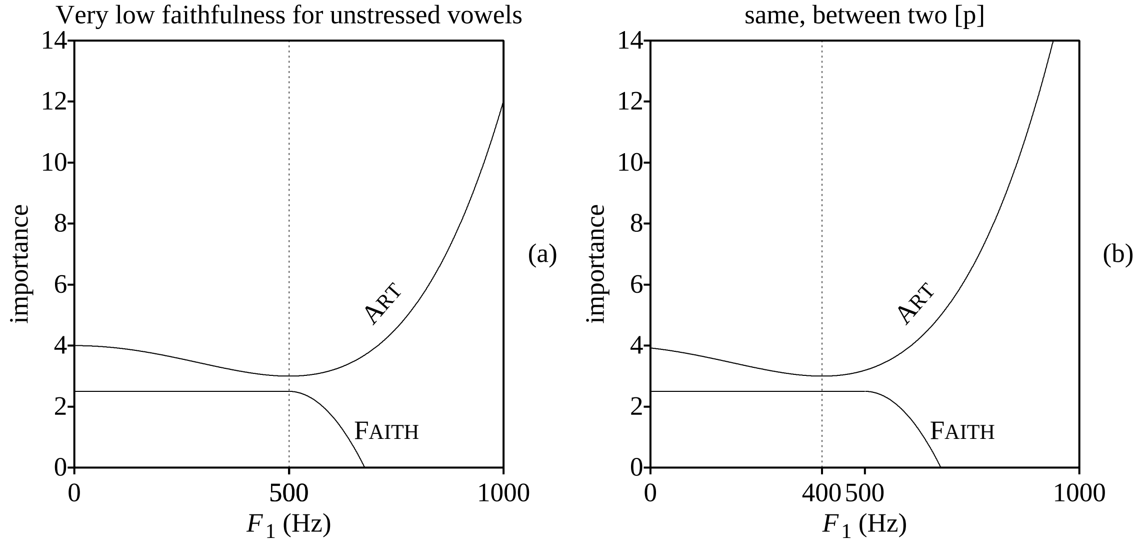


Fig. 10.5 If faithfulness is very weak, the outcome is solely determined by articulatory effort. The resulting F_1 is found where the effort is minimal.

We can explain lexical vowel reduction in the following way. The relative importance of articulatory effort and perceptual confusion is language-specific. In some language, therefore, the acoustic faithfulness constraints in unstressed syllables may be lower than in figure 10.4b; they could even be as low as in figure 10.5a. In this case, effort is always stronger than confusion, so that the result is determined by least effort: the speech-neutral position of the jaw, leading to a F_1 of 500 Hz, and an acoustic result of [ə], independently of the exact shape of the faithfulness curve, i.e. independently of the specified height of the vowel. Thus, this situation signifies the merger of all vowels in unstressed position.

The actual least-effort position of the jaw and other articulators will depend on the neighbouring consonants, so that the actual realization of the reduced vowel varies according to the phonetic context. For instance, in [pəp], F_1 will be lower than 500 Hz, as we can tell from figure 10.5b, where the curve of the effort constraint was copied from figure 10.4c. In this case, too, the underlying vowel quality has no influence at all on the resulting vowel.

Not that this sudden transition from a continuous reduction to a discrete reduction can only occur in an OT optimization scheme, not in a utilitarian scheme, which would minimize the sum of the effort and confusion costs.

10.6 Typologies of vowel systems

The argument can be extended for peripheral vowels other than /a/. Peripheral front vowels are specified for maximum F_2, given their values of F_1. A high F_2 (with constant F_1) is achieved by a combination of wide pharynx (so that the tongue body does not have to be constricted too much), strongly bulging tongue, and strong lip spreading; relaxing these conditions in any way will lower the F_2. Peripheral back vowels are specified for minimum F_2, which, with constant F_1, is achieved by strong lip rounding and a back closure, the location of which depends on F_1. Any more neutral vocal-tract shape would give a higher F_2. So we see that the peripherality of front unrounded and back rounded vowels is subject to the same mechanisms as the lowness of /a/.

We have thus found formal functional explanations for the following facts of language:

- Vowels are less peripheral in unstressed than in stressed position (for Dutch: Koopmans-Van Beinum 1980). In Italian, the system of unstressed vowels is usually described with merged |ɛ| - |e| and |ɔ| - |o|, but |a| in unstressed position is pronounced at the height of stressed |ɛ| and |ɔ| (Leoni, Cutugno & Savy 1995).
- Vowels are more peripheral when spoken in isolation than when embedded in an utterance (for Dutch: Koopmans-Van Beinum 1980).
- Long vowels are more peripheral than short vowels. Even some languages whose vowel systems can be described as sets of long-short pairs, often have less peripheral short vowels. Latin and Czech, for instance, have lowered short |i| and |u|.
- The vowel triangle is larger for large inventories than for small ones. This is somewhat hard to prove, because three-vowel systems tend to be reported as { "a", "i", "u" }. Nevertheless, if we count [ɪ] and [ʊ] as high vowels, 2 of the 18 three-vowel systems in Maddieson (1984) are reported as having no high vowels (Alabama and Amuesha), whereas only 1 of the 299 larger systems lack high vowels (Squamish). Similarly, if we count [ɐ] and [æ] as low vowels, 1 of Maddieson's 18 three-vowel systems is reported as lacking low vowels (Tagalog), as is 1 of the 299 larger systems. If we count [ɪ] and [ʊ] as non-high, and [ɐ] and [æ] as non-low, 5 of the 18 three-vowel systems lack high or low vowels or both, as do only 18 of the 299 larger systems; these 18 include a few systems whose larger number of vowels comes from including some very rare vowels (Alawa) or some vowels that only occur in assimilated loans (Quechua).
- In a large inventory, vowels are closer together than in a small inventory. This is a trivial observation in theories that build vowel systems around the "features" or "particles" **a**, **i,** and **u**, like Dependency Phonology (Anderson & Ewen 1997) and Particle Phonology (Schane 1995), but in a theory that refuses to accept innate feature values, it has to be explained separately (§10.5.4).

10.7 Conclusion

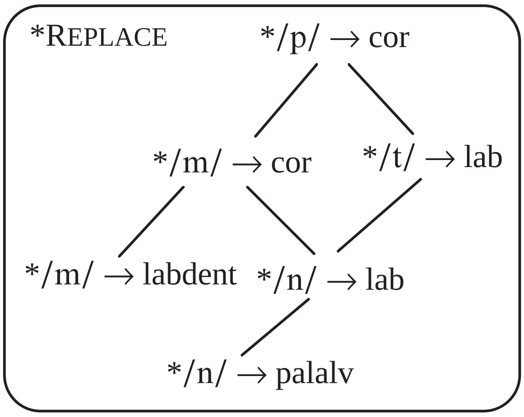
The model can be extended to other cases, most notably the interaction between *PRECISION and *SYNC. Like the acquisition of coordination facilitates recurrent use of combinations of articulatory gestures, the acquisition of categorization facilitates recognition of the discrete elements that make up the utterance, and is translated into a reranking of the *PRECISION and *SYNC constraints by changing the boundaries between which the articulations are constrained in order to produce a reproducible percept. Another field where the balancing model will lead to an optimum is the interaction between the informational constraint (maximum entropy, §9.6) on the one hand, and minimization of effort and categorization on the other.

We have thus seen that phonetic explanations can well be formulated with constraint ranking; that the constraint formalism allows robust identification of working points, unlike cost-minimization strategies; and that the transition between continuous and discrete phenomena is expected and natural.

11 Typology and the local-ranking hypothesis

Abstract. The language-specific freedom of constraint ranking is determined by the local-ranking principle, which states that constraints that differ in a single parameter may be universally ranked with respect to one another. From all the grammars allowed by the local-ranking principle, languages tend to choose a grammar in which many constraints can be generalized over their arguments or environments.

To get a detailed example of a near-universal hierarchy of place faithfulness, we can combine (9.22), (9.39), and (9.44) into the following partial grammar:



(11.1)

The lines in this figure connect pairs that vary along a single perceptual dimension (place), or that vary minimally in their environment (plosive/nasal), or that vary minimally in their degree of specificity. These minimally different pairs could be locally ranked according to universal principles of commonness (§9.6), environment-dependent contrast (§9.9), or the distinction between essentials and side-issues (§9.10).

As already touched upon in §7.6, §9.1.2, §9.9, and §9.14, the remaining pairs in (11.1) cannot be ranked locally in this way, and we will propose that speakers and listeners cannot rank them in this way either. This leads to the hypothesis that phonology can rank but not count, or, more accurately:

Local-ranking principle (LRP):

“Universal rankings are possible only within a single constraint family, for the same feature or gesture, for the same sign of the articulatory or perceptual deviation”, e.g. (7.7), (7.11), (7.12), (7.15), (7.26), (7.33), (8.5), (9.7), (9.12), (9.18), (9.45), (9.57), (10.2). (11.2a)

“A near-universal ranking is possible only between a pair of constraints whose arguments or environments differ minimally”, e.g., (7.19) (for lip vs. blade), (9.23), (9.29), (9.32), (9.42), (10.8). (11.2b)

“Cross-linguistic variation is expected for other pairs, though *tendencies* are expected for rankings based on global measures of effort or contrast, and the strength of the tendency depends on the difference between the two global measures”, e.g. (7.3), (7.19) (for blade vs. velum), (4.18), (9.3), (9.38), (9.41). (11.2c)

Of course, the transitivity of the strict-ranking scheme causes such rankings as *REPLACE (place: bilabial, coronal / plosive) >> *REPLACE (place: bilabial, labiodental / nasal) to be near-universal, too.

The LRP is a special case of Ladefoged’s (1990) statement: “there is no linguistically useful notion of auditory distinctiveness or articulatory economy in absolute terms”. Instead of going along with Ladefoged’s pessimistic view of the possibility of doing anything interesting with these principles in phonology, we can be glad that the LRP allows the linguist in her quest for universals to restrict herself to local, more manageable, variations, instead of tediously trying to measure the ingredients of equations (4.24) and (7.4).

11.1 Freedom of ranking

By itself, nearly every constraint can be ranked very low in one language, and very high in the other.

After the speakers of a language have learned a gesture, the corresponding *GESTURE (*gesture*) constraint is often very low; for other languages, however, it may still be undominated. For instance, a language typically has no apico-palatal closures at all, or it has a more or less complete set like /ɭ/, /ɳ/, /ɗ/, and /ʈ/.

The same is true of the *COORD families. Consider, for instance, the “anterior”-dorsal coordination found in oral suction consonants: a language typically has no click consonants at all, or it has a complete set with three, four, or five anterior releases, one or two dorsal closures, and several ***manners*** chosen from voiceless, voiced, aspirated, nasal, prenasalized, and glottalized.

The same, again, is true of *CATEG constraints: every language makes its own choice of subdividing the continuous parameter of vowel height or the continuous parameter of the voice-onset time of plosives.

Thus, the height of many *GESTURE, *COORD, and *CATEG constraints varies cross-linguistically from maximally high to maximally low. Universal notions of “easy” and “difficult” gestures and coordinations do not play any role in the description of any particular language. At best, these notions could explain statistical tendencies such as the relatively modest rate of occurrence of apico-palatal gestures and velaric ingressive coordinations when compared with, say, apico-alveolar gestures and labial-velar approximants.

We have seen that the possibility of universal ranking within a family is subject to the condition of *ceteris paribus* (“if everything else stays equal”): we can only impose an a-priori ranking on constraint pairs that differ minimally. There is no simple way in which we could predict the universal ranking of the labiality of /m/ and the coronality of /t/. The local-ranking principle proposes that there *is* no such universal ranking; this would mean that we expect that some languages rank the labial parsing constraints as a group above the coronal parsing constraints, and others rank the parsing constraints for plosives as a group above those for nasals:

Typological prediction of the local-ranking principle:

> "Languages can freely rank any pair of constraints that cannot be ranked by the LRP (11.2ab) directly or by transitivity." (11.3)

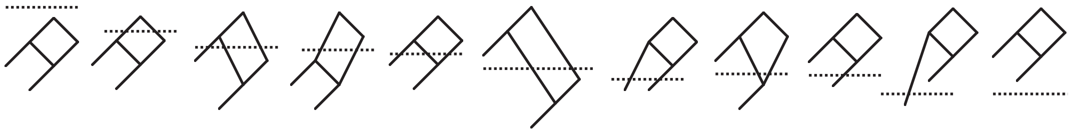
Stated as bluntly as this, (11.3) is too strong; after all, most people would agree that competitive skating is more difficult than riding a bike slowly, and that a horse is more different from a duck than an apple is from a pear. Thus, very large differences of effort and contrast will still be visible in the typology of languages (11.2c). We predict that *only* very large differences of effort and contrast will be visible in the ranking of non-minimally different pairs of constraints.

So it seems that we need only look at the local (one-feature) variation to predict universal or near-universal ranking, and that many of the more distant constraint pairs must be ranked in the grammar of each language. Restricting ourselves to these relative terms relieves us of the task of finding global measures of effort or distinctivity: if languages do not care, why should the linguist?

11.2 Combinatorial typology

Prince & Smolensky's (1993) view of the freedom of ranking goes by the name of ***factorial typology***: if there are four constraints, these can be ranked in 4! (*four-factorial*) = 24 ways. The local-ranking principle, however, restricts the freedom of ranking. If we have two families of three constraints, and the constraints within these families can be ranked according to universal principles, the rankings of each set of three constraints is fixed. The number of possible rankings should then be divided by $2!\cdot 2! = 4$, leaving six ways in which languages are allowed to rank them. In general, with two families of *m* and *n* constraints, we have $\binom{m+n}{m}$ possible rankings: the number of ***combinations*** of *m* elements within a set of $m + n$.

The typical way to test the ranking of *Replace constraints is to split up the family by using a homogeneous *Gesture constraint: all faithfulness constraints ranked above it will be satisfied; those below may be violated. Random variation in the ranking of this *Gesture constraint determines the number of possible languages. For our ranking (11.1), we get 11 possibilities (the homogeneous *Gesture is shown as a dotted line):



(11.4)

For consonants in onset position, the rightmost of this figure usually holds: all place contrasts surface. For consonants in coda position before another consonant, the Parse constraints are ranked lower, and place assimilation may result. The leftmost of these figures depicts the situation in which all coda consonants assimilate to any following consonant.

11.3 Implicational universals

The connections in (11.1) allow us to state the following implicational universals for the assimilation of place (also put forward by Mohanan 1993):

- If plosives assimilate, so do nasals (at the same place). (11.5)
- If labials assimilate, so do coronals (with the same manner). (11.6)

The fact that there is no connection in (11.1) between */m/ → [coronal] and */t/ → [labial], means that (11.5) and (11.6) are ***independent*** of each other: there will be languages where nasals assimilate, but plosives do not, and there will be languages where coronals assimilate, and labials do not, and the possibility of the inclusion of any language in the first group is independent from the possibility of its inclusion in the second group, as shown in §11.2. Thus, we have the following corollary:

Independence of implicational universals:

> "The local-ranking principle ensures that two implicational universals, if not transitively related, are independent of each other." (11.7)

The reverse is also true. Independence of the two implicational universals (11.5) and (11.6) gives the diamond-shaped part of (11.1), not two independent pairs of constraints. Thus, if we try to translate (11.5) and (11.6) into the two independent rankings *REPLACE (place / plosive) >> *REPLACE (place / nasal) and *REPLACE (place: labial, coronal) >> *REPLACE (place: coronal, labial), there would be no ranking that represents a language where all coronals and all nasals assimilate (so that only /p/ surfaces faithfully), in contrast with our prediction of (11.4) and with the universals of (11.5) and (11.6).

11.4 Case: place assimilation of nasal stops

We expect that *REPLACE (place: coronal, labial / plosive) and *REPLACE (place: labial, coronal / nasal), shown in (11.1), can be ranked in either way, depending on the language. That this accurately represents the situation in the languages of the world, will be illustrated with data on place assimilation of nasals in Dutch and Catalan.

In Dutch, nasal consonants at the end of a word have the tendency to change their place of articulation to that of an immediately following consonant. However, this tendency is not the same for all three nasal consonants (/n/, /m/, /ŋ/). The velar nasal /ŋ/ is always realized as a velar, irrespective of the place of the following consonant:

/dɪŋ/ 'thing' + /pɑkə/ 'take' → /dɪŋpɑkə/ 'take thing'
/dɪŋ/ 'thing' + /trɛkə/ 'pull' → /dɪŋtrɛkə/ 'pull thing'
/dɪŋ/ 'thing' + /kɛɪkə/ 'watch' → /dɪŋkɛɪkə/ 'watch thing' (11.8)

The alveolar nasal /n/ takes on the place of any following consonant, which can be velar, uvular, bilabial, labiodental, or palatalized alveolar:

/aːn/ 'on, at' + /pɑkə/ 'take' → /aːmpɑkə/ 'take on'
/aːn/ 'on, at' + /vɑlə/ 'fall' → /aːɱvɑlə/ 'attack'

/aːn/ ‘on, at’ + /trɛkə/ ‘pull’ → /aːntrɛkə/ ‘attract’
/aːn/ ‘on, at’ + /kɛɪkə/ ‘watch’ → /aːŋkɛɪkə/ ‘look at’
/aːn/ ‘on, at’ + /ʀaːdə/ ‘guess’ → /aːɴʀaːdə/ ‘advise’ (11.9)

The bilabial nasal /m/ is always realized as a labial, but may surface as labiodental before labiodental consonants:

/ʊm/ ‘about’ + /poːtə/ ‘plant’ → /ʊmpoːtə/ ‘transplant’
/ʊm/ ‘about’ + /vɑlə/ ‘fall’ → /ʊɱvɑlə/ ‘fall over’
/ʊm/ ‘about’ + /trɛkə/ ‘pull’ → /ʊmtrɛkə/ ‘pull down’
/ʊm/ ‘about’ + /kɛɪkə/ ‘watch’ → /ʊmkɛɪkə/ ‘look round’
/ʊm/ ‘about’ + /ʀɛɪə/ ‘drive’ → /ʊmʀɛɪə/ ‘make a detour’ (11.10)

This situation could be captured by the following naive superficial constraint system (from high to low):

(a) PARSE (dorsal), PARSE (labial), PARSE (nasal)
(b) NC-HOMORGANIC: “A sequence of nasal plus consonant is homorganic”
(c) PARSE (bilabial)
(d) PARSE (coronal) (11.11)

For instance, we see that the sequence /m + k/ must surface as [mk], because that only violates constraint (b), whereas [ŋk] would violate the higher-ranked constraint (a):

/m+k/	PARSE (labial)	NC-HOMORGANIC	PARSE (bilabial)
☞ mk		*	
ŋk	*!		*
nk	*!		*
ɱk		*	*!

(11.12)

On the other hand, /m + f/ must surface as [ɱf], as the highest violated constraint in this case is (d), whereas [ɱf] would violate constraint (b):[1]

/m+f/	PARSE (labial)	NC-HOMORGANIC	PARSE (bilabial)
mf		*!	
☞ ɱf			*

(11.13)

[1] Because of the hybrid formulation, which bypasses the OCP for PARSE constraints, /m/ → [labial] is not violated. See chapter 12.

11.5 Optionality

Language variation can simply be viewed as a variation in the ranking of constraints. For instance, for those speakers whose /m/ is always bilabial, constraint (c) ranks higher than constraint (b). But reranking is possible within a single grammar, too. Native speakers of Dutch often object to the reality of the constraint hierarchy that I showed above for the place assimilation of nasal consonants. Beside the fact that many people maintain that they always pronounce /m/ as a bilabial (and some of them actually do), people express considerable disbelief about the whole theory because “all those assimilation rules are optional”; they state that if they want to speak clearly, there need not be any place assimilation at all. Some opponents restrict their objections to the assimilation of /m/.

They are right of course. If your utterance is not understood at the first try, the importance of perceptual contrast rises with respect to the importance of articulatory effort, and you may repeat your utterance with fewer assimilations and more “parsed” features. In terms of constraint ranking, this means that perceptual constraints rise with respect to articulatory constraints. From the Dutch data, for instance, it seems warranted to state that “homorganic nasal plus consonant” first falls prey to “PARSE bilabial” (people start out saying [ʊmvɑlə] for /ʊm + vɑlə/), and that “PARSE coronal” only wins in situations where separate syllables are “spelled out” ([ɪnvɑlə] instead of [ɪɱvɑlə] for /ɪn + vɑlə/). This stylistic variation is the reason why we can rank “PARSE bilabial” above “PARSE coronal”, although the two can never be in conflict. The strength of the objections to the assimilation of /m/, expressed by some people, can now be seen, not as an overreaction to a mild constraint reranking, but as a defence against the shattering of the illusion of the discrete inviolability of the “PARSE bilabial” constraint.

On the other hand, we could also imagine that there are situations (highly predictable words; singing without the need to be understood) where articulatory constraints may rise with respect to perceptual constraints. In our example, we could expect that the first thing to happen is that the velar nasal assimilates to a following uvular consonant (*onraad* vs. *vangrail*).

11.6 Problems with surface constraints

Most languages do not exhibit the combination of assimilation of /n/ and faithful parsing of /m/. But Catalan (Recasens 1991) and Dutch do. Instead of a cross-linguistically optional assimilation rule, we have a structural constraint, whose ranking determines whether we see the phenomenon: in Dutch and Catalan, it is ranked higher than in the non-assimilating languages (Limburgian), but lower than in the fully assimilating languages, like Malayalam (Mohanan 1986). Cross-linguistic optionality is thus automatically caused by the ranking of the constraints, and not an isolated coincidence.

A problem arises when we extend our example to clusters of plosive plus consonant. In Dutch, these clusters are not subject to the same assimilations as clusters of nasal plus consonant. For instance, though /n + x/ combines to /ŋx/, not /nx/, its counterpart /t + x/ is rendered as /tx/, not /kx/. The only assimilation that targets a plosive seems to be that of the alveolar plosive /t/ to a following /j/, /ʃ^j/ or /t̠^{j}/, which turn it into a palatalized palatoalveolar plosive, which I take to be coronal but not alveolar (of course, the nasal /n/ also assimilates to these three sounds).

We could encompass all stops (nasals and plosives) in a single superficial grammar:

(a) PARSE (dorsal), PARSE (labial), PARSE (nasal)
(b) NC-HOMORGANIC
(c) PARSE (bilabial)
(d) PARSE (coronal)
(e) "A sequence of plosive and consonant is homorganic"
(f) PARSE (alveolar) (11.14)

In terms of functional principles, this is clearly wrong. NC-HOMORGANIC is an ad-hoc constraint, the result of a confusion of articulatory and perceptual constraints (§10.1.2); as such, it is found in the generative literature. For instance, Lombardi (1995) states: "in a language like Diola the constraint causing nasals to assimilate is high ranked, but whatever could cause other consonants to assimilate is low ranked". What the *whatever* is, makes a large difference in explanation. Making the wrong choice here will eventually have repercussions throughout our theory of grammar.

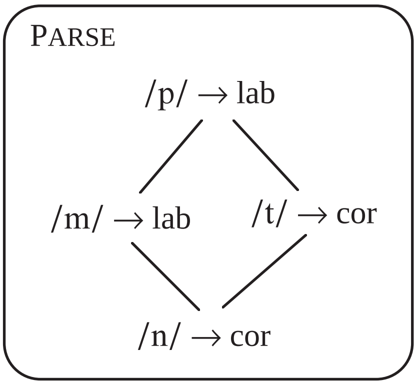
The articulatory gain of the homorganicity of plosive plus consonant must actually be *equal* to the gain of the homorganicity of nasal plus consonant, since it involves exactly the same articulatory phenomena: spreading of a place feature, and deletion of another. It is not the articulatory constraints, but the faithfulness constraints that are ranked differently. So, PARSE (coronal) is more important for plosives than for nasals, because its violation spans a larger contrast for plosives than for nasals. Therefore, the correct ranking is something like (assuming equal articulatory effort for the various oral closing gestures):

(a) /ŋ, k/ → [dorsal], /m, p/ → [labial], /t/ → [coronal]
(b) /p/ → [bilabial]
(c) *GESTURE (tongue tip), *GESTURE (upper lip), *GESTURE (back of tongue)
(d) /m/ → [bilabial], /t/ → [alveolar]
(e) /n/ → [coronal], /n/ → [alveolar] (11.15)

This ranking is not only in accordance with the data (it shares that with (11.14)), but it is also in agreement with the ranking (11.1), which was derived from functional principles.

11.7 Typology of place assimilation of nasals

The constraint ranking found in (11.15) contains some universal rankings, shown in this figure, which abstracts away from the second argument of *REPLACE:



(11.16)

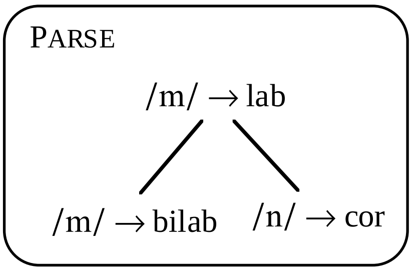
The solid lines in this figure reflect the universal ranking of place-parsing constraints for plosives above those for nasals, and the almost universal ranking of the parsing of labial features above coronal features. Depending on the ranking of the *GESTURE constraints, this predicts the following possible place-assimilation systems:

- Nothing assimilates (Limburgian).
- Only coronal nasals assimilate (Dutch).
- All coronals assimilate, but labials do not (English).
- All nasals assimilate, but plosives do not (Malayalam).
- All nasals and all coronals assimilate (no language known to me).
- Everything assimilates. (11.17)

These are exactly the six that can be expected with a “combinatorial typology”. In those exceptional languages where the dorsal articulator is as commonly used for stops as the coronal articulator, we may find that PARSE (labial) >> PARSE (dorsal) also holds: in Tagalog, for instance, /ŋ/ will often assimilate (though not as often as /n/), and /m/ will not (Schachter & Otanes 1972); this seems to be a counterexample to Jun’s (1995) cautious suggestion that “if velars are targets of place assimilation, so are labials”. Note that with the separate rankings PARSE (lab) >> PARSE (cor) and PARSE (place / plosive) >> PARSE (place / nasal), as proposed by Jun (1995), the possible absence of languages in the fifth category in (11.17) could be explained (see also §11.3). If such languages (where only /p/ and /k/ surface faithfully) do exist, the dependence of contrast on the environment should generally be included in the environment clause of the constraint, as implicit in (11.1); in such a case, influences of the environment are ***additive***, and not subject to strict ranking: they *add* to the ranking of the faithfulness constraint; and implicational universals (11.5-6) would respect this addivity. If, however, such languages do not exist, this would be evidence of a tendency of constraint simplification, and Jun’s rankings would be typologically correct.[2]

For the finer place structure of nasals, we have the following universal ranking, simplified from (11.1):

[2] The current text of this paragraph corrects the earlier version in Boersma (1997a), where I hallucinated that Jun’s rankings would fail to predict the occurrence of the Dutch situation.



(11.18)

Again, the two subordinate specifications are not neighbours (i.e., they differ in more than one respect), and can, therefore, be ranked freely. This gives the following typology for assimilation of nasals to a following labiodental consonant:

- Nothing assimilates.
- Only /m/ assimilates: Central Catalan (Recasens 1991: 252, 256).
- Only /n/ assimilates: many speakers of Dutch.
- Both /m/ and /n/ assimilate: Mallorca Catalan and the other speakers of Dutch.
- Everything assimilates. (11.19)

Thus, we see that the only freely rankable pair of constraints (/m/ → [bilab] and /n/ → [cor]) can be shown to be actually ranked differently in a variety of Catalan and a variety of Dutch.

11.8 Perceptual versus acoustic faithfulness

As we will see in almost every example, the ranking of PARSE is usually determined by its environment. For the assimilation example /atpa/ → [[ap˺_ːpa]], there are two possibilities:

1. It violates PARSE (coronal / _C) or, in a loose declarative way, /t/ → coronal / _ C. This is the approach found in the present book.
2. It violates PARSE (t˺) or *REPLACE (t˺, p˺). The first of these is analogous to Jun's (1995) account of place assimilation. There is an obvious problem in the autonomous ranking of separate place cues: because of the strict-ranking principle of OT, the cues do not additively contribute to the perception of place. I cannot tell whether this is Jun's intention; his use of an environment-conditioned constraint translatable as PARSE (place / onset) suggests that it is not.

The choice between the two approaches may be helped with the following argument: faithfulness is, in the end, a relation between specification and *perception*, not between specification and *acoustics*. Therefore, the effects of categorization should be taken into account. Now, if we accept that [coronal] is a perceptual category, and [t˺] is only an acoustic cue (see §12.3), and if we believe that strict ranking is the way that our grammar works, we must conclude that the grammar contains strictly rankable faithfulness constraints for [coronal], and that there is no evidence for such constraints for [t˺]. If we exclude constraints like PARSE (t˺) from the grammar, the possibility of additive contribution of acoustic cues to perceptual categorization is preserved (analogously to the aspects of *ENERGY, see §7.1).

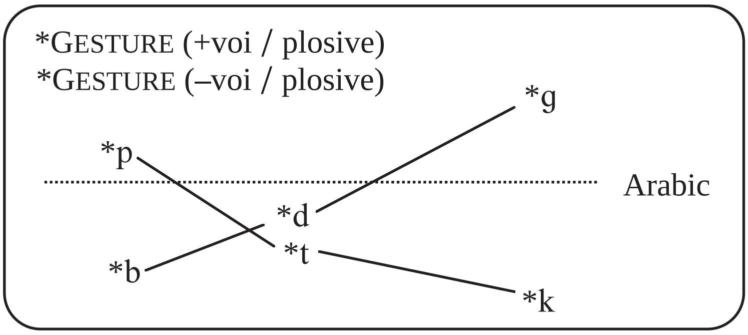
Thus, we opt for PARSE constraints for perceptual features, provided with environment clauses. The interpretation of the environment “C” that occurs in PARSE (coronal / _C), is that it refers to a consonant present in the *output*, not in the input, because the ranking of the faithfulness constraints should reflect the perceptual contrast between the output results [[at˺_ːpa]] and [[ap˺_ːpa]]. The relevant constraint is not *REPLACE (t˺, p˺), but *REPLACE (coronal, labial / C).

11.9 Constraint generalization

Depending on the relative rankings of /m/ → [lab] and /t/ → [cor] and the homogeneous *GESTURE constraint, there must be languages where (11.16) can be simplified as PARSE (lab) >> PARSE (cor) (English) or as PARSE (place / plosive) >> PARSE (place / nasal) (Malayalam). This is a trivial case of generalization, empirically void because no difference can be detected with the full constraint system (11.16); it just means that there are no constraints in between the two, so that they appear as homogeneous. Only if the English and Malayalam-type languages occur much more often than the Dutch-type languages, could we conclude that languages like to use generalized constraints.

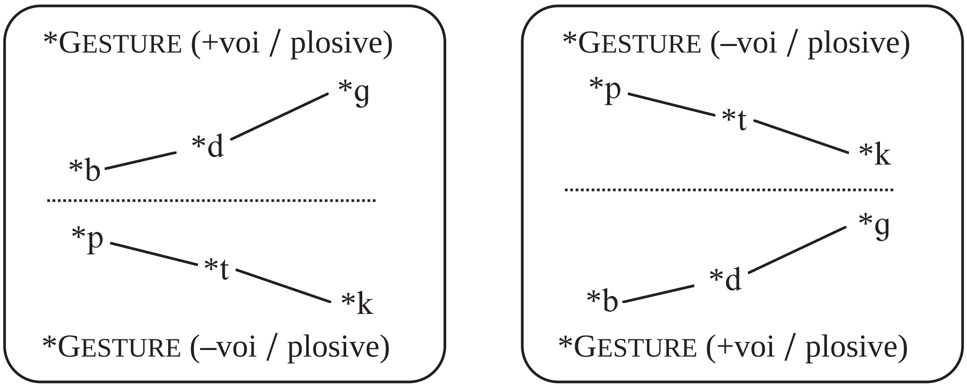
Another trivial case of generalization is the following. The near-universal hierarchy PARSE (place / onset) >> PARSE (place / coda) (which, accidentally, we need in order to derive the direction of place assimilation in §11.4), can be replaced with the ranking PARSE (place / onset) >> PARSE (place) without any empirical consequences, though the number of violation marks can be higher in the second case (if an onset place specification fails to surface). Note that we do not have to stipulate an Elsewhere principle to make this work. With this strategy, only PARSE (place / onset) and PARSE (place) need ever occur in grammars, and the constraint PARSE (place / coda) could be called *ungrounded* in the sense of Archangeli & Pulleyblank (1994). Here, the constraint PARSE (place / coda) can be considered superfluous because one of its simplifications would do as well.

As an example, we will now see how the articulatory problems of the voicing contrast in plosives can be generalized in the grammar. Because the amount of air that can expand above the glottis depends on the place of constriction, and some air expansion is necessary to keep the vocal folds vibrating for some time, a /g/ is more difficult to voice than a /d/ or a /b/ (Ohala & Riordan 1979; see also the simulation in §5.12, and the examples in §). For voiceless plosives, the situation is the reverse of this. Thus, we get the following global hierarchy of articulatory effort (a “phonetic difficulty map” in the words of Hayes 1996b) for voicing contrast in plosives:



(11.20)

The lines in this figure connect universal rankings of voicing difficulty. Note that there are no lines connecting *p and *b, because languages, according to the LRP, cannot rank the effort of the two different gestures (say, pharynx widening and vocal-fold abduction) in a universal manner. Nevertheless, from the fact that more languages have a gap in their plosive system at /g/ than at /k/, and more languages have a gap at /p/ than at /b/, we may conclude that the phonetic difficulties are close to those portrayed in the figure. We can see, then, that Arabic actually respects the global hierarchy: it lacks /p/ and /g/ (and /ɢ/), as shown in the figure with a dotted line, which represents a homogeneous PARSE (±voi) constraint. In general, however, languages are free to rank the two families, so we expect to find lots of the following rankings:


*GESTURE (+voi / plosive)
*g
*d
*b
*p
*t
*k
*GESTURE (–voi / plosive)
*GESTURE (–voi / plosive)
*p
*t
*k
*g
*d
*b
*GESTURE (+voi / plosive)

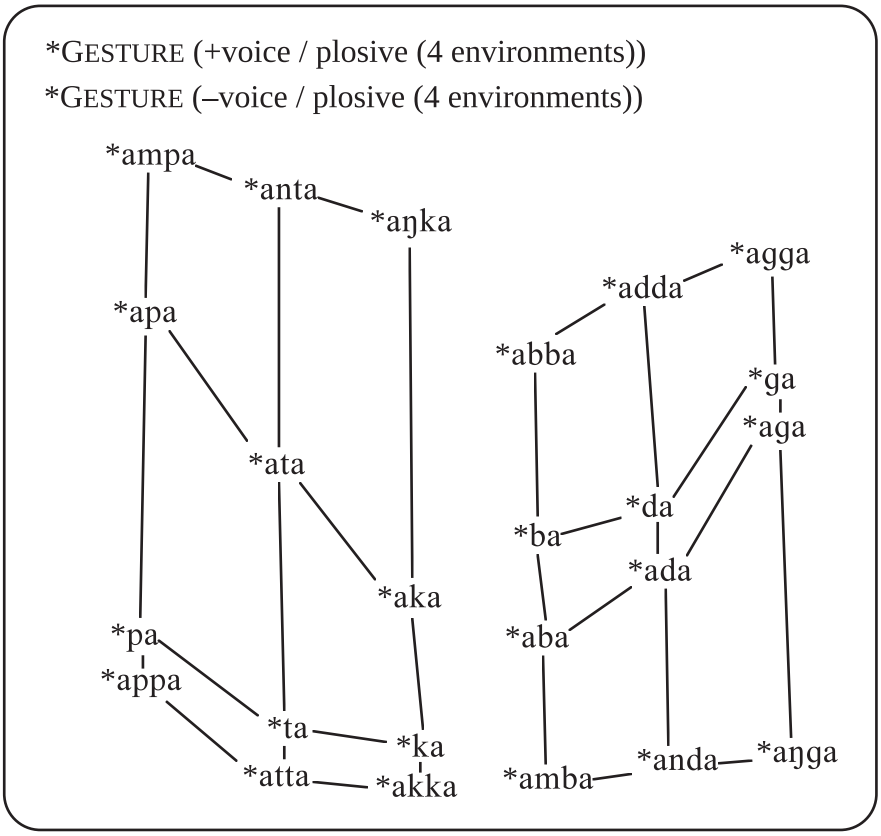
(11.21)

If the global map of (11.20) is correct, we expect to find a larger number of languages of the type pictured on the left of (11.21), than of the type on the right, i.e., we will find more languages with only voiceless stops than with only voiced stops: a tendency expected by the principle of (11.2c).

But there is a difference with the PARSE hierarchy seen before. Once that a gesture has been learned, its *GESTURE constraint falls to a low position in the overall constraint hiererchy. Because voicedness and voicelessness are implemented by very different gestures, the separations depicted in (11.21) are expected to be much more common than a grammar that allows a plosive inventory restricted to [b], [t], and [k]; this is different from faithfulness, because, say, learning of the perceptual feature value [+voice] automatically involves learning of the perceptual feature value [–voice].

11.10 Phonologization

The procedure of the previous section can be extended from single gestures to coordination. The phonetic hierarchy (11.20) would look differently for plosives in initial position (less easily voiced than elsewhere), geminate plosives (hard to voice), intervocalic plosives (easy to voice), and post-nasal plosives (hard to devoice). Hayes (1996b) gives a tentative measure for the effort associated with all 24 cases, based on Westbury & Keating's (1986) aerodynamic vocal tract model, which predicts the possibilities of voicing on the basis of transglottal pressure. Though Hayes uses a global effort measure, we should respect the fact that voicing and devoicing strategies use different gestures, so Hayes' numbers can be pictured as follows, if we take into account the local-ranking principle:



(11.22)

(In Hayes' table, [pa] and [aba] tie, and the seven utterances at the bottom have zero effort.) With the algorithm of random reranking, subject to the local-ranking principle (which fixes the rankings that are expressed with lines in (11.22)), several universals follow automatically:

- There are languages with voiceless plosives in every position except post-nasally (see Pater 1996).
- There are languages which only allow voiceless geminates (Japanese).
- If voiced plosives are allowed initially, they are also allowed intervocalically and postnasally (if plosives are allowed there at all, of course).

- If voiced coronals are allowed, so are voiced labials (in the same position, and if labials are allowed at all).
- Et cetera.

Besides these near-universals, several tendencies can be predicted from the global height of the constraints in the phonetic map (11.22):

- The average *b is ranked lower than the average *p, so gaps at /p/ are more common than gaps at /b/.
- The average *g is ranked higher than the average *k, so gaps at /g/ are more common than gaps at /k/.
- The average *aNÇ̥a is ranked higher than the average *aÇ̬ːa, so voiced geminates are more common in languages with geminates than post-nasal voiceless plosives in languages with post-nasal plosives.
- Et cetera.

The local-ranking principle may lead to a phonological constraint ranking that is very different from the global phonetic ranking in (11.21).

Dutch, for instance, allows /aŋka/ and not /aŋga/, although the map shows that the latter must be much less difficult (and Dutch has some voiced plosives). This is possible because the local-ranking principle allows the right half of the map to be turned counterclockwise by almost 90 degrees, without disturbing the fixed rankings, so that the quartet *agga >> *ga >> *aga >> *aŋga may dominate all other voiced-plosive constraints. These fixed rankings do predict that if a language does not allow /aŋga/ (but does allow post-nasal stops), it also disallows the near-universally worse /aga/, /ga/, and /agːa/. For Dutch, this prediction is borne out: the language simply lacks a /g/ phoneme. Thus, ranking (11.22) allows the generalization of four constraints to the simple *g.

The perfect mirror image of the Dutch example is found in Arabic and was dubbed “very striking” by Hayes (1996b: 10). Arabic has the voiced geminate [bː] but not the voiceless geminate [pː], though the phonetic map shows that *abba is ranked much higher than *appa in a global effort space. Now, the left-hand side of the map (11.21) may be turned clockwise by almost 90 degrees, so that the quartet *ampa >> *apa >> *pa >> *appa may dominate all other voiceless-plosive constraints. These fixed rankings do predict that if a language does not allow /apːa/ (but does allow geminates), it also disallows the near-universally worse /pa/, /apa/, and /ampa/. For Arabic, this prediction is borne out: the language simply lacks a /p/ phoneme. Thus, ranking (11.22) allows the generalization of four constraints to the simple *[–voi / labial plosive].

A word must, then, be said about Hayes’ solution for this phenomenon. To assess the “effectiveness” of the generalized constraint *p, he computes its average ranking number as the average of the ranking numbers of *appa (8), *pa (9.5), *apa (19), and *ampa (24)[3], as counted from the bottom in (11.22); the result is 15.1. The effectiveness of the generalized *b is 11.1, which is the average of the ranking numbers of *abba (18), *ba (13), *aba (9.5), and *amba (4). Now, Hayes’ criterion of ***inductive*** (i.e., learnable) ***grounding*** identifies *p as grounded because its effectiveness is greater than that of all its simpler or equally simple “neighbours” *b, *t, *k, *[lab] and *[–voice]. In the same way,

[3] This is an equivalent reformulation of Hayes’ very different-looking algorithm.

*b is not grounded because all of its neighbours *p, *d, *g, *[lab] and *[+voice] are more effective (a single one would have been enough to make it ungrounded). Hayes proposes that only grounded constraints make their way into the grammar.

There are several problems with Hayes' approach. First, it would mean that *[cor] and *[dors], which we can identify as *GESTURE (blade) and *GESTURE (body), do not occur in the grammar because the more effective constraint *[lab] is a neighbour, an obviously undesirable result in the light of our example of place assimilation. Another serious problem with inductive grounding is that it is a procedure based on a global effort map, and, as such, only capable of deriving tendencies, not universals. For instance, the average ranking of the voicedness constraints in (11.21) is somewhat higher (12.7) than that of the voicelessness constraints (12.3), predicting that there are languages with exclusively voiceless plosives, and no languages with exclusively voiced plosives. Though this is a strong tendency with as few exceptions (Maddieson 1984: Alawa and Bandjalang) as the "near-universal" hierarchies *d >> *b (Proto-Indo-European; Maddieson 1984: Mixe, Cashinahua) and *g >> *d (Maddieson 1984: Acoma), the question is whether languages with a single series of plosives bother at all about making them voiceless or voiced; rather, they are likely not to show active devoicing at all, giving, on the average, a "lax voiceless" stop which does not violate any glottal *GESTURE constraint. The surprise of Westbury & Keating (1986) at finding that most languages with a single stop series have voiceless stops even in intervocalic position, whereas their model predicted that these should be more easily voiced than voiceless, may be due to an oversimplification in their model: even if the transglottal pressure is sufficiently high to allow voicing, a supraglottal closure should be accompanied by an active laryngeal closing gesture in order to withstand the voicing-adverse passive vocal-fold abduction caused by the rising intraglottal pressure, as seen in our comprehensive vocal-tract model (fig. 5.13, §5.12). As an example (with unrealistic figures), consider the passive and active contributions to glottal widening in five obstruents (PCA = posterior cricoarytenoid, IA = interarytenoid):[4]

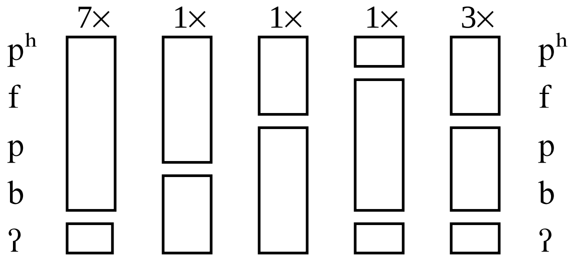
sound	supra-laryngeal closure	passive widening	active widening	muscle	total widening	acoustic result
p^h	closed	3 mm	3 mm	PCA	6 mm	aspirated
f	critical	2 mm	2 mm	PCA	4 mm	voiceless
p	closed	3 mm	1 mm	PCA	4 mm	voiceless
b	closed	3 mm	–3 mm	IA	0	voiced
ʔ	open	0 mm	–2 mm	IA	–2 mm	voiceless

(11.23)

In the column "total widening", we see the glottal strictures in the order of Ladefoged (1973). Gandour (1974), however, notes that the natural classes of initial obstruents in 13 tone-split rules in the histories of various Tai languages point to an order of [p^h, f, p, b,

[4] This simple example ignores supralaryngeal voicing gestures and the muscle-spindle reflex, which may bring the vocal folds together again after 20 ms of passive widening.

ʔ]. These tone splits are collected in the following table, where the natural classes are shown as rectangles:

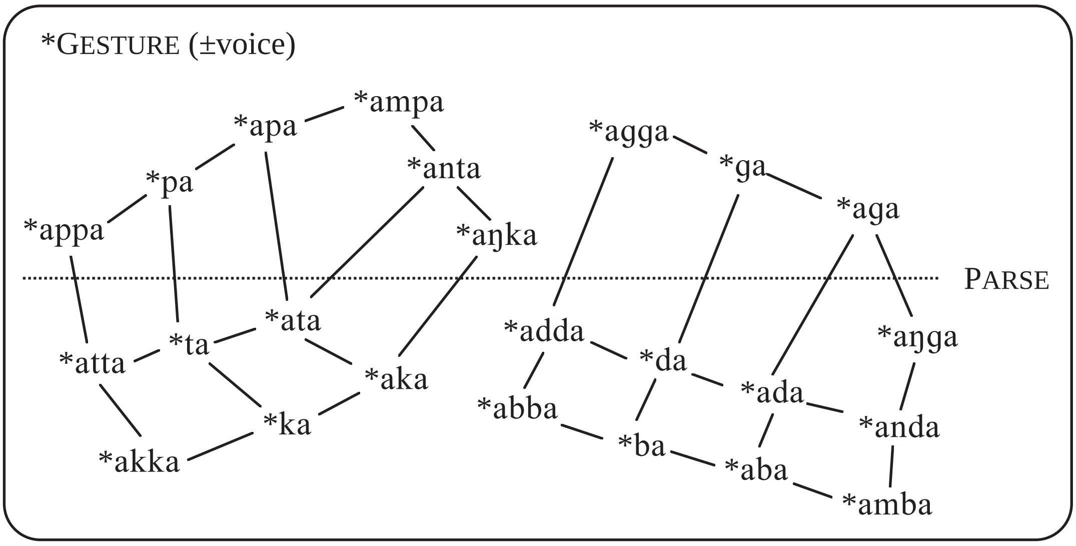


(11.24)

Gandour's solution to the disparity between Ladefoged's order and the Tai data involves a hierarchical ordering between the binary perceptual feature [±vibrating] and the multi-valued articulatory feature [glottal width]. Note, however, that sorting the five obstruents by their degree of *active* widening in (11.23) would also give the order [p^h, f, p, b, ʔ]. If there is some realism in my picture of passive glottal widening, this explains Westbury & Keating's surprise as well as the highly skewed distribution of homogeneously voiceless versus homogeneously voiced plosive systems: an active widening of 0 mm, as may be appropriate in a system without any voicing contrasts, leads to a total width of 3 mm for plosives, as can be seen in (11.23), and these may be considered "lenis voiceless". Thus, this skewed distribution cannot be taken as evidence of a universally ungrounded *[–voice] in systems that have to maintain a faithful voicing contrast in obstruents.

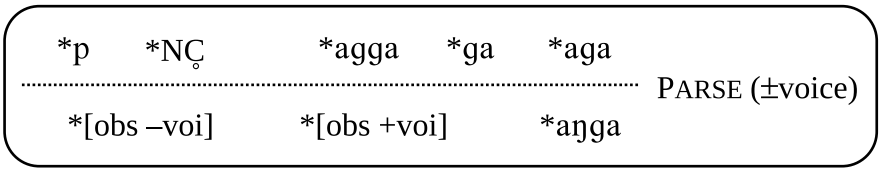
The conclusion must be that inductive grounding does not redeem Hayes' promise (1996b: 5) that "we seek to go beyond mere explanation to achieve actual description". Rather, a much simpler strategy based on local ranking, which does not need a global effort map, correctly generalizes phonetic principles to phonological constraints. Just turn the symmetric diamond ◊ by almost 45 degrees in either direction.

As an example, consider a language which lacks /p/, /g/ (except post-nasally), and post-nasal voiceless plosives. Such a language should be able to exist according to the fixed rankings in (11.22). Without changing the ranking topology of this map, we can transform (11.22) into:



(11.25)

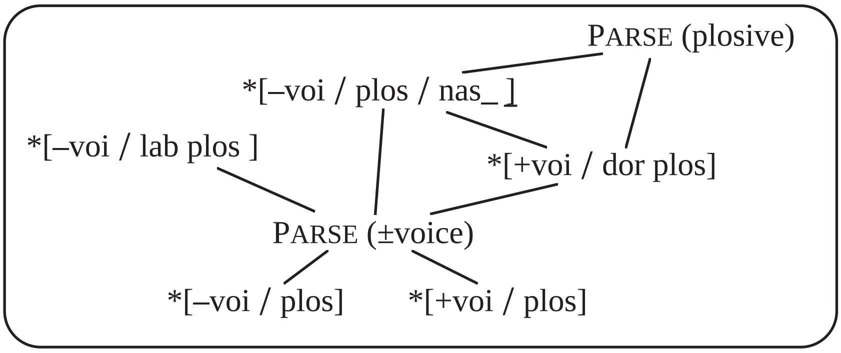
This can be simplified as



(11.26)

Thus, a simplification like (11.26) is allowed by the local-ranking principle. Cross-linguistically, languages seem to prefer these simplifications over drawing a dotted PARSE (±voice) line through the middle of (11.22). This effect cannot be explained by an asymmetry in the learning of voicing versus devoicing gestures, since the language of (11.25) obviously uses both of these gestures to a large extent. Rather, its success lies in the simplification itself: (11.26) needs fewer constraints than the average language that can be derived from (11.22) by restricted random reranking.

The remaining complexity with /g/ in (11.26) can be resolved by noting that if the language had a homogeneous *g constraint, there would be no way to parse a dorsal nasal-plosive sequence, as *[aŋka] is ruled out by *NC̥. Therefore, a strong PARSE (plosive) constraint may force the surfacing of [aŋga]. The following constraint system can handle this:



(11.27)

Note that if PARSE (plosive) is ranked at the top, *[–voi / plos / nas_] must dominate *[+voi / dor plos]; with the reverse ranking, underlying dorsal nasal-plosive sequences would show up as [aŋka] instead of [aŋga]: a minimal difference.

From the 24 articulatory constraints that we started with, only five remain, even in this relatively complex language. The reformulation of the *GESTURE constraints in (11.27) is explained below.

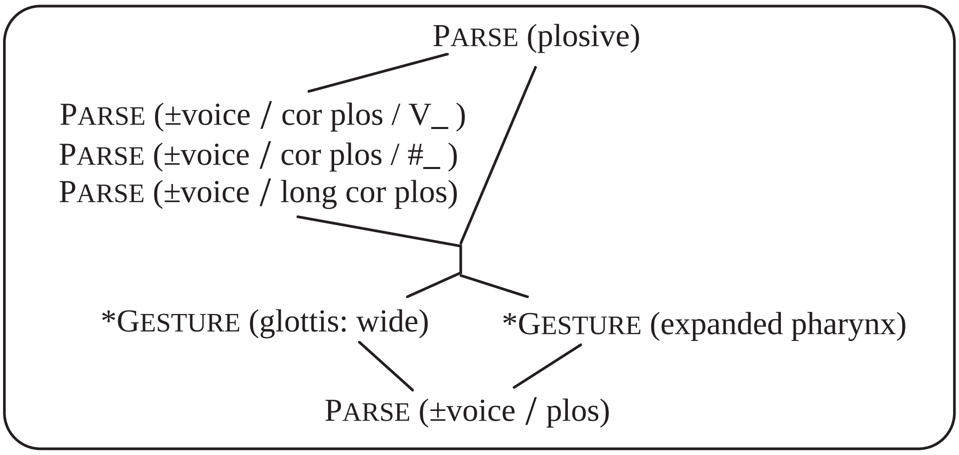
11.11 Homogeneous *G ESTURE or homogeneous PARSE?

The reader may have noticed that in §11.2 to §11.7, a homogeneous *GESTURE constraint was used to divide up interestingly ranked PARSE families, whereas in §11.10 a homogeneous PARSE constraint was used to divide up interestingly ranked *GESTURE families. Clearly, we cannot have both at the same time. In this section, I will solve this mystery and show that a phonetic map like (11.22) and a language like (11.25) can also be described with homogeneous *GESTURE constraints and varying PARSE constraints.

First, we can note that the articulatory constraints in (11.27) are explicitly shown in an “implementational” formulation: *[–voi / lab plos] >> *[–voi / cor plos] means that it is more difficult to make a labial plosive voiceless than to make a coronal plosive voiceless. Of course, this ranking can only be fixed if these formulations refer to the same degree of perceptual voicing for the labial and coronal cases. Thus, more effort is required for the implementation of the [aba] - [apa] contrast than for the [ada] - [ata] contrast, *given that the perceptual contrasts are the same in both cases.* Now, equal contrasts mean equal PARSE constraints (ch. 9), so use of a homogeneous PARSE (±voice) constraint for all places is legitimate.

While the PARSE (voice) constraints are equally high for the various places, the *GESTURE constraints are not. The implementationally formulated constraint *[–voi / lab plos] is really something like *GESTURE (glottis width: 3 mm), and *[–voi / cor plos] is something like *GESTURE (glottis width: 2 mm), which is universally ranked lower, if the gesture is considered made from a state of phonation-friendly vocal-fold adduction.

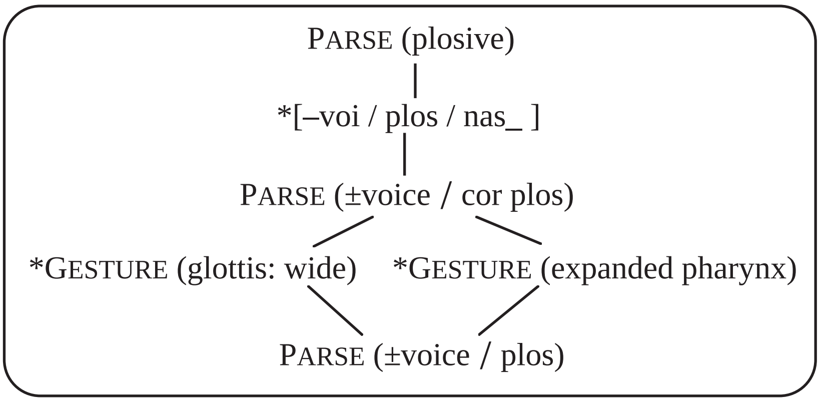
The voicing theory described above is perception-oriented. We can also devise an articulation-oriented theory, namely, one that says that only particular gestures are learned. For instance, if we learn to use the gesture “glottis width: 2mm” for the implementation of voiceless plosives, a /p/ will surface as less voiceless than a /t/. Likewise, with equal voicing gestures (pharynx expansion or so), a /g/ will come out as less voiced than /k/. Thus, PARSE (±voice) will be ranked less high for labials than for coronals, and PARSE (±voice) will be ranked less high for dorsals than for coronals. For post-nasal position, PARSE (±voice) will be ranked very low because post-nasal plosives with a 2mm glottis-width gesture will be voiced in such an environment, so that the perceptual contrast with the result of the expanded-pharynx gesture is very small. A working constraint hierarchy is:


PARSE (plosive)
PARSE (±voice / cor plos / V_)
PARSE (±voice / cor plos / #_)
PARSE (±voice / long cor plos)
*GESTURE (glottis: wide)
*GESTURE (expanded pharynx)
PARSE (±voice / plos)

(11.28)

This yields a language slightly different from (11.27): for labial and dorsal plosives and for post-nasal plosives, no voicing contrast exists, and neither of the gestures will be used for them. The automatic results may be something like [aba], [b̥a], [apːa], [aɡ̊a], [ka], [akːa], [amba], [anda], and [aŋga]; the minimal difference referred to below (11.27) does not exist. Note that it is no coincidence that both *GESTURE constraints in (11.28) seem to be on the same height: if PARSE (±voice) falls below one of them, the voicing contrast is neutralized, so that PARSE (±voice), whose ranking depends on contrast, falls further.

To sum up, the ranking in (11.27) expresses the articulatory problem of implementing the perceptual voicing feature faithfully, whereas (11.28) expresses the resistance against using articulatory gestures that do not result in good perceptual voicing contrasts. Real languages will allow both of these ideas to play a role in the grammar. For instance, the simplest constraint ranking for the languages in (11.27) and (11.28) would be



(11.29)

This uses only six constraints; both (11.27) and (11.28) needed one more. The ranking (11.29) expresses the following ideas: except for coronals, the voicing contrast in plosives, as implemented by a fixed pair of gestures, is so low that is too unimportant to maintain; for coronals, therefore, the contrast is maintained, except in post-nasal position, where the implementation of [–voice] is too difficult.

11.12 Licensing

In the previous section, we noted two different ways to phonologize articulatory constraints.

In the first interpretation, articulatory phonological constraints directly militate against certain fixed articulations. Typical examples are all the constraints proposed in chapter 7, most notably *GESTURE.

The second interpretation emerges from an interaction with perceptual requirements, and sees articulatory phonological constraints as constraints against the effort of implementing fixed perceptual results; their arguments, therefore, are perceptual features, not articulatory gestures. A typical example is *[–voi / plos / nas_]. Generally, we can call these constraints ***licensing constraints*** (since the dependence of their ranking on the environment tells us in what positions the feature is licensed) or ***implementation constraints***.

Def. licensing constraints: *[*f*: *v* / *env*]

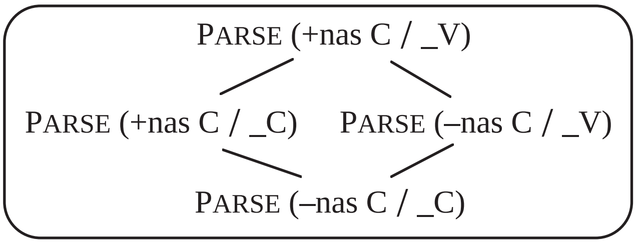
> "The value *v* on a perceptual tier *f* is not implemented in the environment *env*." (11.30)

Licensing constraints seem the only way to reconcile a functional approach with a single system of features: articulatory gestures may be removed from the grammar. However, as seen in §11.11, these licensing constraints are *GESTURE constraints in disguise, and can be universally ranked with the procedures of chapter 7. In §13.2, we will see that the more fundamental *GESTURE constraints are probably needed: the fact that most languages with voiceless nasals also have aspirated plosives; this can most easily be explained directly with the lowness of *GESTURE (spread glottis), and not with constraints like *[asp] and *[–voi / nasal]. Note that because of their grounding in more basic articulatory constraints, a functional ranking of licensing constraints such as *NC̥ >> *VC̥V is legitimate, but a similarly-looking ranking of assimilation constraints such as *[np] >> *[tp] is not: the former ranking may involve articulatory constraints only (as in 11.27), whereas the second ranking crucially involves an interaction with faithfulness constraints.

11.13 Assimilation of nasality

After place assimilation and the implementation of voicing contrasts, our third example involves the assimilation of nasality.

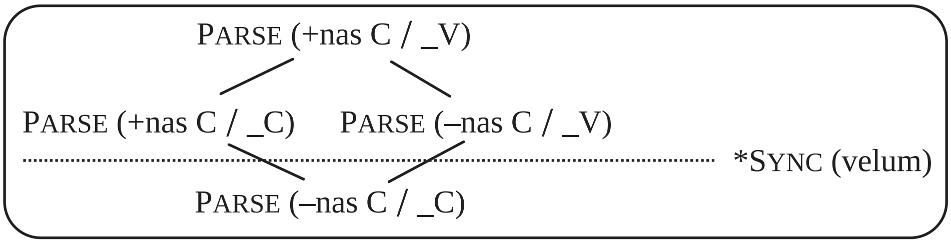
In Sanskrit, word-final plosives assimilate to following nasals: /ak#ma/ → [aŋma]. From the commonness considerations of §9.5 ([+nasal] is less common than [–nasal], because fewer constrasts can be made with [+nasal] sounds than with [–nasal] sounds), we can expect that this is a less offensive change than assimilation of [–nasal], as in /aŋ#pa/ → [akpa]. Also, we can expect that onset specifications are stronger than coda specifications, as with our example of place assimilation. This leads to the following near-universal ranking:



(11.31)

The presence of “C” in the argument of PARSE makes this an explicitly segmental formulation, a shorthand for PARSEPATH (nasal & root) or PARSEPATH (nasal & timing), though it could be replaced with a formulation involving higher prosodic units (by replacing “C” with “μ” or “σ”, for instance).

According to the local-ranking principle, all rankings not shown with straight lines in (11.31) are free. Sanskrit makes the following choice:



(11.32)

The relevant articulatory constraint is not from the *GESTURE family, but from the *SYNC family, and militates against a velar movement inside a CC cluster.

We expect the following typology for assimilation of nasality:

(a) Nothing assimilates (most languages).
(b) Plosives assimilate to a following nasal (Sanskrit).
(c) Coda consonants assimilate their nasality to the following [±nas] consonant (spreading of [–nas] is found in the North-Germanic sound change /ŋk/ → /kː/).
(d) Plosives assimilate to a nasal on either side. (11.33)

There are only four (not six) possibilities, because (c) and (d) both already satisfy *SYNC (velum). Note that none of the four violates FILL (+nas).

The typology (11.33) is equivalent to the following set of independent implicational universals for nasal spreading within consonant clusters:

(a) If [–nas] spreads, so does [+nas].
(b) If [+nas] spreads rightward, it also spreads leftward. (11.34)

11.14 Conclusion

Starting from a typological interpretation of the local-ranking principle, we derived a successful strategy for simplification of the grammar:

The functional view of the phonologization of functional constraints

> "From all the grammars allowed by the local-ranking principle, languages tend to choose a grammar in which many constraints can be generalized over their arguments or environments." (11.35)

12 Correspondence: Segmental integrity versus featural autonomy

Abstract. Segmental effects are caused by "vertical" connections between perceptual tiers, and autosegmental effects are caused by "horizontal" connections between perceptual cues.

This chapter handles the problem of the correspondence between features in the input and features in the output, which has a large bearing on the interpretation of faithfulness.

In chapter 9, I proposed a large number of faithfulness constraints. The workings of some of these are likely to overlap. Though all these constraints can be defended by invoking functional principles, it is not unthinkable that phonology allows only a subset of them to play a role in language. In this section, we will compare two hypotheses for a reduction of the number of necessary faithfulness constraints:

a. Segmental integrity:

> "All featural faithfulness relations are transferred through the segment, which is the complete bundle of simultaneously present features." (12.1)

The typical representative of this approach is the "linear" ***Correspondence Theory*** of McCarthy & Prince (1995), who used the following constraints:

- MAX-IO: **if** the input contains a segment, **then** this segment should also be in the output (like our PARSE, but for segments, not features).
- IDENT-IO (*f*): **if** the input segment **and** the corresponding output segment both contain the feature *f*, **then** the two values of this feature should be equal (like our *REPLACE).

For instance, IDENT-IO (voice) is satisfied if the value for the feature [voice] in the input is equal to the value for [voice] in the ***corresponding segment*** in the output, and it is violated if these values are unequal. But if either the input or the output does *not* contain the bearing segment, the constraint is *not* violated.

b. Featural autonomy:

> "Every specified feature has its own faithfulness constraints, which try to bring it to the surface." (12.2)

Archangeli & Pulleyblank (1994) simply state that "the notion of segment is both inadequate and superfluous" and that phonology works with features, nodes, and links (though they do incorporate a root tier). My account in chapter 9 also brought up featural faithfulness as predominantly autonomous, disregarding correspondence through segments, controlling faithfulness with constraints like:

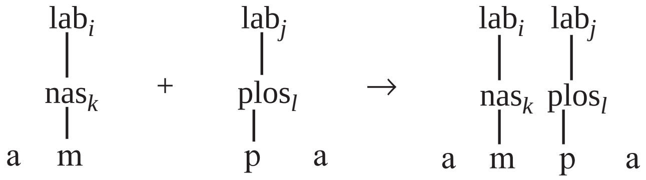
- PARSE (feature: *x*): **if** the input contains the feature value *x*, **then** *x* should also be in the output.

In the examples of chapter 11, however, I tacitly handled the faithfulness of features by using segments as their domains. In the following, we will relieve this tension and consider the relative merits of the linear and the autosegmental approaches.

12.1 Is perception segmental?

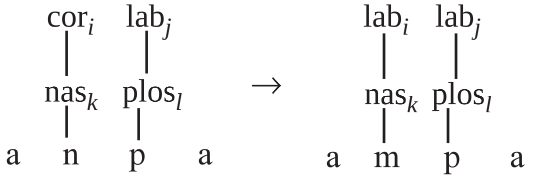
With a distinction between articulation and perception, there is a very simple solution to the everlasting problem of segmental (or *linear*) versus autosegmental processes: the consonant cluster in [ampa] contains a single articulatory labial gesture, but is heard as containing two separate instances of a perceptual feature [labial]. Thus, we can evaluate our faithfulness constraints via linearly ordered segments, and still understand that assimilation is spreading of an articulatory gesture. In this way, we have the best of both worlds.

In chapter 11, we assumed the segmental interpretation of faithfulness to our advantage. For instance, we did not mark the concatenation /ʊm/ + /poːtə/ → /ʊmpoːtə/ with a violation of PARSE (labial). Thus, the correspondence in this example is like in the following diagram (which uses /a/ vowels in line with following diagrams):



(12.3)

Another process discussed earlier, the assimilation /an+pa/ → [ampa], can be seen as a replacement of [coronal] with [labial] on the perceptual place tier, but only if we represent the two [labial] feature values of the output as separate:



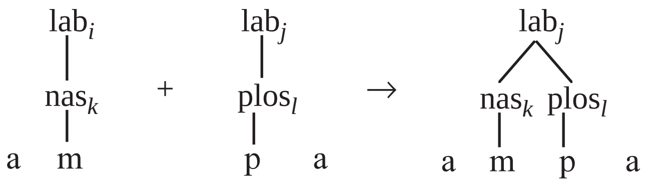
(12.4)

Now, it might just be the case that this is the correct rendering of the perceptual score, and that autosegmental representations respecting the OCP (§1.3.1, ch. 18) are limited to the articulatory score. Such a hypothesis would express a nice functional correlate of the tension between segmental and autosegmental phenomena: there is a single lip gesture, but separate labial sounds.

But in those cases where features are not neatly lined up, it is often difficult to even count the number of segments in an utterance. For instance, does the usual pronunciation of *tense* with an intrusive stop lead to four or five segments on the surface? And there are several other problems with the segmental approach.

12.2 OCP-driven epenthesis

I argued earlier (§11.8) that several place cues can collectively contribute to the perception of a single value of the perceptual place feature. For instance, transition and burst together will make us hear a single instance of [labial] in [apa], microscopically [[ap˺_pa]]. Such a single labial would surely also be perceived in a prenasalized stop as in [a^{m}pa]. But a homorganic cluster as in [ampa] is, in many languages, by far the most common nasal-plosive sequence, and it would be advantageous to the listener to hear them as a cluster with a single place, in accordance with what happens gesturally. Thus, we could re-represent (12.3) with autonomous features as



(12.5)

In a segmental approach, no constraint at all is violated: Max-IO is satisfied because all underlying segments appear in the output, and the resulting [m] corresponds maximally faithfully with underlying /m/. In an autosegmental approach, by contrast, we have only *one* labial feature left in [mp], whereas the two underlying segments /m/ and /p/, coming from two different morphemes with separate lexical representations, contribute *two* labial specifications. Therefore, we have a violation of Parse (labial), and the utterance is indistinguishable from an utterance with a single underlying dorsal gesture (i.e., a tautomorphemic homorganic nasal-plosive cluster). This violation of Parse is a faithfulness problem, so we expect interactions with other constraints, such as Fill.

As an example, consider the following data from Geleen Limburgian, where the diminutive suffix /kə(n)/ shows epenthesis of [s] when attached to a stem that ends in a dorso-velar:[1]

pop (pl. popə) ‘doll’	pøpkə
lɑm̄p (pl. lɑm̄pə) ‘lamp’	læm̄(p)kə
kom̄p (pl. kǿm) ‘bowl’	kǿmkə
bɔūm (pl. bœ́ym) ‘tree’	bœ́ymkə
dû:f (pl. dū:və) ‘pigeon’	dŷ:fkə
ʃtʀɔ̂:t (pl. ʃtʀɔ̂:tə) ‘street’	ʃtʀœ̂:cjə ([c, ɲ] = palatalized alveolar)
bɛt (pl. bɛdə) ‘bed’	bɛcjə
mɑn̄ (pl. mǽn or mɑ́nə) ‘man’	mǽnkə (place assimilation forbidden)
bɑl̄ (pl. bǽl) ‘ball’	bǽlkə
kɑʀ̄ (pl. kǽʀ) ‘cart’	kǽʀkə
jɑs (pl. jæs) ‘coat’	jæskə
ɦū:s (pl. ɦū:zəʀ) ‘house’	ɦøskə (irreg. vowel)

[1] The highly productive diminutive morpheme is expressed as: umlaut (fronting of back vowels); softening (leaving only the sonorant from underlying sonorant + voiced plosive sequences); tone change (changing an underlying circumflex tone into an acute, but not before a voiceless consonant); and the suffix /-kə(n)/. The notations /â:/ and /ɑ́n/ denote the acute (falling or rising) accent (Stoßton), /ā:/ and /ɑn̄/ the circumflex (drawling) accent (Schleifton).

kes (pl. kestə) ‘chest’	keskə
kɑɲ̄c (pl. kǽɲ) ‘side	kæɲ̄cjə
ɦoɲ̄c (pl. ɦǿɲ) ‘dog’	ɦǿɲcjə
wóɲ (pl. wóɲə) ‘wound’	wǿɲcjə
vœʃ (pl. vœʃə) ‘fish’	vœʃkə
bǽlʃ (pl. bæl̄ʒə) ‘Belgian’	bǽlʃkə
blɔk (pl. blœk) ‘block’	blœkskə
ɦɛk (pl. ɦɛgə) ‘hedge’	ɦɛkskə
plɑŋ̄k (pl. plæŋ̄k) ‘plank’	plæŋ̄kskə
dɛŋ̄k (pl. dɛ́ŋəʀ) ‘thing’	dɛ́ŋskə
ɔ́ux (pl. ɔūɣə) ‘eye’	œ́yxskə
lê:x (pl. lê:xtəʀ) ‘light’	lê:xskə

(12.6)

In Correspondence Theory, this epenthesis cannot be represented, because a violation of DEP-IO (= FILL) is always worse than no violation at all, independently of the relative rankings of MAX-IO, DEP-IO, and IDENT-IO (place):

/ŋ+k/	MAX-IO	DEP-IO	IDENT-IO(place)
*☞ [ŋk]			
[ŋsk]		*!	

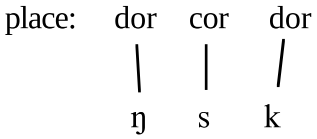
(12.7)

In the purely autosegmental approach, PARSE (dorsal) may be strong enough to force epenthesis:

/ŋ+k/	PARSE (dorsal)	FILL (sibilant)
[Nk]	*!	
☞ [Nsk]		*

(12.8)

With the epenthesis of [s], PARSE (dorsal) is no longer violated, because the two dorsal specifications of /ŋ/ and /k/ are now separated on the perceptual place tier:



(12.9)

Though some theories (e.g., McCarthy 1988) might still consider the two unary [dor] features adjacent because there is no conflicting value on the same tier, we cannot represent them with one specification without getting the ‘gapped’ representation that Archangeli & Pulleyblank (1994) militate against. Going back to the fundamentals (i.e., function), we see that there is perceptual separation on the place tier: there are no separate perceptual coronal and dorsal tiers[2].

12.3 Horizontal and vertical correspondence

In chapter 8, we handled the acoustics-to-perception faithfulness of an utterance consisting of a single feature; in such a case, the question what input feature values correspond to what output feature values, has a simple answer. But if an utterance contains multiple simultaneous feature values across tiers and multiple ordered feature values within tiers, the correspondence question becomes more complicated. The general idea is that it is favourable for the listener to perceive a set of acoustic cues or perceptual features that often occur together, as a single feature.

One aspect of this occurring together is the grouping of simultaneously occurring features, discussed in §9.11. If the "vertical" path constraints are strong, we can expect segmental effects.

The other aspect of occurring together is the grouping of acoustic cues or feature values that occur after one another. If cue A is usually followed by cue B, they may be recognized as a single feature value. If the "horizontal" temporal identity constraints are strong, we expect autosegmental effects.

In OT, every conflict is resolved by a constraint, so the conflict between the segmental representation (12.3) and the autosegmental representation (12.5) must be handled by a constraint as well. I propose the following pair of listener constraints (in the perception grammar) for the temporal correspondence between the acoustic input and the perceptual result

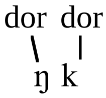
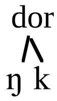
Def. OBLIGATORYCONTOURPRINCIPLE (*f*: *x*; cue_1 | *m* | cue_2)
"A sequence of acoustic cues cue_1, cue_2 with intervening material *m* is heard as a single value *x* on the perceptual tier *f*." (12.10)

Def. NOCROSSINGCONSTRAINT (*f*: *x*; cue_1 | *m* | cue_2)
"A sequence of acoustic cues cue_1, cue_2 with intervening material *m* is not heard as a single value *x* on the perceptual tier *f*." (12.11)

These constraints (which I shall abbreviate as OCP and NCC) promote a maximally easy perceptual organization. The more often the cues occur together, the greater the chance that they are perceived as a single feature value; this is true for simultaneity (segmentalism) as well as temporal ordering (autosegmentalism). They can be universally ranked by such things as temporal distance of the cues, rate of occurrence within morphemes versus across morphemes, etc. For instance, I believe that OCP (place: lab; transition | silence | burst) is ranked so high that the plosive in [apa] is represented almost universally with a single perceptual place specification. Not much lower would be the constraint that forces us to hear a geminate consonant as having a single place value: OCP (place: lab; transition | double silence | burst). Lower still would be the constraint

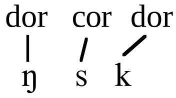
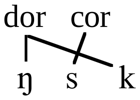
[2] Long-distance "OCP effects" that forbid the use of the same articulator twice within a domain, are due to a *REPEAT constraint that works exclusively with articulatory gestures (chapter 18).

against hearing homorganic nasal-plosive clusters as having a single place value: OCP (place: labial; side | silence | burst). The NCC constraint would be ranked in the other direction: the more intervening material, the higher the ranking. In the Limburgian case, the acoustic input [[ŋ_k]] is perceived with a single dorsal place value (pl = place, sil = silence, bu = burst):

[[ŋ_k]]	OCP (pl: dor; side | sil | bu)	NCC (pl: dor; side | [s_] | bu)	OCP (pl: dor; side | [s_] | bu)	NCC (pl: dor; side | sil | bu)
	*!			
☞ 				*

(12.12)

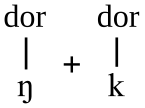
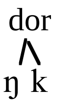
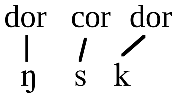
The acoustic input [[ŋs_k] is perceived with two dorsal place values:

[[ŋs_k]]	OCP (pl: dor; side | sil | bu)	NCC (pl: dor; side | [s_] | bu)	OCP (pl: dor; side | [s_] | bu)	NCC (pl: dor; side | sil | bu)
☞ 			*	
		*!		

(12.13)

Note that association lines cross in the second candidate, since there is a single perceptual place tier.

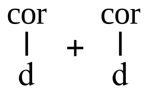
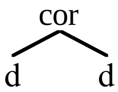
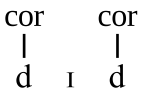
Though OCP and NCC are not constraints of the production grammar, they influence the result of the production grammar, since the perception grammar is used for evaluating faithfulness (fig. 6.1). The following tableau evaluates the Limburgian case again, using the results of (12.12) and (12.13):

	PARSE (place: dorsal)	FILL (noise: sibilant)
[[ŋ_k]] 	*!	
☞ [[ŋs_k]] 		*

(12.14)

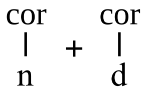
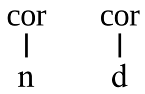
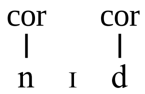
We see that in the first candidate, the highly ranked OCP causes the listener to perceive a single dorsal place value, which again causes a PARSE violation.

Consider now the English past tenses /hɛd-ɪd/ 'headed' versus /kæn-d/ 'canned'. A sequence of two [d] would be perceived with a single coronal place value; with an epenthesized vowel, [dɪd] would be perceived with two coronal place values. If PARSE (place: cor) is ranked higher than FILL (syllable), this situation forces epenthesis between homorganic plosives:

	PARSE (place: cor)	FILL (σ)
	*!	
☞ 		*

(12.15)

Between a nasal and a plosive, no epenthesis occurs. We can explain this if the [nd] sequence is perceived with two coronal place values. This is the case if NCC (place: cor; side | plosive voicing | burst) dominates its OCP counterpart. Such reversal of the two constraints (compared with the plosive-plosive case) is indeed not surprising, considering the universal dependence of the ranking of the OCP on frequency of occurrence.

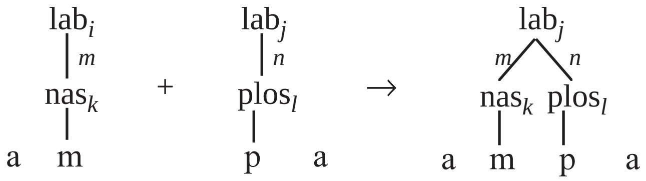
	PARSE (place: cor)	FILL (σ)
☞ 		
		*!

(12.16)

The OCP-based account described here manages the data of Limburgian and English well and makes the typological prediction that if heteromorphemic homorganic nasal-plosive clusters undergo epenthesis, then plosive-plosive clusters undergo epenthesis as well.

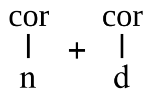
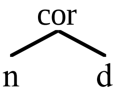
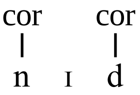
But there is still a problem. There seems to be a segmental intuition that the perceptual loss of identity of the first /d/ in /d+d/ → /dː/ is greater than the loss of identity of /n/ in /n+d/ → /nd/. It would be nice if we could express this intuition with a variation in the ranking of a faithfulness constraint, instead of burdening the listener with a dual-coronal representation of /nd/.

We can respect OCP (place: dorsal) in /nd/ if we notice that no identity is lost on the combined place and nasal tiers. We can rewrite (12.5) as



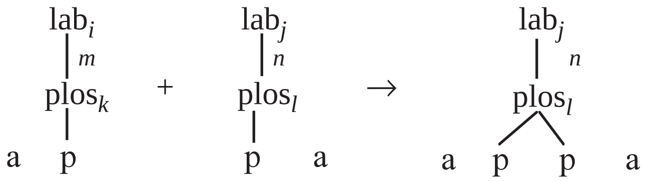
(12.17)

On the combined place-nasal tiers, correspondence is between feature combinations, not between the separate features: it is (nas lab)$_m$, not (nas$_k$ lab$_i$), and the former is preserved in the output, though Parse (lab) is still violated. Back to (12.16): even if OCP (place: cor; side | plosive voicing | burst) dominates its NCC counterpart, the winning candidate is [nd]:

	Parse (nasal × place: nas & cor)	Fill (σ)	Parse (place: cor)
☞ 			*
		*!	

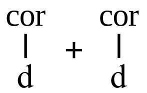
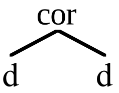
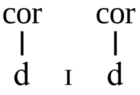
(12.18)

The analogue of (12.17) for plosive-plosive clusters is:



(12.19)

The same constraint system as in (12.18) will now have to evaluate /hɛd+d/:

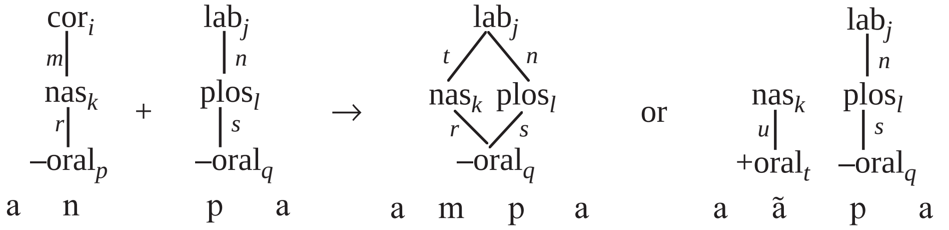
	Parse (nasal × place: plosive & cor)	Fill (σ)	Parse (place: cor)
	*!		*
☞ 		*	

(12.20)

Technically, we could have done the job with the near-universal ranking Parse (cor / plosive) >> Parse (cor / nasal), derived earlier (9.39) from considerations of perceptual confusability, but this is a coincidence that we probably cannot use for all cases.

The somewhat unsettling problem with (12.18) is that even for the seemingly trivial case of /m+p/ → /mp/ we need a ranking like PARSE (nas & lab) >> PARSE (lab), against the functional ranking principle of §9.10.

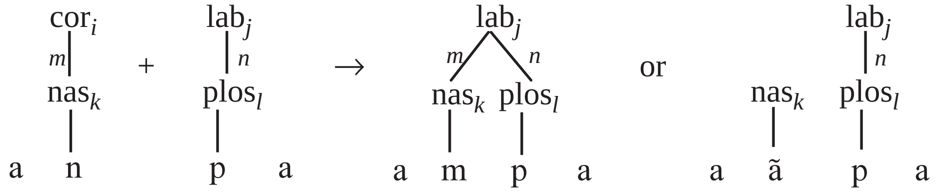
For the assimilation case (12.4) of /an+pa/, we have to understand why the candidate [ampa] is better than [aãpa]. In the segmental account, this is because the non-orality (or consonantality) of /n/ is preserved in [m] but not in [ã][3]. In the autosegmental account, however, non-orality is not even preserved in [ampa], because this feature is shared with [p]:



(12.21)

We see that both candidates violate PARSE (cor), PARSE (nas & cor), PARSE (–oral & cor) (though not shown, this path must be present), and PARSE (–oral), and that [ampa] also violates FILL (nas & lab), while [aãpa] violates FILL (+oral), PARSE (+nasal & –oral), and FILL (+nasal & +oral). Note that this example shows that PARSE (–oral) is not necessarily the same as FILL (+oral) in autosegmental perceptual phonology. Actually, however, FILL (+oral) is not violated in this case, since [ã] must share its [+oral] value with [a]. The real conflict, therefore, is between FILL (nas & lab) on the one hand, and PARSE (+nasal & –oral) and FILL (+nasal & +oral) on the other. Languages that highly estimate the preservation of nasal non-orality, will end up with [ampa]; those that do not like to hear a labial nasal where it is not specified, will end up with [aãpa]; in both cases, cross-tier faithfulness constraints decide the issue. If the /p/ in (12.21) were a fricative, there would only be one change: [ampa] would not violate PARSE (–oral).

The process /an+pa/ → [ampa] can be represented with less violation of correspondence than in (12.21). Though TRANSMIT (place) may be ranked lower than *REPLACE (cor, lab), this situation may well be reversed for the combined feature [place × nasal] (§9.8): TRANSMIT (place × nasal) may be ranked higher than *REPLACE (place × nasal: cor & nas, lab & nas) because within the combined [place × nasal] space, [cor & nas] and [lab & nas] are relatively close together. Instead of (12.21), we get



(12.22)

The input [cor & nas]$_m$ now corresponds to the output [lab & nas]$_m$. The main candidates are evaluated (without any constraints involving [–oral]) according to:

[3] We cannot yet say that consonantality is not subject to the OCP because it belongs in the root node. Such things have to be derived, not posited, in a functional phonology.

/an+pa/	PARSE (nasal)	*GESTURE (blade)	TRANSMIT (place × nasal / _ C)	*REPLACE (nas cor, nas lab / _ C)
anpa		*!		
ampa (12.21)			*!	
☞ ampa (12.22)				*
aãpa			*!	
apa	*!		*	

(12.23)

(The candidate [apa] loses as long as PARSE (nasal) dominates *GESTURE (velum).) This evaluation, involving the correspondence in (12.20), is the interpretation of the example of §9.5 and §11.4, which involved the less accurate constraint *REPLACE (cor, lab / _ C). We see that strong "vertical" constraints like TRANSMITPATH can force segment-like behaviour: the faithfulness constraints in (12.21) look suspiciously like MAX-IO and IDENT-IO (place), but note that we can only use the latter pair of constraints if we do not consider the [ã] in [aãpa] to be a segment (it could be transcribed as [ãpa], with [ã] corresponding to both /a/ and /n/; see §12.5); with our more restricted path constraints, such a stipulation is unnecessary.

Conclusion: we need path constraints whose featural coherence is greater than that of autonomous features, but smaller than that of a segment.

An inherent problem in autosegmentalism. In the autosegmental approach, subtraction may sometimes be evaluated as addition.

The process /dap/ → [dãp] violates FILL (nasal) and FILL (nasal & vowel), whereas /dam/ → [dãm] violates only FILL (nasal & vowel). Therefore, the former process is always worse: insertion of a marked value or privative feature is worse than spreading.

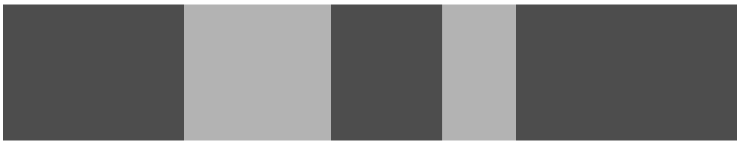
Likewise, the process /dãp/ → [dap] violates PARSE (nasal) and PARSE (nasal & vowel), /dãm/ → [dam] violates only PARSE (nasal & vowel). Therefore, the former process is always worse: deletion of a marked feature is worse than deletion of its association line only.

The symmetry seen in the /dap/ and /dam/ cases is related to the idea that the perceptual contrast between [dap] and [dãp] is larger than that between [dam] and [dãm], a difference that can be ascribed to the general process of lateral inhibition (a nasality contrast is less easy to hear next to a nasal). An asymmetry is due to the markedness of [+nasal] (§9.5): the process /dap/ → [dãp] must be less bad than /dãp/ → [dap], suggesting that for marked feature values, PARSE violations are worse than FILL violations.

But problems arise with /pap/ and /mam/. Let us assume that the distinction between [pap] and [pãp] is larger than the distinction between [mam] and [mãm].

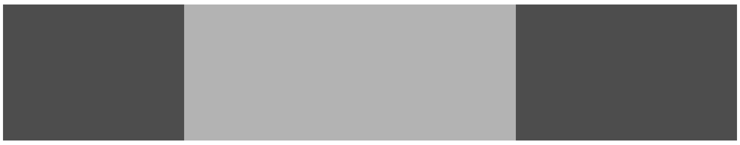
The process /pap/ → [pãp] violates FILL (nasal) and FILL (nasal & vowel), whereas /mam/ → [mãm] violates PARSE (nasal) and FILL (nasal & vowel). The violation of

PARSE (nasal) can be illustrated with the following metaphor. Suppose we start with a sequence of dark-light-dark-light-dark rectangles:



(12.24)

If we paint the middle rectangle in a light shade of grey, one dark rectangle is lost:

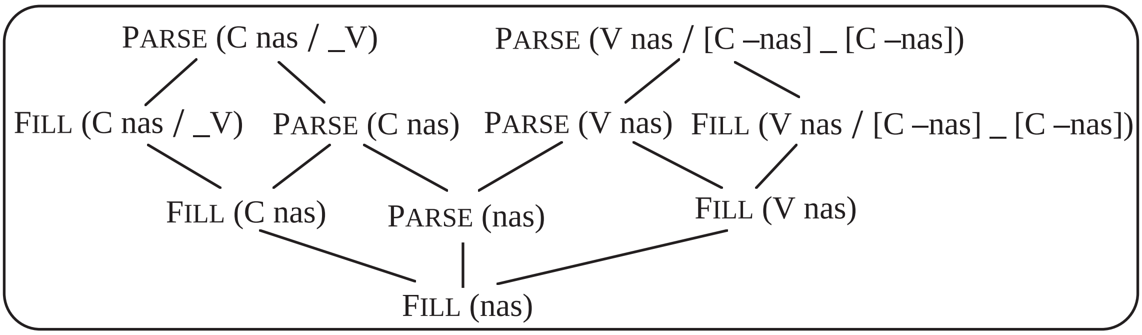


(12.25)

As we see, however, one light rectangle is also lost. Adding nasality to the vowel in [mam] thus violates PARSE (nasal). Now, if the [pap] - [pãp] distinction is larger than the [mam] - [mãm] distinction, the change /pap/ → [pãp] is more offensive than the change /mam/ → [mãm], so that FILL (nasal) must dominate PARSE (nasal).

The process /pãp/ → [pap] violates PARSE (nasal) and PARSE (nasal & vowel), whereas /mãm/ → [mam] violates FILL (nasal) (like going from (12.25) to (12.24)) and PARSE (nasal & vowel). If the latter process is less bad than the former, PARSE (nasal) must dominate FILL (nasal), so there is a contradiction with the previous pair.

We can get out of the predicament only by assuming such rankings as PARSE (+nas & vowel / [–nas & cons] _ [–nas & cons]) >> PARSE (+nas & vowel / [+nas & cons] _ [+nas & cons]), and the same ranking for FILL, together with low rankings of PARSE (nasal) and FILL (nasal). So, we can finally replace the naive nasality faithfulness rankings of §9.8 with an accurate grammar (cf. 11.31) that handles all cases of the spreading of nasality to adjacent plosives and vowels:



(12.26)

12.4 Floating features

The faithfulness of floating features cannot be represented at all within Correspondence Theory, because these features are by definition not underlyingly connected to a segment, which makes IDENT-IO insensitive to them. This was already recognized by McCarthy & Prince (1995); the solution they suggest is the generalization of the MAX and DEP constraints to autosegmental features. However, this would involve not just a simple generalization of Correspondence Theory to the featural domain, because some constraint families will show considerable overlap: the separate need for the Ident-IO family (together with Max-IO) will be severely reduced as a consequence of the existence of the Max (*feature*) family (though in a *comprehensive* approach we may need all of them).

Zoll (1996) explicitly evaluates correspondence through output segments, even for floating features (if these dock onto segments). For instance, Zoll argues that in Inor, the verb /kəfəd/ plus the masculine floating affix [round], which together give [kəf^{w}əd] (because [round] will dock on the rightmost labial or dorsal segment), should be analysed as if both underlying /f/ and underlying [round] correspond to the output segment [f^w]. This would lead to the following evaluation:

/kəfəd/ + [round]	Max (Seg)	Max (Subseg)	Ident (F)
☞ kəf^{w}əd			
kəfəd		*!	
kəf^{w}əz			*!
kəf^{w}ə	*!		

(12.27)

Two remarks are in order.

First, Zoll holds the underlying /f/ to correspond to surface [f^w] without violating Ident(F) (a constraint that requires that the featural make-up of corresponding segments should be the same), because Zoll "follow[s] the proposal of Orgun 1995 and 1996 in assessing violations of Ident(F) only in cases of absent or differing specifications, but not when the output correspondent is more specified than the input". As we have seen in §9.5, we can explain such an asymmetry between input and output without such stipulations: it follows directly from the markedness of the feature [round] in the average utterance and the listener's optimal recognition strategy, which leads to the near-universal ranking Parse (round) >> Fill (round). In other words, it is worse to replace /f^w/ with [f] than to replace /f/ with [f^w]. Thus, the segment-based constraint Ident(F) is superfluous.

More important is the fact that both (sub)segmental Max constraints can be replaced with featural correspondence constraints. In the winning candidate, Parse (round) is satisfied. The only problem with [kəf^{w}əd] is that [round] has been linked with a labial consonant; but this is less bad than linking [round] with the coronal consonant (in Inor), although that is final. The complete constraint system gives:

/kəfəd/ | [round]	PARSE (cor)	FILL (noise)	FILL (rnd &cor)	PARSE (rnd)	FILL (rnd & lab) FILL (rnd & dor)	*SHIFT (σσ)	*SHIFT (σ)
☞ kəf^{w}əd					*		*
k^{w}əfəd					*	*!	*
kəfəd				*!			
kəf^{w}əz		*!					*
kəfəd^w			*!				
kəf^{w}ə	*!						*

(12.28)

I misused the constraint PARSE (cor) for assessing the loss of the final segment in [kəf^{w}ə]; the question of featural or temporal segmentality (i.e., whether we should have taken PARSE (root) or PARSE (timing) instead of PARSE (cor)) is independent from the question of featural correspondence discussed here. The *SHIFT family evaluates the suffixal specification of [round], as suggested by the "|" in the representation; note that this constraint is vacuously satisfied if the floating [round] does not surface, and that it rates [k^{w}əfəd] as worse than the winner (§9.13), it will be replaced with a continuous family.

12.5 Fusion

In the simple fusion /n+b/ → [m] (e.g., Tagalog /maŋ+bili/ → [mamili] 'buy'), one segment disappears.

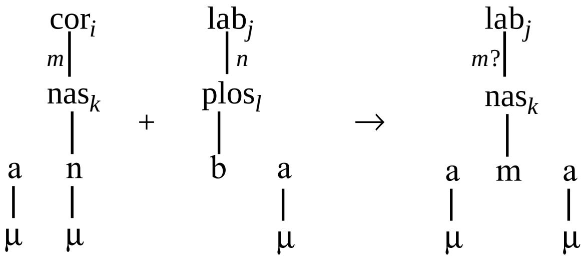
First, assume that the deleted segment is /b/. In Correspondence Theory, this means that there is one violation of MAX-IO. This must be brought about by a higher-ranked constraint, say the anti-cluster constraint *CC. However, because underlying /n/ now corresponds to surface [m], we also have a violation of IDENT-IO(place). In that case, as (12.29) shows, the candidate [n] would always be better, independently of the ranking of MAX-IO and IDENT-IO (place).

The second strategy would be to assume that the deleted segment is /n/. In this case, the output candidate [m] must correspond to the input /b/, violating IDENT-IO (nasal). Correspondence Theory would then predict the output [b], independently of the ranking of MAX-IO and IDENT-IO (nasal). Thus, the output [m] cannot possibly win, unless it corresponds to both input segments:

/n_i+b_j/	*CC	MAX-IO	IDENT-IO (place)	IDENT-IO (nasal)
[$n_i b_j$]	*!			
[m_i]		*	*!	
☞ [n_i]		*		
[m_j]		*		*!
☞ [b_j]		*		
[m_{ij}]		?	?	?

(12.29)

To represent the fusion /n+b/ → [m] correctly, Correspondence Theory would have to be extended appreciably, because it is no trivial matter to decide whether MAX-IO or IDENT-IO are satisfied or not in (12.29). The autosegmental account, by contrast, views features as independent of segments. The fusion process is shown as follows:



(12.30)

PARSE (coronal) and PARSE (plosive) are violated, but the universal frequency-based rankings of PARSE (labial) above PARSE (coronal) and PARSE (nasal) above PARSE (plosive) guarantee the output [m]:

/n+b/	*CC	PARSE (μ)	PARSE (labial)	PARSE (nasal)	PARSE (coronal)	PARSE (plosive)
[nb]	*!					
☞ [m]		*				*
[n]		*	*!			
[b]		*		*!	*	
[d]		*	*!	*!		

(12.31)

So, fusion is most easily described with PARSE constraints for fully autonomous features.

12.6 Phonetic substance of epenthesis

There is a technical problem that purely segmental (linear) theories have trouble handling. An assessment of the featural content of epenthesized segments is impossible within Correspondence Theory: IDENT-IO is insensitive to any extra segment in the output, exactly because the epenthesized segment has no correspondent in the input. In the autosegmental account, the most faithful epenthesis candidate (i.e. the one that violates the least severe FILL constraint) will be the one that adds the fewest features or paths to the output (unless, of course, the epenthesis is meant to separate identical elements, as in §12.2).

12.7 Subsegmental satisfaction by segmental deletion

As we can see from its definition, IDENT-IO can be satisfied by means of the deletion of a segment. An example of this may be found in Limburgian, where the /n/ in the masculine singular ending of articles and adjectives is only licensed by following laryngeal consonants and coronal stops: /dən/ ‘the’ + /dāːx/ ‘day’ becomes /dəndāːx/ ‘the day’ (likewise: dən-tīːt ‘the time’), but /dən/ + /bɛ̂ːʀ/ ‘bear’ becomes [dəbɛ̂ːʀ]: rather than deleting only the coronal gesture, which would give *[dəmbɛ̂ːʀ], the whole segment is deleted (likewise: də-ʃtɛīn ‘the stone’).

Segmental account. Apparently, IDENT-IO outranks MAX-IO (we use an ad-hoc nasal-consonant (NC) homorganicity constraint to make [nb] ill-formed):

/dən+dāːx/	IDENT-IO(place)	NC-HOMORGANIC	MAX-IO (ənC)
☞ dəndāːx			
dədāːx			*!

(12.32)

/dən+bɛ̂ːʀ/	IDENT-IO(place)	NC-HOMORGANIC	MAX-IO (ənC)
dənbɛ̂ːʀ		*!	
dəmbɛ̂ːʀ	*!		
☞ dəbɛ̂ːʀ			*

(12.33)

Thus, in this case, the segmental account seems appropriate. We will now see that all attempts to describe the phenomenon with the assumption of featural autonomy, are problematic.

Autosegmental account. If autosegments were autonomous, constraint satisfaction by deletion could not occur: independently of the ranking of the two PARSE constraints involved, *[dəmbɛ̂ːʀ], which violates PARSE (coronal), would always be a better candidate than [dəbɛ̂ːʀ], which violates both PARSE (coronal) and PARSE (nasal). To solve this problem, we could put the constraint *GESTURE (velum) in between the two PARSE constraints:

Version 1 (covertly segmental)	PARSE (cor/_V)	*GESTURE (blade)	PARSE (cor/_C)	*GESTURE (velum)	PARSE (nas/ə_C)
☞ dəndāːx		*		*	
dədāːx		*	*!		*
dənbɛ̂ːʀ		*!		*	
dəmbɛ̂ːʀ			*	*!	
☞ dəbɛ̂ːʀ			*		*

(12.34)

All rankings in this tableau are crucial: any other ranking of the same constraints would give a different result. The idea is that the inviolable parsing of the place features of the onset consonant (/d/) forces the tongue-tip gesture and thereby *licenses* the surfacing of coronality in the nasal consonant (because the two segments share the same gesture). A nice result, and we can relate the rarity of this phenomenon to the critical ranking that is needed: even if we assume that PARSE (cor / _V) is universally undominated, there are 24 possible rankings of the four remaining constraints, and only one of those rankings produces the correct result. Unfortunately, there is a flaw. In a truly autosegmental framework, [dəndāːx] actually violates PARSE (coronal), according to the OCP; in §12.2, it was proved that Limburgian considers a homorganic nasal-plosive sequence to have a single [coronal] specification. But [dəndāːx] does not violate the segmental-integrity constraint PARSE (nasal & coronal), which is part of the specification and requires the co-occurrence of two perceptual features:

Version 2 (illogical)	PARSE (cor/_V)	*GESTURE (tongue tip)	PARSE (cor)	PARSE (nas & cor)	*GESTURE (velum)	PARSE (nas/ə_C)
☞ dəndāːx		*	*		*	
dədāːx		*	*	*!		*
dənbɛ̂ːʀ		*!			*	
dəmbɛ̂ːʀ			*	*	*!	
☞ dəbɛ̂ːʀ			*	*		*

(12.35)

Rather strange in this proposal, however, is the crucial ranking of the more general Parse (nas) below the more specific Parse (nas & cor), allowing a *Gesture constraint to intervene, contrary to the universal logical ranking defended in §9.10. It seems we will have to use a constraint against [m]: not against [m] in this position in general ([əmb] is an otherwise licit sequence), but against [m] where there is no underlying labial nasal; in other words, Fill (nas & lab), which is unviolated:

Version 3	Parse (cor/_V)	*Gesture (tongue tip)	Parse (cor)	Fill (nas & lab)	Parse (nas/ə_C)	*Gesture (velum)
☞ dəndāːx		*	*			*
dədāːx		*	*		*!	
dənbɛ̂ːʀ		*!				*
dəmbɛ̂ːʀ			*	*!		*
☞ dəbɛ̂ːʀ			*		*	

(12.36)

The two “nasal” constraints have been crucially reranked. An undominated Fill (nasal & dorsal) constraint is needed as well. This account takes care of the fact that Limburgian is adverse to nasal place assimilation in general. The obvious functional reason for ranking Fill (nas & lab) so high is that the result of violating it is the creation of an otherwise licit path (or the creation of an existing phoneme, so to say), thus crossing the border between two main categories.

The crucial ranking of Parse (nas/ə_C) >> *Gesture (velum) in (12.36) is needed to ensure the surfacing of the /n/ is [dəndāːx]. In (12.35), the reverse ranking was needed to get rid of the [m] in [dəmbɛ̂ːʀ]. There are three reasons to prefer (12.36):

1. With (12.36), we understand the general resistance of Limburgian against place assimilation of nasals. No association lines should be added.
2. In (12.35), Parse (nas & cor) is crucially ranked with respect to *Gesture (blade). In (12.36), Fill (nas & lab) is not crucially ranked with the constraints to its left. Therefore, (12.36) is the simpler grammar.
3. If we accept the ease of correspondence between /n/ and [m], we cannot use ParsePath or FillPath, but should use TransmitPath and *ReplacePath instead. This gives the same ranking as with FillPath:

Version 4	PARSE (cor/_V)	*GESTURE (tongue tip)	PARSE (cor)	*REPLACE (nas cor, nas lab)	TRANSMIT (nas/ə_C)	*GESTURE (velum)
☞ dəndāːx		*	*			*
dədāːx		*	*		*!	
dənbɛ̂ːʀ		*!				*
dəmbɛ̂ːʀ			*	*!		*
☞ dəbɛ̂ːʀ			*		*	

(12.37)

Whether we represent this phenomenon with PARSE (nas & cor) “we are only interested in /n/ if it stays coronal”, or as FILL (nas & lab) “do not create an [m] where there is no /m/”, both the marked PARSE ranking and the combinatory FILL constraint express an attitude to the segment that is contrary to the idea of autonomous features.

Though the above example seems to make a case for the “segmental” approach, Lombardi (1996) notices that there are no languages that satisfy a final-devoicing constraint by deletion of underlying voiced segments only. Thus, a grammar that allows /at#/ to surface as [at], but forces /ad#/ to become [a], does not occur. Nevertheless, this is what a ranking of IDENT-IO (voice) above MAX-IO would have to give:

/at#/	CODAVOICELESS	IDENT-IO (voice)	MAX-IO
☞ at			
a			*!

(12.38)

/ad#/	CODAVOICELESS	IDENT-IO (voice)	MAX-IO
ad	*!		
at		*!	
☞ a			*

(12.39)

If the typological interpretation of Optimality Theory, namely that all thinkable rankings give possible grammars and that all possible grammars are given by a thinkable ranking, is correct, the non-existence of the above grammar must lead us to conclude that IDENT-IO (voice) is not a viable constraint. If we consider, instead, the feature [voice] as an autonomous autosegment, we can replace the offensive constraint with PARSE (voice); even if we rank this above PARSE (segment) (which is the same as MAX-IO), there is no deletion:

/at#/	CodaVoiceless	Parse (voice)	Parse (segment)
☞ at			
a			*!

(12.40)

/ad#/	CodaVoiceless	Parse (voice)	Parse (segment)
ad	*!		
☞ at		*	
a		*	*!

(12.41)

This gives the correct result (final devoicing), since deletion of the segment is no way to satisfy Parse (voice).

12.8 Conclusion

To sum up: the overall rarity of featural constraint satisfaction by deletion of a segment, and typical autosegmental effects such as fusion, OCP-driven epenthesis, and floating features pose insuperable problems to a linear version of Correspondence Theory.

So we use Parse (*feature*), and if we need control over the exact location of features in the output, which is the rationale behind any segmental approach, we can use path constraints like Fill (*feature*$_1$ & *feature*$_2$). The idea is that all aspects of segmentality are as violable as any other constraints.

The grammar of most languages apparently handles both *segmental* effects, which are caused by “vertical” (simultaneous) connections between perceptual tiers, and *autosegmental* effects, which are caused by “horizontal” (sequential) connections between perceptual cues. Phonology thus seems to involve all the faithfulness constraints that we can derive from general principles of human perception, which, after all, is capable of grouping simultaneous as well as sequential events.

13 Degrees of specification

Abstract. Underspecification is not a separate principle of phonology, but is an illusion created by normal interaction of faithfulness constraints. Instead of representational solutions, we can maintain degrees of specification.

Of all the representations in figure 6.1, the ***underlying form*** or ***perceptual specification*** is probably the least accessible to direct investigation, so it seems appropriate to ask what a functional theory of phonology has to say about what this representation looks like. The main dispute in phonological theory about the underlying form centres around the number of feature values that the underlying form contains. In all current theories of ***underspecification***, however, segments are either completely specified underlyingly for a given feature, or not specified at all for that feature. In this chapter, I will defend the position that this all-or-none strategy common to these current theories, can be replaced with a strategy of varying the ***degree*** to which the separate features are specified.

The term *underspecification* is used for two not necessarily related phenomena: the fact that some features are redundant in underlying representations (e.g., the segment /m/, being a sonorant, does not have to be specified for [+voice]), and the fact that some features (like [coronal]) are more likely not to surface than some other features (like [labial]). In this chapter, I shall address both of these phenomena.

13.1 Different feature systems for inventories and rules

In a formal formulation of a phonological rule, a natural class is often represented by a bundle of features. Such a bundle specifies the features common to the segments that undergo the rule. Usual phonological practice uses the same features for rules as it does for describing the contrasts in sound inventories:

> “redundant phonological features are mostly inert, neither triggering phonological rules nor interfering with the workings of contrastive features.” (Itô, Mester & Padgett 1995, p. 571)

However, the number of features used for describing sound inventories is usually the minimum that is needed to catch all the possible contrasts. There is no *a priori* reason why these should be the same as those needed in rules. For instance, languages might never contrast more than two values for the feature [voice]; nevertheless, the involvement of segments bearing this feature in phonological processes like voicing assimilation is likely to depend on the actual implementation of the voicing feature in the language at hand. I will show that there are also *empirical* reasons for not assuming the identity of distinctive and inclusive features.[1]

[1] cf. Archangeli & Pulleyblank (1994: 52): “both unpredictable, lexically specified F-elements as well as completely predictable F-elements may play either active or inert roles in the phonologies of different languages”.

13.2 Redundant features

The segment /m/ is allegedly underspecified for the feature [voice]. From the functional viewpoint, however, it is completely specified as /labial, nasal, stop, voiced, sonorant, consonant, bilabial/: a complete set of perceptual features. Voicing is an inalienable facet of the listener's idea of how the segment /m/ should sound, i.e., if it is not voiced, it is less of an /m/. The non-redundant feature [sonorant] might be sufficient, because an unvoiced [m̥] is not sonorant any longer, but an /m/ made non-sonorant will be more /m/-like if it is voiced than if it is voiceless, so [+voiced] is independently needed. Consider the situation where the common cold obstructs your nasal tract. The following tableau shows the three relevant candidates, and the solution that you are likely to choose:

/m/ + cold	*GESTURE (open nasal tract)	PARSE (nasal)	PARSE (voice)	*GESTURE (lowered velum)
[m]	*!			*
☞ [b]		*		*
[p]		*	*!	*

(13.1)

Though the articulatory distances of both [b] and [p] to [m] are comparable, the perceptual distance of [b] to [m] is much smaller than the [p] – [m] distance. We see that the superiority of [b] over [p], can only be explained if the constraint PARSE (voice) is allowed to compete, i.e., if the feature [voice] is present.

Of course, they who consider this strategy a part of phonetic implementation, which would be a stratum that is ordered *after* redundant feature values have been filled in, would consider this example phonologically irrelevant. Therefore, I will have to address the positive evidence that has been adduced for the underspecification of voicing for sonorants.

The idea that some features are redundant in underlying representations, is based on two, not necessarily related, reasons: redundancy for describing inventories, and inertness in phonological rules. I will tackle both.

The inventory argument: "in many segment inventories, all sonorants are voiced but obstruents exhibit a voiced/voiceless contrast; therefore, sonorants are not underlyingly specified for voice".

To make a sonorant, like /m/, voiceless, you have to actively widen your glottis to a large extent; otherwise, because the airflow is not seriously obstructed above the larynx, the vocal folds will not cease to vibrate. In an obstruent, like [b] or [p], voicelessness is brought about more easily, because the labial closure decreases the glottal airflow, which disfavours vocal vibration; instead, sustaining the vibration now requires some extra effort. In other words, for a voiceless [m̥] we need aspiration, and for voiceless [p] only a condition that we will vaguely call "obstruent-voiceless", and we can assume a fixed ranking of the directly articulatory constraint *GESTURE (glottis: spread) above the implementationally formulated (licensing) constraint *[–voiced / obstruent] (see §11.12).

On the perceptual side, we have the PARSE (voice) constraints. Now, voiceless nasals are barely audible in many situations, and their place distinctions are nothing to write home about either (Ohala 1975). By contrast, voiceless plosives have salient release bursts with strong place cues. So, according to the minimal-confusion hypothesis, we can rank PARSE (voice / nasal) below PARSE (voice / obstruent).

The common inventory described above is a result of the following ranking, where we assume that the categorization is so unrestrictive as to allow the recognition of /m/, /m̥/, /b/, /p/, and /p^h/, and that all three voicing features are subject to the same PARSE (voice) constraint (the licensing constraint has been replaced with its appropriate articulatory constraint):

input	output	*GESTURE (spread glottis)	PARSE (voice/plos)	PARSE (voice/nas)	*GESTURE (obs –voi)
/m/	☞ [m]				
/m̥/	[m̥]	*!			
	☞ [m]			*	
/b/	☞ [b]				
/p/	[b]		*!		
	☞ [p]				*
/p^h/	☞ [p]		*		*
	[p^h]	*!			*

(13.2)

The resulting inventory is { m, b, p }, independent of the ranking of the rightmost two constraints. If we reverse the first two constraints, the inventory will be { m, b, p, p^h }. So four of the six possible rankings give an inventory that contains more voicing contrasts in obstruents than in sonorants, and even the inventory with the aspirated obstruent does not contain a voiceless sonorant. The two remaining possible rankings, however, will show us that nothing special is going on. First, if we rank both *GESTURE constraints (in their fixed order) above both PARSE constraints (in *their* fixed order), the inventory will be { m, p }. Finally, if we rank both PARSE constraints above both *GESTURE constraints, we get { m, m̥, b, p, p^h }. Apart from the richness of some of these inventories along the voicing dimension for obstruents, which is a result of the assumptions mentioned earlier,[2] the four types of inventories predicted here are exactly the ones that are attested in actual languages. The typological predictions are:

[2] If we had added the *GESTURE (+voi / obs) constraint, which can be ranked below *GESTURE (–voi / obs), we would have generated the inventories { m, p, p^h } and { m, m̥, p, p^h }; if we had restricted the categorization of the voicing dimension, we would have gotten { m, b, p^h } and { m, m̥, b, p^h } as well.

- As an automatic result of the fixed ranking of the two PARSE constraints and the fixed ranking of the two *GESTURE constraints (and not of an inherent property of sonorants), /m̥/ is rare in inventories. Of the 317 languages considered in Maddieson (1984), only 3 have /m̥/.
- If a language has voiceless sonorants like /m̥/, it must also have aspirated plosives like /pʰ/.[3]

This predicted implicational universal is borne out by the facts (all the languages mentioned also have a series of voiced nasals):

- Of the three languages with /m̥/ in Maddieson (1984), only Otomi and Klamath are presented with aspirated plosives, whereas Hopi is only reported to have plain voiceless stops. However, Voegelin (1956) explicitly states that exactly those Hopi dialects that have voiceless nasals, also have pre-aspirated plosives that contrast with plain plosives. In the description of Toreva Hopi, Voegelin considers two possible analyses of the stop inventory: either voiceless nasals /m̥/ and pre-aspirated plosives /ʰp/, or the phoneme sequences /mh/ and /hp/.
- Klamath (Barker 1964) has a series of nasals that are "preaspirated and voiceless throughout", and postaspirated plosives.
- In Tenango Otomi (Blight & Pike 1976), initial sequences of /h+m/ are realized as [m̥m] and /m+h/ often as [mm̥]. Medial plosives are "frequently preaspirated".
- In Temoayan Otomi (Andrews 1949), both nasals and plosives may "unite with h or ʔ to form a sequence", meaning /hm/ and /pʰ/, respectively.
- In Welsh, post-aspirated nasals alternate with post-aspirated plosives: /ən + pʰɔrθmadɔg/ → /əmʰɔrθmadɔg/ 'in Porthmadog'.
- In Iaai (Maddieson & Anderson 1994), voiceless nasals may be analysed as /hm/ sequences phonetically (because voicing starts halfway the closure) as well as phonologically (because they alternate with /m/ in the same way as vowels alternate with /hV/ sequences). Still, all voiceless plosives, except the dental, have long releases and long voice-onset times (i.e., they are aspirated).
- Jalapa Mazatec (Silverman, Blankenship, Kirk & Ladefoged 1994) has, besides voiceless nasals, full series of plain voiceless as well as aspirated plosives.
- In Burmese (Cornyn 1944, Sprigg 1965, Okell 1969), there are "voiceless" or "preaspirated" nasals, with a voiced second half, as confirmed by the measurements by Dantsuji (1984, 1986) and Bhaskararao & Ladefoged (1991), contrasting and morphologically alternating with voiced nasals in much the same way as aspirated plosives do with plain ones.
- In Tee (Ladefoged 1995), the only voiceless nasal is /n̥/. Ladefoged is not explicit about the VOT of the voiceless plosives (there are voiced ones, too), though he transcribes them as /p/ etc.
- Angami (Bhaskararao & Ladefoged 1991) has completely voiceless nasals whose second part also has oral airflow (no information about plosives).
- Xi-de Yi (Dantsuji 1982) has voiceless nasals, and an aspiration contrast in plosives.

[3] We must make an exception for final voiceless sonorants as may occur after voiceless obstruents in French, which has no aspirated plosives. Like final devoicing of obstruents, this may be caused by the universal spreading of the breathing position of the vocal folds after the utterance.

- Mizo (= Lushai) (Weidert 1975), has a series of nasals whose first part is voiceless. Bhaskararao & Ladefoged (1991) call them “voiceless (unaspirated) nasals”, in order to contrast them with the voiceless and postaspirated nasals of Angami (no information about plosives).

Thus, most of these languages with voiceless nasals also have aspirated plosives, whereas less than 30% of the 317 languages of Maddieson’s (1984) database have aspirated plosives.[4] To what extent this supports our prediction, is hard to find out precisely, because many of the above languages belong to one family (Tibeto-Burman), which may have a skewed distribution of aspirated plosives. Furthermore, in many of these languages the timing of the glottal gestures with respect to the oral gestures often differs between nasals and plosives. Thus, most of these languages use different glottal-oral coordinations for voiceless nasals and aspirated plosives, which is a somewhat surprising phenomenon. According to Ohala (1975), “voiceless nasals should be partly voiced, because otherwise we would hear no place distinctions”.

The activity argument: “the feature [+voice] can spread, but only from obstruents; sonorants, therefore, do not contain the feature [+voice]”.

This argument is due to a failure to appreciate the difference between articulatory and perceptual features. Voiced obstruents are implemented with active gestures to facilitate voicing under the adverse conditions of a supralaryngeal obstruction, such as an extra adduction of the vocal folds to compensate for the raised intraglottal pressure, a slackening of the pharyngeal and oral walls, and a lowering gesture of the larynx. Whatever combination of these tricks is used by the speaker (or the language), this “obstruent-voiced” gesture may spread to a preceding obstruent, making that one voiced as well: /s + b/ → [zb]. For sonorants, by contrast, such a gesture is less needed, and if the gesture is not there, it does not spread: /s + m/ → [sm]. The *perceptual* feature [voice], however, is present in both [b] and [m], because the vocal folds vibrate in both sounds, which leads to the perceptual impression of periodicity. If we make a distinction between articulatory gestures and perceptual features, there is no need to assume an underlying [+voice] only in voiced obstruents and a redundancy rule that should assign [+voice] to sonorants at the end of the derivation.

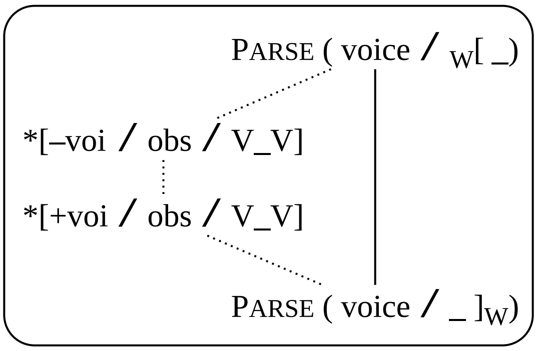
In a framework with underspecification and rule ordering, we would expect the default rule to be able to occur before the spreading rule. Thus, spreading of [+voice] from sonorants is expected to occur, and because of this, Steriade (1995) proposes a feature [expanded larynx] and a feature [voice], both of which should be able to spread. In a framework with a distinction between articulatory and perceptual features, this would not be expected. We must review, therefore, the evidence that has been brought up for the spreading of [voice] from sonorants.

First, Steriade (1995) mentions the English morpheme plural morpheme, which shows up as [+voiced] after voiced obstruents and sonorants ([bʌg-z] ‘bugs’, [k^hɔːl-z] ‘calls’), but as [–voiced] after voiceless obstruents ([tʃ$^\text{h}$ɪk-s] ‘chicks’). This can be analysed, however, with a voiced morpheme /z/, with spreading of [–voice] from

[4] The 30% is probably an underestimation caused by the common linguistic practice of transcribing aspirates as plain voiceless stops in languages without aspiration contrasts.

voiceless obstruents. Confirmation of this analysis is found with the morpheme /θ/, which, being specified as [–voice] shows no voicing after sonorants ([hɛl-θ] ‘health’), nor, for that matter, after voiced obstruents ([brɛd-θ] ‘breadth’).

Another example is final voice neutralization. In Sanskrit, word-final obstruents assimilate their voicing features to those of any following sound, be it an obstruent, a sonorant consonant (but /k+m/ → [ŋm]), or a vowel. In Limburgian, word-final obstruents “assimilate” to following plosives and vowels; before fricatives and sonorant consonants, they are voiceless. Neither of these cases has to be described as spreading from a sonorant, because in both Sanskrit and Limburgian, utterance-final obstruents devoice, which, together with the “assimilations” mentioned earlier, leads to a complete voice neutralization of word-final obstruents. Therefore, Parse (±voi / _ $]_W$) must be ranked very low, probably as a generalization of utterance-final voice neutralization: words are often utterance-final, so their final obstruents are less likely to show up as voiced than their initial obstruents, even if a voicing contrast is maintained at the end of a word but not at the end of an utterance, so Parse (±voi / _ $]_W$) must be ranked lower than Parse (±voi / $_W[$ _), and the local-ranking principle (ch. 11) does the rest. The data of Limburgian can now be explained by the following ranking (the interaction with fricative devoicing is too complex to discuss here):



(13.3)

where the solid line depicts a fixed ranking, and the dotted lines depict language-specific rankings.

The Sanskrit data are found by generalizing the right-hand environment to all sonorants. The typology suggested by the two languages derives from the near-universal ranking *[–voi / obs / V_C son] >> *[–voi / obs / V_V]. If sonorants could spread their voicing gesture, we would have to realize that sonorant consonants need a stronger voicing gesture than vowels, so that we should expect the ranking *[+voi / obs] >> *[+voi / C son] >> *[+voi / V] to be active. The typology that can be derived from this ranking would predict that there are languages where sonorant consonants spread, but vowels do not: the reverse situation from the Limburgian case. Only if such languages exist, would it be reasonable to believe in the spreading of [+voice] from sonorants.

13.3 Weak features

In our account, **specifications are constraints**. Some features, like [coronal], are less likely to surface than some other features, like labial. For instance, /n/ is specified as being coronal from the beginning, but a higher-ranked gesture-minimizing constraint can cause the underlying value not to surface (§11.4). So, Dutch /n/ only surfaces as a coronal if it cannot get its place specification from a following consonant. Underspecification theories "explain" this by stating that /n/ is not specified at all for place underlyingly, so that its place specification does not have to be erased by the following consonant, which would be an undesirable structure-changing process. Afterwards, a ***default rule*** would fill in the coronal place specification. Kiparsky (1985), who analysed the similar data of Catalan, would describe this situation with the following ordered set of rules:

1. (Underlying specifications:) /ŋ/ is specified as using the dorsal articulator and the velar place of articulation, /m/ is specified as using the labial articulator (lower lip) but has no specification for place of articulation, and /n/ is not specified for any articulator or place at all.
2. (Feature-filling assimilation rule:) every nasal consonant, if not yet specified, takes on the articulator and place of the following consonant.
3. (Feature-filling default rules:) a labial stop (plosive or nasal) that is not yet specified for place of articulation, is bilabial, and a consonant not yet specified for place at all, is coronal and alveolar.

A major drawback of such an approach is that rule 2 produces a result that can be expressed as a well-formedness condition on clusters of a nasal plus a consonant, i.e., it ensures that clusters originating when two words are concatenated, adhere to the same phonotactic constraints that hold inside morphemes. Thus, rule 2 seems to be goal-oriented (the goal being the fulfilment of the preference for homorganic clusters), but does not refer explicitly to that goal. Optimality Theory and other constraint-based theories promote these goals to the status of the actual building blocks of phonological description. In the approach of §11.4, underspecification is taken care of in a natural way: /n/ is not really unspecified for place, but the place specification for /n/ just ranks lower than many other constraints, likewise, bilabiality of /m/ emerges although its specification is weak.

Thus, underspecification is not a separate device, but an automatic result from the general theory.

13.4 The lexicon

There is one area where underspecification is still useful: the efficient storage of forms in the lexicon. For instance, a morpheme with /m/ will only contain the minimal information needed to reconstruct this segment: perhaps the specification /nasal + labial/ or just the specification /m/. In both cases, these specifications must be pointers to a fully specified list of the perceptual features that are desired in the output, like [voice].

In Chomsky & Halle (1968), the specification of the most common (or *unmarked*) values of all features could be left out of the underlying representation (*m* for "marked"), for the sake of even more lexical efficiency:

	/t/	/ɛ/	/n/	/s/
coronal	+			
voiced				
continuant				*m*
strident				
nasal			*m*	
vocalic				*m*
sonorant				
high		*m*		
back				

(13.4)

The empty cells would be filled in by redundancy rules, such as [+son] → [+voi], [+nas] → [+son], ∅ → [–voi], etcetera (note the subtle difference between "plus" and "marked"; also note that the values for [vocalic] for the first two segments could be filled in on the basis of the default CV syllable). It was not suggested that the marked values were phonologically more active than the unmarked values. The phonetic form [t^{h}ɛn^{t}s] is derived by late rules that govern the aspiration of plosives in onsets of stressed syllables, and the insertion of an intrusive stop in sequences of sonorant plus /s/.

Our example /tɛns/ would in such a theory be specified as a sequence of "oral plosive" plus "mid vowel" plus "nasal" plus "fricative", in this order and without overlap. We could run this specification through the English constraint system. All the consonants would become coronal, not labial, because *GESTURE (coronal) is ranked below *GESTURE (labial), or because FILL (coronal) is ranked below FILL (labial). The resulting output would be [[t^{h}ɛɛ̃n_ts]], like in the real world. So we could ask whether the underspecified input is real or not. The question cannot be answered in general, because it depends on what criteria of simplicity you apply. As always in phonology, there is a trade here: the simplicity of the underlying form shifts the burden of stress to the recognition phase: many FILL constraints are violated in deriving an actual output from an underspecified input. If the simplicity of recognition is the criterion, the underlying form should be maximally similar to the surface form. If the underlying form is /t^{h}ɛɛ̃n_ts/, no constraints are violated in the resulting tableau. With a "tableau of tableaux" criterion of lexicon formation (Prince & Smolensky 1993), this underlying form would be optimal.

Opting for /t^{h}ɛɛ̃n_ts/ as the underlying form, however, does not take account of the speaker's intuitions as to the phonological make-up of this morpheme. Spoken backwards (in a language game), for instance, the word is not [[st_nɛ̃ɛ^{h}t]], but [[snɛ^{ʔ}t]], which suggests an underlying /snɛt/, with an appreciable degree of segmental organization (i.e., high path constraints).

13.5 Optionality and stylistic variation

In rule-based theories, rules either do or do not apply. If a rule does not apply, it is not in the grammar. If a speaker sometimes does apply the rule, and sometimes does not, it has to be marked in her grammar as ***optional***. This is a separate device again.

In a theory based on constraint ranking, there is no such built-in phenomenon as optionality (except if we accept stochastic evaluation, as I will in chapters 14 and 15). A constraint does not have to leave the grammar if it becomes weaker. It may even still be active, but less visibly so. The rule-based counterpart of this change in visibility would be a change in the environment of the rule, a change which can not be related in any principled way to the function of the rules.

In §11.5, I showed that even within a language, constraint ranking can show variation, and that (it will come as no surprise) the division between articulatory and perceptual constraints plays an interesting role there.

13.6 Privative features

Steriade (1995) states that the need for underspecification theories is much diminished if most features are seen as privative. For instance, if [nasal] is considered a privative feature, this would “explain” the fact that nasality can spread, but non-nasality cannot. But as seen in §9.5, this effect is related to a listener strategy based on commonness, and is expressed in the grammar as the fixed ranking *DELETE (+nasal) >> *INSERTPATH (+nasal & place). The same goes for the /ɛ/ in /tɛns/: it can be underlyingly specified as [–nasal], but *REPLACE (ɛ, ɛ̃) is ranked below *SYNC (blade: open | closed, velum: closed | open).

13.7 “Trivial” underspecification

According to Steriade (1995), “plain coronals are trivially, inherently, and permanently lacking in specifications for the features [labial] or [tongue root]”. But coronals are specified for [labial] in the sense that the lips cannot be closed during the burst of [t]: as we saw in §1.2.4, the articulatory underspecification is restricted by the needs of perceptual invariance, i.e. the variable α in a dominated *REPLACE (t, α) cannot be perceptually too far away from [t]. Because the spectrum of the burst of [t] is determined most prominently by the location of the release, and less so by secondary constrictions, the illusion of underspecification comes to the surface.

13.8 Invisible specifications

In §10.1, I argued that the /s/ in /tɛns/, though not rounded at the surface, may be underlyingly unspecified for the feature [round]. In /usu/, the lips may stay rounded throughout the coronal constriction, and in /isi/, they may stay spread, so there is no empirical difference between specifying /s/ as [+round] or [–round]. Even the fact that an isolated utterance /s/ is pronounced without rounding, can be attributed to the ranking *GESTURE (lips: rounded) >> PARSE (±round / sibilant).[5] In a sense, the grammar that uses an underlyingly unrounded /s/ is simpler than the grammar that uses a rounded /sʷ/, because the former grammar inflicts a smaller constraint violation for a maximally faithful rendering of the underlying form. However, no empirical effects are associated with this "minimization of grammatical stress".

13.9 Conclusion

In functional phonology, listener-based constraint rankings replace the "unmarkedness" that other theories ascribe to certain features or feature values, and that they try to build into their representation of phonology. The explanation for each of these rankings has to be sought in the coincidental properties of the human speech channels and the human ear, not in the human language faculty. The default assumption must be that the input contains full specifications of all feature values, though some of these specifications are so weak that they can easily be overridden by articulatory constraints. These weaknesses cannot be stipulated, but can be derived instead from considerations of perceptual contrast.

[5] In English, this is not quite true, because an isolated utterance /ʃ/ is pronounced with lip rounding.

Part III

GRAMMAR

Part II developed a theory of what phonology would look like if it were governed by functional principles. This may be fine as an exercise of the bottom-up construction of an ideal world, but the resulting functional theory of phonology will be an acceptable alternative to generative approaches only if it is capable of describing phonological structures and processes with an equal or higher amount of empirical adequacy, efficiency, and explanatory force. Chapters 14 to 19 will show that Functional Phonology can stand up to this test and clarify several hitherto recalcitrant phonological issues.

14 Learning a production grammar[1]

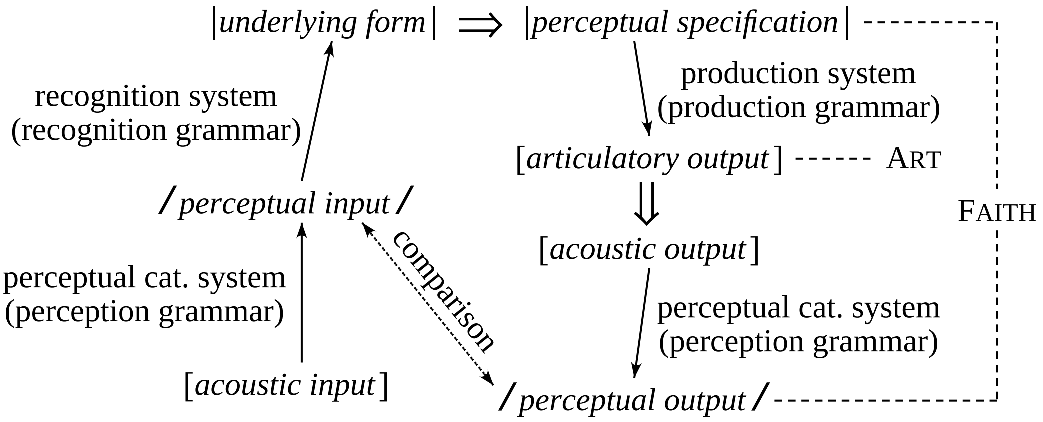
Abstract. Learners start with empty grammars, and have to learn both the constraints and their rankings. A convergent and robust gradual learning algorithm exists.

In this chapter, I will describe how learners can acquire the articulatory and perceptual constraints of segmental phonology and their interactions; specifically, I will show that the substantial content of these constraints does not have to be innate for the learner to be able to acquire an adequate grammar from a realistic amount of overt data.

I will assess the empirical adequacy of the functional learning algorithm with respect to existing algorithms that assume the innateness of constraints, and I will show that it is convergent, realistic, and robust.

14.1 Grammar model

Figure (14.1), repeated from (6.1), shows the functional concept of the linguistically relevant systems, processes, and representations of speech production and perception. The acquisition of a production grammar is the subject of this chapter. For the situation of learning, we will have to focus on the ***comparison*** module:



(14.1)

Thus, the learner can compare her own output, *as perceived by herself*, with her perception of an adult utterance, and take a ***learning step*** if there is a ***mismatch*** between these. Note that **this comparison is different from the comparison between the perceptual specification and the perceptual output**, as evaluated by the faithfulness constraints: the child should learn to imitate the adult system of faithfulness violations.

For the learning situation, two representations have to be added to a tableau like (6.3):

[1] An abridged version of this chapter will appear as Boersma (to appear). Computer versions of the algorithms in this chapter (Triggering Learning Algorithm with or without greediness and/or conservatism; Error-Driven Constraint Demotion; Minimal or Maximal Gradual Learning Algorithm) are available in the Praat program, http://fonsg3.hum.uva.nl/praat/.

[*model utterance*] /*model perc*/ $\lvert spec \rvert$	A	B
☞ [art_1] /$perc_1$/		←∗
√ [art_2] /$perc_2$/	∗!→	

(14.2)

This tableau shows a ***learning pair***: a model (adult) utterance and the corresponding learner's utterance (perhaps from an imitation one way or the other), with the following representations, all of which can be identified in figure (14.1):

(1) The *model utterance*, as it is available to the ear of the learner; not, therefore, the articulatory representation, but the acoustic form that is a direct consequence of that articulation. This is the acoustic input to the learner's perceptual categorization system, and is able to slowly change that system during acquisition.
(2) The *model perception*: the adult utterance as perceived by the learner. This is the output of the learner's perceptual categorization system.
(3) The specification is by definition the input to the production grammar. In early learning, this may be a concatenation of words as stored in the lexicon, equal to the adult overt forms in isolation, as perceived (categorized) by the child (the model perception, in other words). Later on, this may be a more abstract underlying form.
(4) Many candidate articulations.
(5) For each candidate articulation: the corresponding output, as perceived by the learner's categorization system.

The learner normally assumes that the perceived model is the ***correct form***. She knows that she is in error when her own output, as perceived by herself, is different from that form. If any of the other candidate outputs **is** equal to the correct form, learning may occur. In tableau (14.2), for instance, the learner's output /$perc_1$/ may be different from /*model perc*/, but the less harmonic candidate /$perc_2$/ may be equal to /*model perc*/. If that is the case, the learner will identify the correct form with /$perc_2$/, as is indicated in (14.2) with a check mark. The model utterance has now become a ***trigger***: the grammar will be changed. The main thing that is wrong with the current grammar, is the crucial mark incurred by constraint A for the correct form. We will call this the offending mark, and constraint A the offending constraint, or simply the ***offender***. The learner's strategy could simply be to execute the following learning step:

Minimal Gradual Learning Algorithm (Minimal GLA):

"Lower the ranking of the offender (by a little amount)."

(14.3a)

Realizing that constraint A is ranked too high or constraint B is ranked too low, we could propose an alternative strategy:

Maximal Gradual Learning Algorithm (Maximal GLA):

"Lower the rankings of all the constraints violated in the adult form, and raise the rankings of all the constraints violated in the learner's form (by a little amount)." (14.3b)

In tableau (3), the direction of this ranking change is shown by arrows in the cells with the violation marks.

As we will see, their graduality makes these algorithms robust (resistant against erroneous input or parses), and causes them to exhibit several types of realistic behaviour. Moreover, both algorithms have a property quite desirable for all learning algorithms: they can correctly and quickly learn any target grammar starting from any initial grammar (with the same constraint set). The Maximal algorithm also behaves well in situations of variation, as I will show in chapter 15.

As we will see, the ***local-ranking principle*** (ch. 11) allows the learner to manage the acquisition of continuous constraint families, and speeds up the acquisition process.

14.2 Learning in functional phonology

In functional phonology, we have an infinite number of learnable constraints. This state of affairs may pose problems for learnability, because several extant learning algorithms (Gibson & Wexler 1994; Tesar & Smolensky 1993, 1996) need a finite number of innate constraints to work. In the rest of this chapter, I will show that neither ***finiteness*** nor ***innateness*** is needed for the acquisition of phonology. I will start with the core algorithm, because that will be used for illustrating every learning stage.

14.2.1 The grammar

I will first present a view of OT grammar and constraints that is fully compatible with nearly all work in OT so far, but allows us to understand more easily several aspects of learning and reranking of constraints. The grammar, in this view, consists of a language-specific finite set of constraints $\{ C_1, C_2, ..., C_N \}$. Each constraint C has two parts: a pair-evaluation function and a ranking value.

Pair-evaluation function. The evaluation function can compare the harmonies of two output candidates with respect to C. For gestural constraints, the function is ART-EVAL (C, $cand_1$, $cand_2$), where $cand_1$ and $cand_2$ are two articulatory candidates. The function ART-EVAL returns an answer to the question which of the two candidates is the better with respect to C, or whether they are equally harmonic. For faithfulness constraints, the function is FAITH-EVAL (C, *spec*, $output_1$, $output_2$), where *spec* is the perceptual specification, and $output_1$ and $output_2$ are the perceptual results of two candidates. This function tells us which of the two output candidates, or neither, matches the specification best in a certain respect.

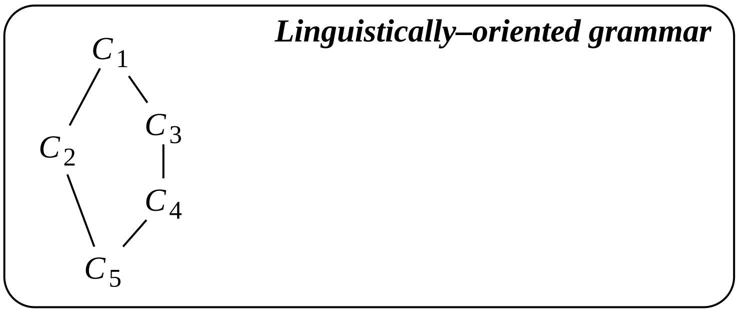
Ranking value. Each constraint is ranked along the ***continuous ranking scale***, with a real value between, say, 0 and 100 (though values outside this range must not be excluded). We will see that the continuity of this scale allows us to understand several real-life phenomena.

Comparing a pair of candidates. With the above ingredients, the procedure for evaluating the relative harmony of two candidate outputs runs as follows. Given a perceptual specification and two candidates, all N constraints are asked for their opinions. The constraints that measure no difference of harmony between the two candidates, remain silent. The other constraints issue a ***protest***. For instance, if candidate A is less harmonic than candidate B with respect to C_2 and C_5, and B is less harmonic than A with respect to C_3, the constraints C_2 and C_5 issue a protest against A, and C_3 issues a protest against B. The loudness of each protest (the ***disharmony***) is determined by the ranking value of the constraint that issues it, and the loudest protest wins. For instance, if C_3 has the highest disharmony of the three protesters, candidate B will be banned, and A emerges as the more harmonic of the two.

Evaluation of all candidates. If GEN (the OT candidate generator) generates 10 candidates, the most harmonic one can be found by starting with a comparison of the first two of these candidates, and going on by comparing the more harmonic of this pair with the third candidate, and so on. After having compared the tenth candidate with the best one of the first nine, we will have identified the winner.

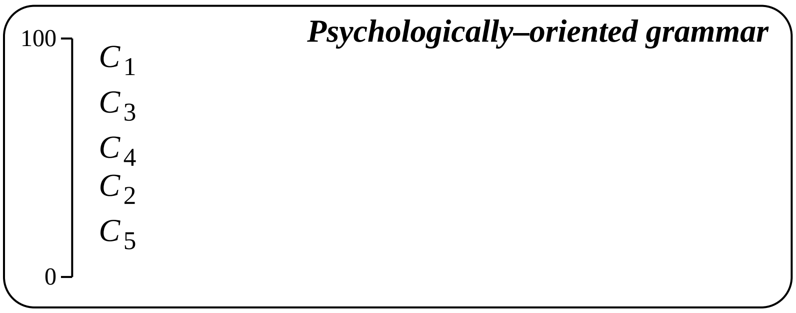
Tableaux. The result of the procedure described above will usually be the same as the result of the usual tableau evaluation. However, as an actual algorithm for finding the winning output, it uses less information (no counting of marks) and memory resources (the evaluations of all the candidates) than the batch algorithm of computing the winner from a tableau. Of course, tableaux were never meant to suggest a psychological reality, and they are particularly useful for explicit communication between linguists. The aim of the current chapter, however, is to show that learning a grammar is as natural as learning anything else, so we need an *explanation* of the learning process in realistic terms.

The view of constraint ranking as exemplified in tableaux is a hybrid representation of grammar: it is meant to represent both the behaviour of the speaker and the properties of the language. A classical OT tableau is totally ranked, i.e. the constraints C_1, C_2, and C_3 written from left to right along the top of the tableau are taken to represent the total ranking $C_1 \gg C_2 \gg C_3$. There exists a device for indicating in a tableau that two adjacent constraints are not crucially ranked with respect to one another, namely, drawing a dotted vertical line between them instead of a solid line. However, this does not work for all non-crucial rankings. For instance, if we have five constraints with crucial rankings $C_1 \gg C_2$, $C_1 \gg C_3$, $C_3 \gg C_4$, $C_2 \gg C_5$, and $C_4 \gg C_5$, the freedom of ranking cannot be represented by the usual linear ordering. Instead, the following topology of crucial rankings is more informative:



(14.4)

For the linguist, it is an interesting quest to find out which constraints are crucially ranked in a language and which are not. However, if the speaker's constraints have ranking values associated with them, she need not know topologies like (14.4), and there is no point for us in assuming that she does. Rather, her grammar will look like



(14.5)

Note that though the constraints in (14.5) are written on different heights than in (14.4), the two grammars are empirically indistinguishable if used with the fixed-winner systems most phonologists are concerned with. The main reason why I think that (14.5) better represents the cognitive capabilities of the speaker, is that the learning algorithm that comes with it, has the ten desirable properties described in the next section; other evidence for the reality of continuous ranking, though, is found in pragmatic reranking, optionality, and so on, as we will also see.

14.2.2 Gradual learning algorithms

The selection of the winning candidate in Optimality Theory can be seen as an instance of the principle "minimize the largest problem": the winner is the candidate whose largest constraint violation is lower than the largest constraint violation of its peers (disregarding shared marks). Likewise, the natural learning algorithm for an Optimality-Theoretic grammar could be equally egalitarian: again, "minimize the largest problem":

Minimal Gradual Learning Algorithm. "If the learner's current grammar computes a "winner" that is different from the correct (adult) output form (as perceived by the learner), look for the offending crucial mark that the current hypothesis incurs on the correct output form, and move the responsible constraint down by a small step along the continuous ranking scale." (14.6a)

However, the following algorithm will be seen to be more realistic and show faster recovery from errors:

Maximal Gradual Learning Algorithm. "If the learner's current grammar computes a "winner" that is different from the correct (adult) output form (as perceived by the learner), move the rankings of the constraints violated in the correct adult output form (after mark cancellation) down, and move the rankings of the constraints violated in the incorrect learner's output form up, by a small step along the continuous ranking scale." (14.6b)

These simple algorithms will prove to have the following desirable properties:

(1) The algorithms are ***convergent***: they can learn all OT grammars, from any initial state, without ever getting trapped in local maxima.
(2) The learning process is ***conservative***: the hypothesis is changed minimally on each step. This ensures the ***stability*** of the learning process: most of the learned relations are remembered.
(3) The algorithms are ***local***: only the wrong winner and the correct output form have to be compared. This keeps the necessary computational load within bounds.
(4) The algorithms are ***oblivious*** of previous evaluations of the data: all experience is laid down in the ranking value of every constraint along the scale. This keeps the acclaimed memory resources within bounds.
(5) The algorithms are ***robust***: if the reranking step is small with respect to the average distance between adjacent constraints on the scale, errors in the input have a small impact as long as they stay a minority; if they do happen to worsen the grammar, correct data received later will remedy this situation.
(6) Since the ranking scale is continuous, the algorithms can model the decrease in ***plasticity*** that comes with the years: for older learners, the reranking steps get smaller.
(7) As we will see, a probabilistic interpretation of constraint distances (in the case of the Maximal algorithm) covers several real-life phenomena: closely ranked constraints can sometimes be reversed during evaluation. This causes a ***repulsion*** of crucially ranked pairs (and thus increases robustness) and an ***attraction*** of pairs connected by mispronunciations or misperceptions.
(8) The algorithms can easily be made to allow the input to the grammar to be the overt adult form, until the learner derives underlying forms from the patterns that she discovers in the language
(9) The algorithms can easily be made to actively preserve markedness relations (i.e., honour the local-ranking principle) until the phonology gets really symbolic.
(10) The algorithms can easily model ***unlearning***: gestural constraints can drift up, faithfulness constraints can drift down.

The distinction between articulatorily oriented and perceptually oriented constraints leads to a straightforward account of learning in six phases.

14.2.3 Three production modes

In all learning phases, I will distinguish three modes of sound production:

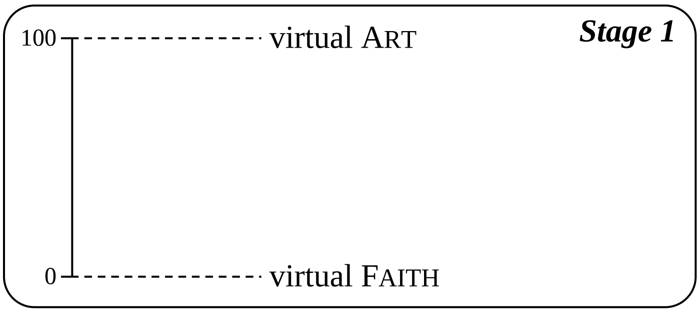
(1) Sound production for the sake of ***sound production***. In this “playing” mode, the child learns the relation between her articulations and their acoustic-perceptual results.
(2) Sound production for the sake of ***communication***. The child’s perception of the adult model utterance is the underlying form, and normal faithfulness constraints apply. The child’s speech, like that of the adult, is the result of an interaction between articulatory and faithfulness constraints.
(3) Imitation for the sake of ***imitation***. In this mode, the child learns to produce articulatory gestures that she does not yet use in her speech. Another person’s utterance is the underlying form; faithfulness is ranked higher than in normal speech,

> because the very goal is imitation, not communication (even adults rerank their faithfulness constraints according to the pragmatics of the situation: higher in addressing a crowd, lower in saying an everyday prayer). Also, articulatory constraints may be ranked lower in imitation, because the child need not give any attention to semantics, syntax, or the lexicon (and even adults rerank their gestural constraints, e.g. they raise them when intoxicated).

I will now identify six observable ***stages*** in the acquisition of (auto-)segmental phonology, and the five developmental ***steps*** by which the learner goes from one stage to the next. These stages are not to be identified with the stages of the development of vocal production (uninterrupted phonation, interrupted phonation, articulation, prosody, babbling, words) in the baby's first year (MacNeilage, 1997; Koopmans-Van Beinum and Van der Stelt, in press), since I am focusing on the development of linguistic communicative performance. Nor is the ordering presented here meant to be fixed or universal in the sense of Jakobson's (1941) proposal; my main point will be that the acquisition of articulatory coordinations and perceptual categorizations naturally evolves into an adult grammar of learned functional constraints.

14.2.4 Stage 1: an empty grammar

Acquisition starts with a stage of unlearned perceptual categorization and articulatory coordination. No real constraints exist yet, because no gestures and no categories have been learned. However, we can say that unlearned gestures would pose maximum difficulties for the speaker, i.e. the virtual articulatory constraints ("ART", typically *GESTURE) are undominated. Likewise, we can say that unlearned perceptual categories pose maximum difficulties for the listener, i.e. the virtual faithfulness constraints ("FAITH", typically *DELETE) are at the bottom of the hierarchy. We can illustrate this situation with the following virtual constraint hierarchy:



(14.7)

This is the true initial state of the learner: a reservoir of latent articulatory constraints at the top, and a reservoir of latent faithfulness constraints at the bottom: the child will not be able to recognize or produce speech. In such a grammar, no constraints are visible: it is truly a ***tabula rasa***; its substance will have to be learned later.

The learner may well be in stage 1 for a certain phonological feature and in a later stage for another. For instance, a child may have acquired some vowels and coronal stops, but not yet the distinctive sibilancy of /s/ nor, of course, its rather involved articulatory implementation. Using an OT tableau as a descriptive device for this situation, we can represent the child's handling of the adult English form [siː] 'see' as

[siː] /tiː/=|tiː|	*DELETE (coronal)	*GESTURE (blade: close & open)
√ ☞ [tiː] /tiː/		*

(14.8)

The adult produces [siː], which is a language-independent IPA shorthand for "lung pressure (= release diaphragm and external intercostals), plus fronted and high tongue body (= pull genioglossus and lower longitudinals), plus tongue-tip grooving, adduction, and opening (= upper transverse tongue fibers etc.), plus glottal adduction (= interarytenoids etc.)", as well as for its automatic acoustic result. The child perceives /tiː/, which is short for "coronal (= high-frequency noise), plus high front vowel (= low-F_1 high-F_2 periodic)". Until further notice, the learner's underlying form, now |tiː|, will be equal to this perceived input (as shown by the equal sign). The child generates an articulatory candidate [tiː] and perceives this as /tiː/, fully faithful to her underlying form (no mark for *DELETE) as well as to the perceived adult form (hence the check mark: no learning). The child does not generate a candidate [siː], because not even the perceptual input contains a sibilant.

14.2.5 Step 1: perceptual categorization and faithfulness constraints

Step 1 in the perception grammar: acquisition of perceptual categorization.
The nativist idea of the universality of phonological feature values is widely held (e.g. Hale & Reiss, to appear; Gussenhoven & Jacobs, to appear). It has been found (Eimas, Siqueland, Jusczyk & Vigorito 1971; Streeter 1976) that very young infants categorize the voice-onset-time (b-p-p^h) continuum according to criteria that reflect the clustering of plosives that Lisker & Abramson (1964) found in the languages of the world (Cho & Ladefoged 1997 found a much more continuous VOT distribution); this capability is lost in adults, who typically recognize only the contrasts that occur in their language (Abramson & Lisker, 1970). However, I will take the less expensive stance that the learner of speech must get by with independently needed strategies of perceptual and cognitive categorization. First, voice-onset-time is not a homogeneous perceptual dimension: it is difficult to regard the [b]-[p] contrast (vocal murmur versus silence) as continuously related to the [p]-[p^h] contrast (sonorancy versus aspiration noise); in fact, Kuhl & Miller (1978) and Kuhl & Padden (1982) found human-like voicing categories in untrained chinchillas and macaques, suggesting that infant VOT categorization refers to general properties of mammalian audition, not to a linguistic capability specific to the human species. Secondly, for the truly continuous perceptual dimensions of vowel quality (auditory spectrum), we do not find any initial categorization in human infants, and we find clustering only in locations predicted by the hypothesis of contrast maximization (the corners of the vowel triangle, its periphery, and equal height divisions in symmetric systems); a specifically human phenomenon of decreased discriminability of vowel qualities occurred only in the immediate vicinity of language-specific vowel categories (Kuhl, 1991). Thirdly, even complicated multidimensional categorization seems not to be

restricted to the human species, since Japanese quail have been trained to develop a human-like category /d/, generalizing across different vowel contexts (Kluender, Diehl & Killeen 1987). For more discussion, see Kuhl (1979), Jusczyk (1986), Werker (1991), Jusczyk (1992), Vihman (1996), MacNeilage (1997), and Behnke (1998).

Thus, the little linguist will listen to her language environment and learn that speakers tend to centre the perceptual feature values of their utterances at certain locations along the continuous perceptual dimensions. This will lead her into constructing perceptual categories around these locations, especially when she realizes that the categories are to be associated with differences of meaning and can be used for disambiguation of speech utterances. This part of the learning process can be simulated with any neural-net classification procedure (e.g. Grossberg 1976; Carpenter & Grossberg eds. 1991; Behnke 1998) that is told to handle acoustic and lexical similarity. The process can also be described within an Optimality-Theoretic model of a perception grammar, namely as the lowering of initially undominated (virtual) *CATEG constraints (ch. 8); this approach also explains the phenomenon that far outside any language-specific categories, adults perceive the same continuous acoustic feature values as infants (Best, McRoberts & Sithole 1988).

Perceptual categorization is a prerequisite for ***lexicalization*** and the acquisition of production; for instance, the learner should have some values for the perceptual voicing feature before she can replace her unspecified lexical /b̥/ tokens with the correct choice between /b/ and /p/, and before she will start taking the trouble to practice the necessary glottal and supraglottal gestures. In the majority of cases, lexicalization of a contrast precedes its production; for instance, Amahl (Smith 1973, p. 3) merged all initial |w| and |f| into [w], but when he started to produce a correct [f], he did so ‘across the board’ in all words that had |f| in the adult language. However, the production of a contrast sometimes precedes its lexicalization: when Amahl (Smith 1973, pp. 54, 77, 97), mastering final [n], acquired the final [nd] cluster ([wɛnd] ‘friend’, [laund] ‘round’), he generalized this to words with final |n| in the adult language ([b̥aund] ‘brown’), which suggests that his lexicon did not yet reflect the adult choice between |n| and |nd|. Other examples of lexicalization lagging behind production in Amahl’s speech, can be found in Braine (1976), Macken (1980), and Vihman (1982). These facts are important to the hypothesis of functional phonology, as they lend support to the existence of a perception grammar (fig. 14.1). As Braine (1976: 495) puts it, “Auditory encoding laws (...) would have adult articulatory features on the left-hand side, indirectly specifying the acoustic input by specifying how adults make it, and the child’s auditory features on the right-hand side”. This ***perception hypothesis*** was evident in the analysis by Waterson (1971), who

> “wants to explain phonological development in terms of a gradual loosening of constraints on the complexity of internal lexical representations. Permitted complexity constraints are in turn assumed to reflect limitations on what the child is capable of perceiving linguistically, at any given time.” (Queller 1988: 465)

I will take this loosening of constraints literally as OT-style demotion of categorization constraints. The *linguistically* is crucial here: even if the child could hear a certain difference, it is often advantageous (as with Labovian near-merger: §0.2.5, §17.1.4) to ignore it in the communicative situation.

Step 1 in the production grammar: the emergence of faithfulness constraints. The learner will intend not only to perceive, but also to produce the carriers of meaning (or meaning *differences*) in her language environment. Therefore, she will ***create*** a set of faithfulness constraints (against deletion, insertion, and replacement) for the newly categorizable feature values, or combinations of these. These constraints evaluate the relation between the underlying form (still the adult model, as categorized by the child) and the perceptual output (the sound of the child's utterance, as categorized by herself).

Since the child cannot yet produce the newly developed category, say the sibilant noise of /s/, the learning tableau is especially simple:

[siː] /siː/=|siː|	0 *DELETE (sibilant)
☞ [tiː] /tiː/	←*

(14.9)

Compared to (14.8), the perceptual input has now been enriched with the feature [sibilant] and is rendered as /siː/, and the underlying form has changed to match this perceptual input. The learner now seems to notice two things: her output /tiː/ is different from her underlying form |siː|, giving a *DELETE violation, and her output is also different from the perceptual input /siː/, resulting in an ***error*** that should lead to a learning step. Note that as the perceptual specification is still intimately connected with the adult model, a *DELETE violation and an output mismatch cannot be distinguished; indeed, they are one and the same thing, and so are unfaitfhulness and errors.

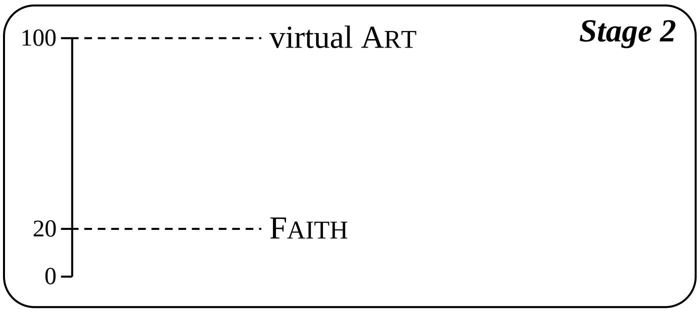
The error will lead to the learning step of promoting the violated constraint. If the initial ranking is 0, as in (14.9), the new ranking will be above 0: the learner will modify the strengths of some neural connections by an amount proportional to the degree of ***plasticity*** of the neural network. If the current plasticity is 1 on our constraint-ranking scale, the learner's single error-correction step will have the consequence of moving the violated constraint by an amount of 1 up the ranking scale.

Restricted generator. Like (14.8), but for a different main reason, tableau (14.9) still does not even *generate* the faithful utterance [siː] as a candidate, because this would require the learner to use a coronal grooving gesture that she has not yet linked to the sound of /s/. This point may require some elaboration. One might think that GEN should be able to generate an output candidate that contains [s], and that the constraint *GESTURE (tongue blade: groove) would be needed to reject that output. According to §14.1, however, GEN generates articulatory implementations, and there is no point for a speaker in including an unlearned (virtual, latent) gesture in the set of candidates. This is an example of the difference between ***universality*** and ***innateness***, two terms that have often been confounded. In the style of Ellison (to appear):

Universality versus innateness 1: GEN. "If GEN is a model for the speaker's generation of output candidates, it will generate no candidates with unmastered articulatory gestures. If, on the other hand, GEN is a device for free hypothesization by linguists, it may generate any articulation that linguists can think of; but this universality will not reflect innateness." (14.10)

14.2.6 Stage 2: violated faithfulness

During the acquisition of its categorization, the faithfulness constraints for a language-specific feature value percolate up from the bottom of the grammar. However, the corresponding articulatory constraint is still found at the top of the hierarchy, because the learner's motor control has not yet mastered the relevant gestures. This situation can be pictured as



(14.11)

This hypothesis is consistent with the proposal by Gnanadesikan (1995), Smolensky (1996b), and others, that in the initial state markedness (or structural) constraints outrank faithfulness constraints, but inconsistent with Hale & Reiss (1996, to appear), who propose that the reverse holds true (see §14.2.7).

Evidence that articulatory constraints crucially outrank faithfulness constraints at this stage (and of the temporal order of perceptual vs. articulatory acquisition), is found in the phenomenon that children tend not to accept an imitation of their own speech by an adult (the 'fis' phenomenon: Berko & Brown 1960: 531). For instance, if the child produces [diː] for the English utterance /siː/ 'see', she may still object to her father pronouncing this word as [diː], a fact that shows that the child's perceptual target must already be /siː/, and that her output [diː] is an unfaithful rendering of it (tableau 14.9).

14.2.7 Step 2: sensorimotor learning

With high faithfulness constraints, but still higher articulatory constraints, the learner experiences a lot of tension in her system: an average utterance will come out so unfaithful that listeners have trouble understanding her. The acquisition of motor skills will remedy this situation. The articulatory constraints are lowered:

> "the child's "tonguetiedness," that overwhelming reality which Stampe and Jakobson both tried to capture with their respective formal structures, could be handled more felicitously if one represented the heavy articulatory limitations of the child by the formal device of output constraints [...]. The child's gradual mastery of articulation then is formalized as a relaxation of those constraints" (Menn 1980: 35-36)

Again, I will take this relaxation of constraints literally in OT terms.

The learner will be able to learn the relation between articulation and perception with the help of articulatory variation. The learner can interpolate and extrapolate, but because of the many non-linear relationships between articulatory and perceptual parameters, she will also have to ***play*** a lot.

Once the learner knows that she can produce something that she will perceive as /s/, the constraint *[sibilant] will enter her grammar, perhaps at the height of *GESTURE (blade: groove), if that is the way she chooses to implement this sound. If the learner is a child that has not practised the blade-grooving gesture before, the articulatory constraint will enter at the top of the grammar (with a ranking value of, say, “100”), still resulting in the unfaithful output /tiː/:

[siː] /siː/=|siː|	100 *[sibilant]	20 *DELETE (sibilant)
☞ [tiː] /tiː/		←*
√ [siː] /siː/	*!→	

(14.12)

In this tableau, the finger points to the learner’s output (her production, as perceived by herself), and the check mark identifies the form that the learner assumes to be correct: the adult output form, as perceived by the learner. Thus, tableau (14.12) results from a ***pair*** of utterances: one by the adult, one by the learner. Given the learner’s capabilities of perceptual categorization, we can assume that she notices the discrepancy.

With the Minimal learning strategy, the learner would tackle the largest problem, getting get rid of the offending crucial mark (the exclamation mark in the row with the check mark), i.e. moving the offending constraint *[sibilant] down the ranking scale, perhaps by practising the grooving gesture. Practice will modify the strengths of some neural connections by an amount proportional to the degree of plasticity of the neural network. If the current plasticity is 1 on our constraint-ranking scale, the learner’s single error-correction step will have the consequence of demoting the offending constraint by 1:

[siː] /siː/ |siː|	99 *[sibilant]	20 *DELETE (sibilant)
☞ [tiː] /tiː/		*
√ [siː] /siː/	→*!	

(14.13a)

With the Maximal learning strategy, on the other hand, the learner will tackle the problem not only by moving *[sibilant] down, but also by moving *DELETE (sibilant) up:

[siː] /siː/=|siː|	99 *[sibilant]	21 *DELETE (sibilant)
☞ [tiː] /tiː/		*←
√ [siː] /siː/	→*!	

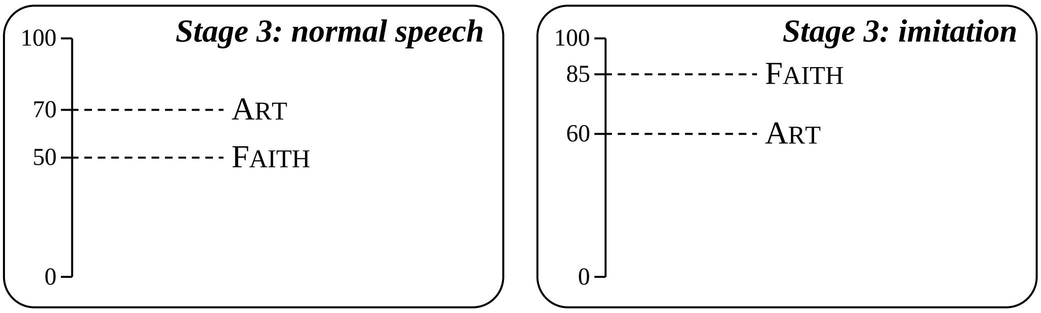
(14.13b)

Until §14.7, I will assume the Maximal algorithm, since the Minimal version will prove incapable of handling optionality (§14.2.12; ch. 15).

Now, fundamental questions arise: do these tableaux show an output mismatch or a *DELETE violation? Is the remedy to practise a gesture or to demote *GESTURE? In other words, does the learner acquire a production system or a grammar? From the UG standpoint, these have been clearly distinguished: Hale & Reiss (1996, to appear) would argue that in (14.12), faithfulness actually has to dominate the structural constraint because the learner perceives /siː/, and knows that the output should be /siː/; so /siː/ would be the output of the ***grammar***, and /tiː/ the output of the ***body***. The idea is that the recognition process mirrors the production process, so that they should share (a part of) the grammar, and the up and down arrows in (14.1) should be intimately connected (for the top part). This may be true for the more "lexical" parts of phonology, but if we see the grammar as a description of human behaviour, we should include the more superficial parts that allow functional explanation, all the way up to the actual articulation. Moreover, the procedure of ***robust interpretive parsing*** (Smolensky, 1996a; Tesar and Smolensky, 1996; Tesar, to appear) allows initially high-ranked structural constraints in a grammar that can be used for production as well as for comprehension.

14.2.8 Stage 3: faithful imitation

After some practice (e.g., 30 unit steps as in (14.12)), the articulatory constraints become so low, and faitfhulness constraints so high, that a special situation of FAITH raising and ART lowering would result in a faithful rendering of the perceptual target, though during normal speech production the articulatory constraint still dominates:



(14.14)

This mode-dependent constraint reranking counters a certain criticism against ascribing the large input-output disparity to performance difficulties:

> "claiming that children don't produce, say, a particular segment because their motor control hasn't yet mastered it, can run afoul of the fact that children who systematically avoid a given structure in their linguistic productions can often easily imitate it." (Smolensky 1996a, p. 720)

If we accept the possibility of pragmatics-based reranking, the argument vanishes. What's more, we have a second explanation (after play) for an otherwise awkward bootstrapping problem: how could we know that we should practise a certain gesture to drag down a constraint, if we do not know that its demotion will result in a more faithful grammar? Answer: we perform an experiment by temporarily ranking a faithfulness constraint above it, and this experiment results in a correct output (which, still equivalently, satisfies *DELETE):

imitation of [siː] /siː/=|siː|	85 *DELETE (sibilant)	60 *[sibilant]
[tiː] /tiː/	∗!	
√ ☞ [siː] /siː/		∗

(14.15)

Thus, free variation of faithfulness adds to the learner's confidence in her choice of articulatory gestures to practise.

14.2.9 Step 3: the learning curve

Some practice of producing the new gesture (perhaps aided by a few faithful imitations) will dramatically facilitate its implementation, and some practice of perceiving the new category will raise its importance in communication. These events draw the anti-gesture constraint and the relevant faithfulness constraint into each-other's vicinity:

[siː] /siː/=|siː|	61 *[sibilant]	59 *DELETE (sibilant)
☞ [tiː] /tiː/		←∗
√ [siː] /siː/	∗!→	

(14.16)

The result is still unfaithful (= incorrect). The next learning step will bring the two constraints on an equal height. Two synaptic strengths are never exactly equal, so we should assume that this situation gives rise to two possible evaluations, a faithful one and an unfaithful one, both with a probability of 1/2. The faithful (= correct) evaluation gives

[siː] /siː/=|siː|	60 *DELETE (sibilant)	60 *[sibilant]
[tiː] /tiː/	∗!	
√ ☞ [siː] /siː/		∗

(14.17)

and the unfaithful (and incorrect) evaluation gives

[siː] /siː/=|siː|	60 *[sibilant]	60 *DELETE (sibilant)
☞ [tiː] /tiː/		←∗
√ [siː] /siː/	∗!→	

(14.18)

Because the learner will still interpret half of her utterances as incorrect, another learning step will soon follow, giving a stable grammar:

[siː] /siː/=|siː|	61 *DELETE (sibilant)	59 *[sibilant]
[tiː] /tiː/	∗!	
√ ☞ [siː] /siː/		∗

(14.19)

Learning has succeeded, no errors are to be expected, and the grammar will not change any further. Note that learnability along this discrete ranking scale prohibits the existence of ***ties***, in the sense of constraints with equal harmonies that pass the buck to lower-ranked constraints:

No ties: the marks incurred by a constraint can never be cancelled by the marks incurred by a different constraint with the same ranking. (14.20)

Stochastic evaluation. The process described above shows some unrealistic behaviour: the learner has a 0% correct score for some time, then a 50% correct score for a very short time, and then a 100% correct score for the rest of the time. Real learning shows a much smoother behaviour. We therefore interpret constraint ranking in a probabilistic manner. With a neural-net analogy, the loudness of the protest of a constraint is the value of an inhibitory postsynaptic potential: it depends on the synaptic strength (the ranking as specified in the grammar) as well as on some things like the incidental amount of locally available neurotransmitter. At evaluation time, therefore, the disharmony (the "effective" ranking) of the *i*-th constraint C_i is something like

$$disharmony\ (C_i) = ranking\ (C_i) + rankingSpreading \cdot \mathbf{z} \qquad (14.21)$$

where **z** is a Gaussian random deviate with mean zero and standard deviation 1. Now the problem of ties automatically vanishes: the probability that the disharmonies of two different constraints are equal, is zero. The learning algorithm is very resilient against the actual value of *rankingSpreading*, but in the examples of this chapter, I will take it to be 2.0.

We will also have fuzzy reranking steps. In the above examples, the plasticity was held constant at 1 per learning step. A more realistic view holds that it contains a noise component, and that it decreases with the age of the learner (children learn and unlearn some things faster than adults):

$$plasticity = plasticity_0 \cdot (1 + relativePlasticitySpreading \cdot \mathbf{z}) \cdot \left(\frac{1}{2}\right)^{age/plasticityHalfTime} \qquad (14.22)$$

In this formula, $plasticity_0$ is the ***day-one plasticity***: the value on the first day. As far as learnability is concerned, *relativePlasticitySpreading* may be anything, including zero, but I will keep it fixed at 0.1. Finally, *plasticityHalfTime* is the time needed for the plasticity to decrease by a factor of two: if day-one plasticity is 1, and the plasticity half-time is 1500 days, the plasticity will have decreased to 1/2 at an age of 1500 days, and to 1/4 at an age of 3000 days; the advantage of this decrease is that young learners learn fast, and older learners learn more accurately. For this chapter, I will ignore this exponential decrease in plasticity, setting *plasticityHalfTime* to an infinite value.

Now, with a ranking spreading of 2.0, the grammar of tableau (14.19) is not yet stable. When it comes to an actual evaluation, *DELETE (sibilant) may be ranked at, say, 61 – 1.34, and *[sibilant] may be ranked at 59 + 0.78 or so, which is higher. This will result in an incorrect (= unfaithful) output again:

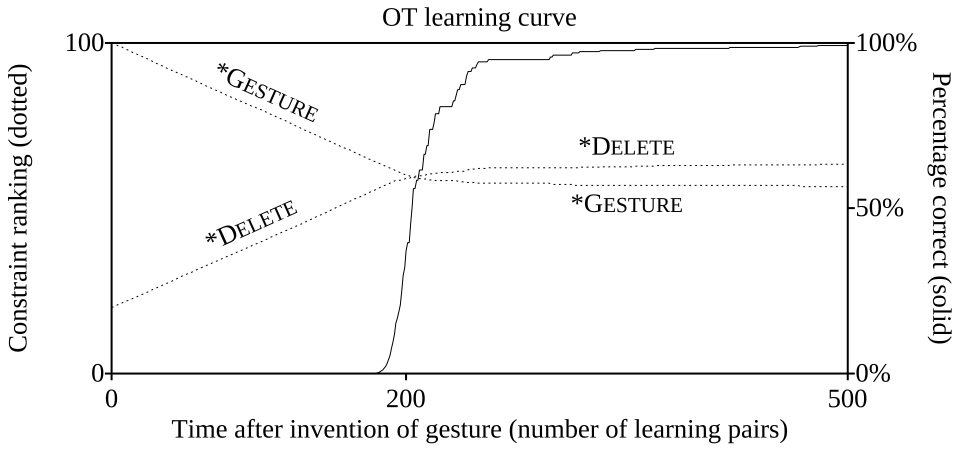
[siː] /siː/=|siː|	59 + 0.78 *[sibilant]	61 – 1.34 *DELETE (sibilant)
☞ [tiː] /tiː/		←*
√ [siː] /siː/	*!→	

(14.23)

So the gestural constraint is demoted again, and the faithfulness constraint is raised. As the distance between the two constraints increases, the probability of a reversal decreases, and the reranking will nearly come to a halt when *[sibilant] has fallen below 55, and *DELETE (sibilant) has risen above 65:

Constraint repulsion. "A crucially ranked pair of constraints repel one another and end up maintaining a safety margin." (14.24)

Simulation (14.25) shows how the ranking of *[sibilant] becomes lower as a function of time (= the number of received /s/ data), and how the percentage of correct productions rises smoothly from 0% to 100% (with a plasticity of 0.2).[2]



(14.25)

We see that the 50%-correct point is reached only a short time after the first correct utterance, but that it takes a relatively long time after that before speech becomes flawless

[2] The algorithm for the rankings G and D of the two constraints *GESTURE and *DELETE is:

```
$G[0] := 100$; $D[0] := 20$; $p := 0.2$;      (":=" is the assignment operator)
for $t$ from 1 to 1500
    if $G[t-1] + 2\cdot\mathbf{z} > D[t-1] + 2\cdot\mathbf{z}$
    then $G[t] := G[t-1] - (1 + 0.1\cdot\mathbf{z})\,p$;  $D[t] := D[t-1] + (1 + 0.1\cdot\mathbf{z})\,p$
    else $G[t] := G[t-1]$;  $D[t] := D[t-1]$
```

Realizing that the spreading of the difference of two Gaussian distributions with spreadings of 2 is $2\sqrt{2}$, we can compute the percentage correct at time t as $1/2\cdot(1 - \mathrm{erf}\,(\sqrt{2}/2\cdot (G[t] - D[t]) / (2\sqrt{2})))$.

(even with constant plasticity), because the probability that a datum triggers a grammar change decreases.

Figure (14.25) shows the learning curve as seen from the learner. From the point of view of the adult, there are two complications. First, the gesture-invention day cannot be seen. Secondly, whether the learning curve rises as steeply as in the figure, depends on whether the learner has had enough time since her categorization day to lexicalize all occurrences of /s/.

Continuous constraint families. Even with fuzzy ranking and demotion, the /siː/ example has a knack of discreteness. After all, it is not so important *whether* we can produce sibilant noise, but *how similar* our output is to sibilant noise. For instance, the candidate [θiː] would be more faithful than the candidate [tiː] because it has a greater chance of being categorized as /siː/. With discrete *DELETE constraints, we would say

[siː] /siː/=|siː|	*GESTURE (blade: groove)	*DELETE (noise)	*GESTURE (blade: protrude)	*DELETE (sibilant)
[tiː] /tiː/		*!	*	*
☞ [θiː] 40% /siː/			←*	
√ [siː] /siː/	*!→			

(14.26)

but a more realistic account would involve a continuous noisiness scale. If course, there is a universal local ranking *DELETE (noise) >> *DELETE (sibilant), or, more explicitly locally ranked: *DELETE (noise: ≥ mellow) >> *DELETE (noise: ≥ sibilant) “it is more important to have at least mellow noise than to have at least sibilant noise”. If you cannot get it all, settle for a little less.

This is an example of ***acoustic faithfulness***: the speaker minimizes the confusion probability by seeking maximum distances between sounds along a continuous acoustic dimension. Generally, the more effort we spend, the less confusing (unfaithful) we are. The optimum is found where the problem of articulatory effort *equals* the problem of confusability (chapter 10).

We also understand now why /s/ is the universally unmarked fricative: *not* because it would be easier to make than /θ/, but because it sounds better: the alveolar fricative is 10 dB louder than the dental fricative.

Probabilistic categorization. As another example, consider the acquisition of the voicing contrast in plosives. When a Dutch learner already has correct categories /d/ and /t/, but does not yet master the voicing and devoicing gestures, both inputs will be pronounced with neither gesture, typically as a lenis voiceless plosive [d̥] (like English initial “d”), which may be perceived (by the adult, and also already by the learner) as the fortis voiceless plosive /t/ in, say, 70% of all cases, and as the lenis voiced plosive /d/ in the other 30% of the cases. Because the majority of utterances will give /ta/, the learner will probably evaluate her candidates as if [d̥] would result in /t/.

Consider the input /ta/. Before production, the learner will evaluate the main candidates as

[ta] /ta/=|ta|	*[–voiced / plosive]	*DELETE (–voice)
√ ☞ [d̥a] /ta/		
√ [ta] /ta/	∗!	

(14.27)

In 70% of the cases, the winner's perceptual output /ta/ equals the prediction of this tableau and also equals the correct adult form, so these cases will cause no change in the grammar (= would not force learning of a devoicing gesture). In the remaining 30% of the cases, however, the learner will perceive the actually produced output as /d/. This will call for a reanalysis, now with a violation of *DELETE (–voice):

[ta] /ta/=|ta| reanalysis	*[–voiced / plosive]	*DELETE (–voice)
☞ [d̥a] /da/		←∗
√ [ta] /ta/	∗!→	

(14.28)

The winning articulatory candidate [d̥] has not changed. However, it is no longer equal to the correct adult form, so it will induce reranking. The net result of a large number of /ta/ input data is the demotion of the gestural constraint past *DELETE (–voice), caused by the simultaneous acquisition of a devoicing gesture and strengthening of the percpetual voicing feature. Thus, error-driven learning manifests itself as minimization of confusion probabilities.

In general, error-driven learning is used by humans to increase the reproducibility of the external results of their motor actions. In phonology, it acts as a local strategy for implementing the globally defined functional principle of minimization of confusion by way of perceptual invariance.

Now consider the input /da/. The learner will initially evaluate the candidates as

[da] /da/=|da|	*[+voiced / / plosive]	*DELETE (plosive)	*DELETE (+voice)	*[+voiced / / –obstruent]
☞ [d̥a] /ta/			←∗	
√ [da] /da/	∗!→			
[na] /na/		∗!		∗

(14.29)

For purposes of illustration, this tableau shows the candidate [na], which represents a way of faithfully parsing the [+voice] specification. If *DELETE (plosive) (or *INSERT (nasal)) dominates *DELETE (+voice), this candidate cannot win. Note that the constraint *[+voiced / –obstruent] is still ranked very low; in fact, it is universally ranked below

*[+voiced / plosive] because of the monotonically decreasing relation between the ease of phonation and the degree of supralaryngeal constriction.

The result of (14.29) will give rise (under Minimal GLA) to a demotion of *[+voiced / plosive], because that is the constraint that incurs the crucial mark (the column with the exclamation mark “!”) for the correct form (the row with the check mark “√”); or it will give rise (under Maximal GLA) to a promotion of all the constraints violated in the learner’s form, i.e. *DELETE (+voice), and a demotion of all constraints violated in the adult form, i.e. *[+voiced / plosive]. In 30% of the cases, however, the produced result will be analysed as /da/, contrary to the predicted output /ta/ (in the row with the pointing finger), and the reanalysis will give

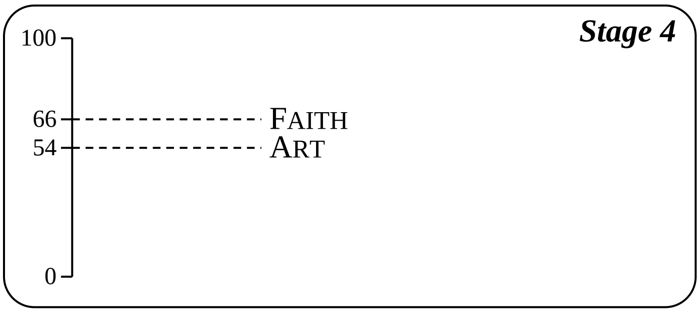
[da] /da/=|da| reanalysis	*[+voiced / / plosive]	*DELETE (plosive)	*DELETE (+voice)	*[+voiced / / –obstruent]
√ ☞ [d̥a] /da/				
√ [da] /da/	*!			
[na] /na/		*!		*

(14.30)

This reanalysis will cause the output to be equal to the correct form /da/, and no demotion of *[+voiced / plosive] will occur.

14.2.10 Stage 4: faithfulness outranks gestural constraints

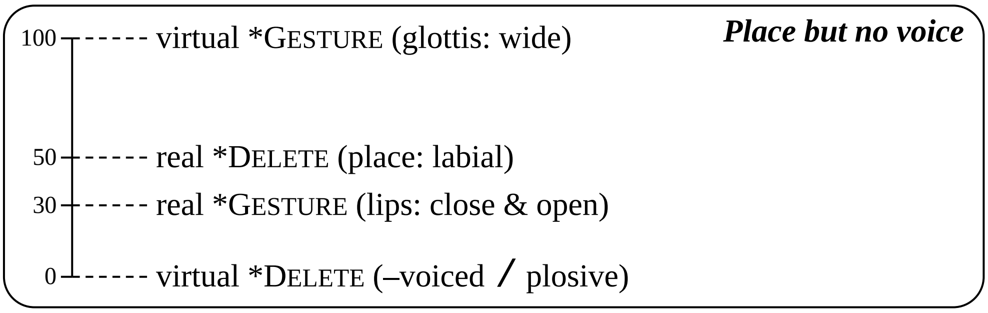
With faithfulness still in the same position as in stage 2, the output is now equal to the specification, which still equals the perceived adult model:



(14.31)

All relevant faithfulness constraints outrank all relevant articulatory constraints. Learning seems to have succeeded. There are no *DELETE violations, because the learner’s perceptual output equals her underlying form. And there are no output mismatches (learning triggers), because the learner’s perceptual output equals her perceptual input (the overt adult model utterance, as perceived by the learner). As the underlying form is still identified with the perceptual input, these two statements refer to the same phenomenon. Until now, we have used the powerful device of an OT tableau with five representations just for massaging a grammar into a state of pronounceability. We will need its full resources when grammar (14.31) will have to change by the only way it can ever change: breaking the identity between the perceptual input and the underlying form.

Timing. The learning phases may have different timing for different features. If place contrasts are mastered, but voicing contrasts are not, we have the following ranking:


Place but no voice
100
virtual *GESTURE (glottis: wide)
50
real *DELETE (place: labial)
30
real *GESTURE (lips: close & open)
0
virtual *DELETE (–voiced / plosive)

(14.32)

Phonation. Some innate sound-producing articulations can be used for communicative purposes exceptionally fast: crying requires glottal adduction, so the articulatory constraint *GESTURE (glottis: adduct) must be low at the time of the first steps into language. As soon as the perceptual features [voiced] and [sonorant] have been acquired, the implementational constraints *[+voiced / –obstruent] and *[sonorant / –obstruent] must be low, because *GESTURE (glottis: adduct) is low.[3] Note that the voicing of non-obstruents is not automatic; it is just **easy** because the necessary gesture (controlled interarytenoid activity to move away from the neutral breathing position of the vocal folds) is mastered early. For phonation, therefore, step 1 will immediately lead to stage 4.

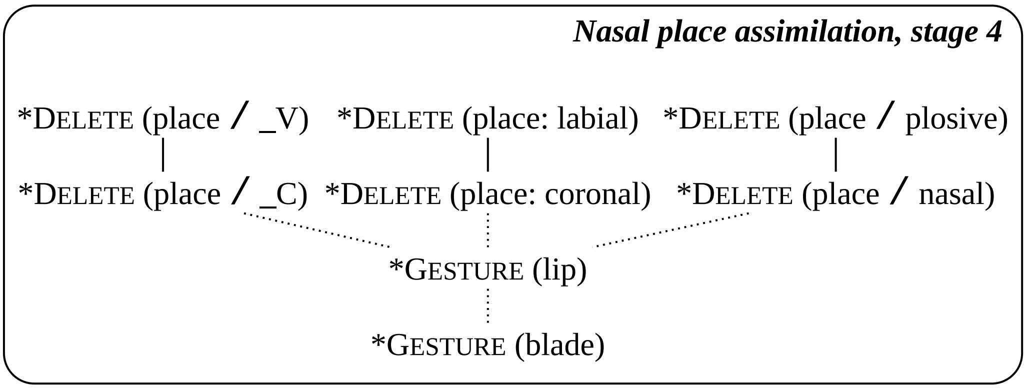
14.2.11 Step 4: sentence-level phonology

The form of a word in isolation is usually acquired earlier than the form of a word in the sentence. Since the learning strategy has involved a fairly high ranking of faithfulness (in order to overcome the articulatory problems), we can expect the child to perform overly faithful in stage 4. For instance, Dutch four- or five-year-olds tend not to implement adult sentence-level phenomena like degemination and nasal place assimilation.[4] In these adult processes, certain position-dependent faithfulness constraints fall below gestural constraints as a result of the low perceptual contrast between [np] and [mp] or between [pp] and [p], which admits the replacement of the faithful form by an articulatorily easier candidate. For the learner, this means that she will eventually lower some faithfulness constraints.

As an example, consider the acquisition of the place assimilation of Dutch /n/ to following labial or dorsal consonants (plosives do not assimilate; labials and dorsals do not assimilate). In stage 4, the grammar is

[3] I'd like to reserve the term *sonorant* for the perceptual feature that refers to full periodicity (voicing) with clearly defined spectral components (formants). I use *obstruent* as a cover term for all the articulations that do not allow air to flow freely between the upper larynx and the outer air.

[4] Judging from my own observation of the speech that prevails in my children's school classes.



(14.33)

This grammar already shows four local rankings that the learner can be assumed to know, based on the dependence of the confusion probability on the availability of acoustic cues (for position-dependent *Delete and manner-dependent *Delete; see Jun 1995), on an adaptation by the listener to asymmetries in frequency of occurrence (for place-dependent *Delete; see §9.5), and on an alleged effort difference between two articulators (reflecting a markedness relation identified by Prince & Smolensky 1993: §9).

In this grammar, the learner will produce an underlying |an#pa| as [anpa], and she will categorize this as /anpa/. Previously, the learner may only have taken isolated word forms as evidence for learning phonology. From the moment that the learner takes into account sandhi phenomena, an error will be generated when the learner hears that an adult speaker uses something that the learner perceives as /ampa/ (we know that the child can hear the difference, because some languages do assimilate and some don't). Instead of questioning her underlying form (which in isolation would be a correct word form), the learner will signal an offending faithfulness constraint in her analysis:

[ampa] /ampa/ |an#pa|	*Delete (place / nas)	*Delete (place: cor)	*Delete (place / _C)	*[lab]	*[cor]
☞ [anpa] /anpa/				*	←*
√ [ampa] /ampa/	*!→	*→	*→	*	

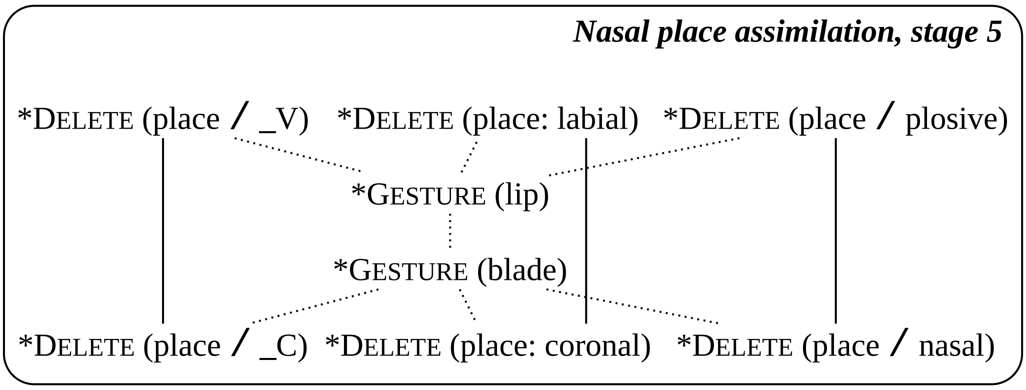
(14.34)

With *Delete violable, the connection between perceptual input and underlying form has been severed, as has the link between faithfulness and correctness.

Three *Delete constraints incur marks in the correct form. With the Minimal algorithm, the crucial mark is incurred by the constraint that happens to be the highest. In this example, *Delete (place / nasal) would be demoted, and the violations in the grey cells would be ignored.[5] This will happen again and again, until *Delete (place / nasal) falls below *Delete (place: cor); at the next error, *Delete (place: cor) will be demoted. The two *Delete constraints will tumble down in turns, until they come past *Delete (place / _C), which will then be demoted at the next error. The three

[5] The error-driven constraint demotion (EDCD) algorithm of Tesar & Smolensky (1996) would immediately drop all three *Delete violators in tableau (14.34) below the constraint that incurs the highest mark in the winner that is not cancelled by a corresponding mark in the output; therefore, below *[coronal]. The EDCD step, therefore, changes the grammar in such a way that the correct adult form becomes more harmonic than the learner's original winner. The properties of the two algorithms are compared in §14.5.

constraints, now at approximately equal heights, will trickle down the hierarchy until they have fallen some safe distance below *[cor], after which they will stop; §14.7 will show that the whole procedure converges to the correct ranking. The resulting grammar will be:



(14.35)

With the Maximal algorithm, on the other hand, all three *DELETE constraints that incur uncancelled marks in the correct form are demoted; the one constraint has an uncancelled violation mark in the learner's winner (*[cor]) is promoted. The three *DELETE constraints will trickle down the hierarchy, and *[cor] will trickle up, until the three have fallen some safe distance below *[cor]. If *[cor] rises above *[lab] and this is inappropriate, other learning steps will raise *[lab] above *[cor] again; §15.4 will show that the whole procedure converges to the correct ranking: in this case, (14.35).

Thus, both gradual algorithms converge (in the case of non-variable data). In grammar (14.35), all the local rankings have been preserved[6].

14.2.12 Stage 5: alternating levels of constraints

We now have virtual ART >> real FAITH >> real ART >> real FAITH >> virtual FAITH. Ignoring the latent faithfulness constraints, such a grammar exhibits a chain of three crucial dominations (from a universal viewpoint): it has a DEPTH of three. Phonology has now gone from the word to the sentence, and postlexical phenomena have been learned.

Optionality. The optionality of a thing like place assimilation of nasals can be explained by a pragmatically determined reranking of constraints. In a communicative situation that requires extra understandability, all *DELETE constraints may go up by a distance of, say, 20 along the continuous ranking scale, and several *GESTURE constraints will fall prey to this rising faithfulness. However, markedness relations are preserved here, too: high-ranked *GESTURE constraints and low-ranked *DELETE constraints will experience a relatively low degree of optionality.

Errors. There is a complication with the Minimal algorithm here. If an adult speaks with unusually high faithfulness, the learner will have to base her analysis on an equally faithfulness-adapted evaluation. If, instead, she analyses the adult forms with her usual grammar, this will result in a discrepancy between her output and the adult model output,

[6] The local rankings actually allow a language with assimilation of /m/ and not of /n/. The non-existence of such a language must lead us into questioning the status of the alleged local ranking of the gestural constraints.

and she will consequently demote a *GESTURE constraint. This kind of demotion will also generally occur when adults make outright mistakes and the learner accepts these mistakes as correct data (we assume that if the learner notices the mistake, she will see no reason to change her grammar).

If there are many spurious constraint demotions, this will lead to errors by the learner. For instance, a spuriously lowered *GESTURE constraint may increase the number of times that a related lower *DELETE constraint is ranked above it at evaluation time. This will cause the *DELETE constraint to be pushed below the safety margin again. Under the Minimal algorithm, the net result would be that two constraints have fallen somewhat down the hierarchy, and they may be going to push down some other constraints that are crucially ranked below them. The conclusion must be that in the presence of unsignalled errors, a large part of the constraint hierarchy would be drifting down. This downdrift, which can also be a consequence of real optionality, does not occur under the Maximal algorithm (see chapter 15).

14.2.13 Step 5: emergence of underlying forms

The learner grows to see patterns in the words of the language, based on morphological alternations, and may construct more or less abstract underlying forms. The input to the learner's grammar shifts from the adult word form to this new underlying representation. The output is no longer equal to the input; as in step 4, the learner may learn that faithfulness constraints can be violated in an adult grammar, but this step will also introduce morphologically conditioned ranking, output-output constraints, perhaps suspension of local ranking, and language-dependent symbolic relations that are hard to describe with functional constraints.

For word-internal phenomena that allow descriptions in terms of functional constraints, like vowel harmony, step 5 will introduce the necessary negative evidence against certain word-internal combinations of gestures (for an example, see §14.3.5).

14.2.14 Stage 6: the adult phase

The adult phase has heavily interacting gestural and faithfulness constraints for postlexical phonology and some autosegmental lexical phenomena, and a language-specific symbolic constraint set based on alternations in the lexicon and some sentence-level phenomena. Some typical levels of constraints are, from top to bottom:

- Depth 0: unlearned gestures (virtual and other).
- Depth 1: obligatory perceptual features.
- Depth 2: difficult gestures.
- Depth 3: unimportant perceptual features.
- Depth 4: easy gestures.
- Below all: unlearned categories (virtual). (14.36)

At the top, we see the latent articulatory gestures: most of them play no role for speakers until they want to learn a new language; likewise, the latent faithfulness constraints at the bottom can be considered to be outside the grammar. These two types of virtual constraints are maximally high or low, not because they have been ranked as such, but because the speakers have no experience whatsoever with them (unless they have ghost segments in underlying forms). Therefore, these constraints have no claim to ***psychological reality***. For all practical purposes, we can assume that the speaker only uses the "real" constraints in the middle, and these have obviously emerged during the processes of perceptual categorization and motor learning:

Finiteness. "Gestural as well as faithfulness constraints are learned, not innate. Each language uses a finite set of these constraints, learned in the processes of perceptual categorization and motor learning." (14.37)

As crucial rankings deeper than four or five levels seem to be quite rare, the question arises whether a grammar consisting of crucial rankings (instead of rankings along a continuous scale) could be psychologically real and learnable. Such a grammar would be expressible with unviolated declarative constraints; the "unimportant perceptual features" of depth 3, for instance, could be reworded in terms like "pronounce except if". In this book, however, I will stay with OT.

14.2.15 Second-language acquisition

After the learning of the first language, a part of the initial state is still there: unused gestures will still be invisible to the grammar, i.e., they will be represented by undominated constraints; and unlearned categories will still be invisible, i.e., their faithfulness constraints will be ranked at the bottom; their workings in the initial stage of second-language acquisition are fully automatic.

The average Germanic adult is still in the initial stage with respect to e.g. the acquisition of ejectives. Upon hearing [k'a], she will first analyse this as /ka/ and imitate it as [ka] (stage 1); after some more exposure, she will recognize the ejective burst and categorize it as /k'a/ (stage 2), though her pronunciation will still be [ka]; after some practice, she will be able to pronounce [k'a], first in imitation (stage 3), later on in communicative situations (stage 4); at some time during the stages 2, 3, or 4, she will have lexicalized the ejective, so that after learning the patterns of the language, she may know that some underlying /k'/ tokens have positionally neutralized [k] variants (stage 5, 6).

14.2.16 Acoustic versus linguistic faithfulness

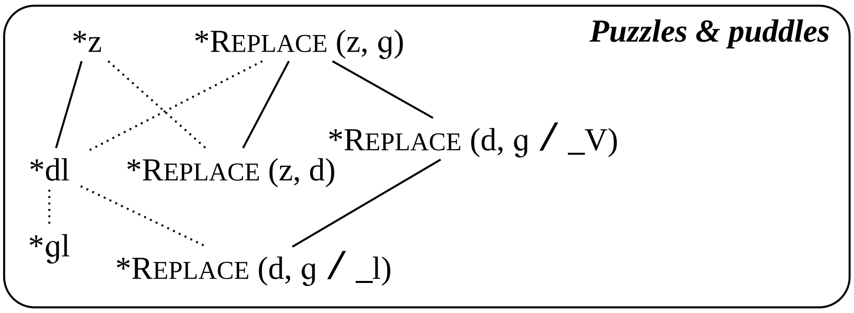
A criticism uttered by Smolensky (1996a) against the performance hypothesis (the hypothesis that performance problems account for the relative poverty of production with respect to perception), is the fact that children's replacements of certain gestures do not seem to have anything to do with articulatory problems. For instance, Smith (1973: 150) mentions a child that renders the adult /θɪk/ 'thick' as [fɪk], but not as a result of problems with the production of a dental fricative, since the same child renders the adult /sɪk/ 'sick' as [θɪk].

The solution of this problem lies in a perceptual asymmetry. Initially, the child will try to imitate the adult model as faithfully as possible, but this faithfulness will be evaluated in acoustic terms, since adaptation to different speakers must also be learned. Consider an example from a different domain: the adult utterance [ɛ] may have a first formant of 600 Hz, defining the vowel as lower-mid. If it is true that the child's initial classification system yields psychophysical rather than phonetic categories (§14.2.5), the child will reproduce this formant and produce a vowel which adults may perceive as [e] because they have learned to correct spectral structure for the length of the vocal tract of the speaker, probably with the help of the fundamental frequency; and indeed, the child has articulated a higher mid vowel. Thus, the adult utterance [i e ɛ a] is perceived by an adult as [high higher-mid lower-mid low] on her perceptual relative [vowel height] tier, and it is perceived by a young child as [250Hz 400Hz 600Hz 850Hz] on her perceptual absolute F_1 tier; in their reproductions of this utterance, both the adult and the child may be absolutely faithful in their own terms. The adult/child difference seems to reflect the usual order of the acquisition of entities versus relations in cognitive development.

The same reasoning may apply to the imitation of adult [θɪk]: the child may perceive her candidate [fɪk] as closer to the original than her candidate [θɪk] because of its somewhat lower spectral content. We thus expect chains like [ɔ] → [o] → [u] and [s] → [θ] → [f], and the child's productions must be considered as more or less equal to her underlying forms (= perceptual targets).

14.2.17 Puzzles

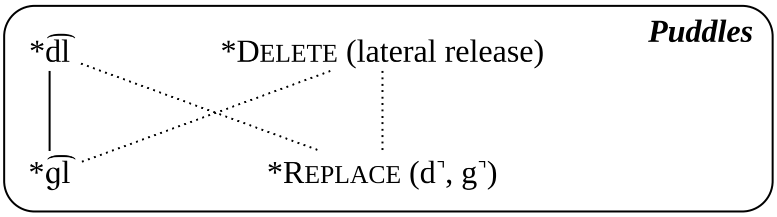
Another apparent chain shift is Amahl's (Smith 1973: 55, 149) rendering of 'puzzle' as [pʌdl] and 'puddle' as [pʌgl]. Smith's derivational phonology has to take recourse to the ***counterfeeding*** ordering of the rules d → g / _l and z → d / _l. Likewise, a naive OT account would have that if [dl] is the optimal candidate for /zl/, it would also be the optimal candidate for /dl/. However, there is a partly universal ranking of functional constraints that produces the attested facts:



(14.37a)

We can consider the ranking of *z above *dl as near-universal, because [z] requires a precise (controlled) movement of the articulator, which is not needed in the ballistic thrust of [d] (Hardcastle 1976); of course, this is the reason why Amahl (and most children) pronounced *all* |z| as [d]. The ranking of *REPLACE (z, g) above the other faithfulness constraints is also universal, because the perceptual distance between /z/ and /g/ (different place and manner) is larger than that between /z/ and /d/ (same place) or between /d/ and /g/ (same manner); this corresponds to a property of Gnanadesikan's (1997) ***ternary scales*** (which are a first cautious step in the ultimately inevitable generalization from binary hybrid features to multi-valued perceptual features). As noted by McCarthy (1998), this type of chain shift is one of the few examples of opaque rule ordering that a monostratal OT grammar can handle. The constraint *REPLACE (d, g / _l) is very low, because the perceptual distance between /dl/ and /gl/ is very small (Kawasaki 1982). Most rankings in (14.37a) are, therefore, expected to occur during the acquisition of any language. The ranking of *dl above *gl could due to the asymmetric motor learning of the gestures associated with the abundant initial |gl| and the absent initial |dl| in English. The remaining crucial property of (14.37a) is the high ranking of *REPLACE (z, g), even before [l]. The functional correlate of this high ranking is the strength of the coronal place cue in [zl]: the continuancy and sibilance of /z/ ensure a good acoustic reflex of coronality, quite differently from that of /d/, which almost vanishes before [l].

Chain shifts are often associated with the idea of contrast preservation. As fas as Amahl is concerned, Kiparsky & Menn (1977: 69) hold that he “displaced the contrast”. However, when Amahl started to pronounce |z| as [d], his ‘puzzle’ surfaced with an adult-like fricative, but he kept on merging |dl| with |gl|. Also, he always seemed more interested in faithful production than in contrast preservation: at first, he pronounced |s| as /t/ (faithful place), and later as /l/ (faithful place and continuancy). More likely, Amahl’s [dl] for |zl| is an instance of his general [d] for [z], and his [gl] for |dl| is a result of his perception of the adult [dl] as a laterally released stop. Smith’s (1973: 150) assertion that “this clearly is false” since the child distinguished adult [dl] and [gl], is only valid for those who believe in universal underlying feature values. English |dl| sequences are very limited in distribution, and it is no surprise that a child should perceive them primarily as ‘laterally released stops’ [d˺_ˡ], with a secondary place cue (the transition from the previous vowel) that distinguishes them from [g˺_ˡ], giving the ranking *DELETE (lateral release) >> *REPLACE (place: coronal, dorsal / unreleased). Thus, the relevant part of (14.37a) will look like



(14.37b)

14.3 Example: acquisition of tongue-root harmony

The example of the previous section involved the learning of the implementation of a single feature value (e.g., sibilant noise) by a possibly complex combination of gestures (e.g., a grooved tongue tip held at a critical position near the dental alveoli, combined with sufficient lung pressure, a closed velum, and an open glottis). In this section, I will show how the local-ranking principle helps in learning the ranking of continuous constraint families.

The example of this chapter will be the acquisition of a tongue-root-harmony system. The largest vowel space for tongue-root-harmony languages that we will consider

(Archangeli & Pulleyblank 1994; Pulleyblank 1993, 1996; Pulleyblank, Ping, Leitch & Ọla 1995), is a product of the dimensions of vowel height (low, mid, high), place (front, back), and tongue-root position (advanced = ATR, retracted = RTR):

	front		central	back	
	ATR	RTR		RTR	ATR
high	i	ɪ		ʊ	u
mid	e	ɛ		ɔ	o
low			ə ATR a RTR		

(14.38)

In order to find out what articulatory constraints are involved in the production of these sounds, we will have a look at the gestural and acoustic correlates of the height and tongue-root features.

Phonetically, a specified vowel height (F_1) can be implemented with the help of an oral tongue-body constriction and a mid-pharyngeal width adjustment. Abstracting away from the effects of lip spreading, tongue-body position, and damping, a specified F_1 must be implemented by adjusting the *quotient* of the cross-sectional area of the oral tongue-body constriction and the area at the mid pharynx: increasing these two areas by the same factor will roughly leave the F_1 unchanged, because the relative deviations of the resonance frequencies of a tract from those of a straight tube with the same length depend, in first approximation, only on the relative areas of the various regions, not on their absolute areas (Fant 1960: 63–67; Flanagan 1972: 69–72). The following table, therefore, shows an idealized account of the gestures that implement four F_1 values:

			retracted implementation			advanced implementation		
1-dim height	F_1	A_{phar}/A_{oral}	A_{phar}	A_{oral}	sound	A_{phar}	A_{oral}	sound
high	250 Hz	20				10 cm^2	0.5 cm^2	[i]
higher-mid	400 Hz	4	2 cm^2	0.5 cm^2	[ɪ]	8 cm^2	2 cm^2	[e]
lower-mid	600 Hz	0.6	1.2 cm^2	2 cm^2	[ɛ]	4.8 cm^2	8 cm^2	[ə]
low	850 Hz	0.1	0.8 cm^2	8 cm^2	[a]			

(14.39)

From this table, we see that the ternary “height” feature in (14.38) corresponds to the degree of oral constriction, and that the “TR” feature corresponds to the width of the pharynx. For instance, [e] has the same F_1-based height as [ɪ] (400 Hz), but it has the same constriction-based height as [ɛ] (2 cm^2).

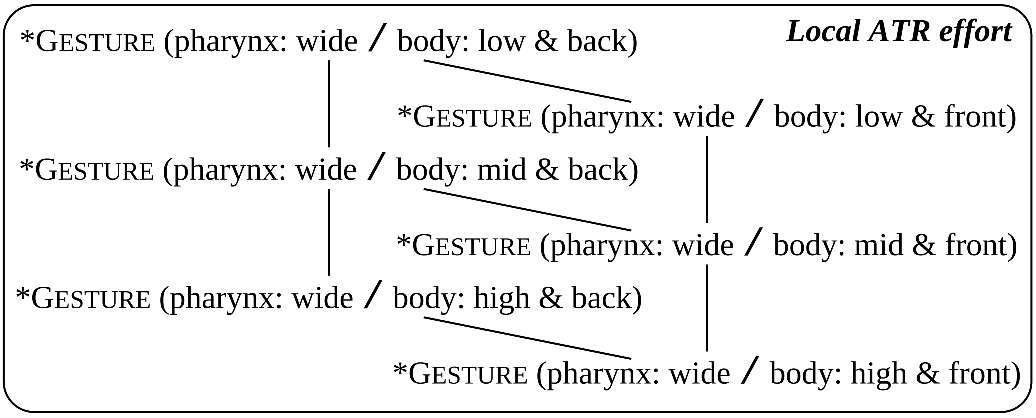
Many tongue-root harmony languages lack the advanced low vowel /ə/, or the retracted high vowels /ɪ/ and /ʊ/, or both of these sets. Archangeli & Pulleyblank (1994) ascribe these asymmetries to the following phonetically motivated ***grounding conditions*** on possible ***paths*** (simultaneous pairs of feature values):

1. LO/RTR: “if a vowel is low, then it has a retracted tongue root”
2. HI/ATR: “if a vowel is high, then it has an advanced tongue root” (14.40)

In later work, Pulleyblank (1993, 1996) translates these grounding conditions directly into OT-able constraints with the same names. From the viewpoint of Functional Phonology, however, these constraints must be regarded as ***surface constraints***: they adequately describe a tendency that occurs in the data, but their functional explanation may involve articulatory as well as perceptual arguments. The difference between such surface constraints and constraints directly derivable from functional principles may be subtle, but will be seen to have empirical consequences under a strict-ranking regime (§14.3.3).

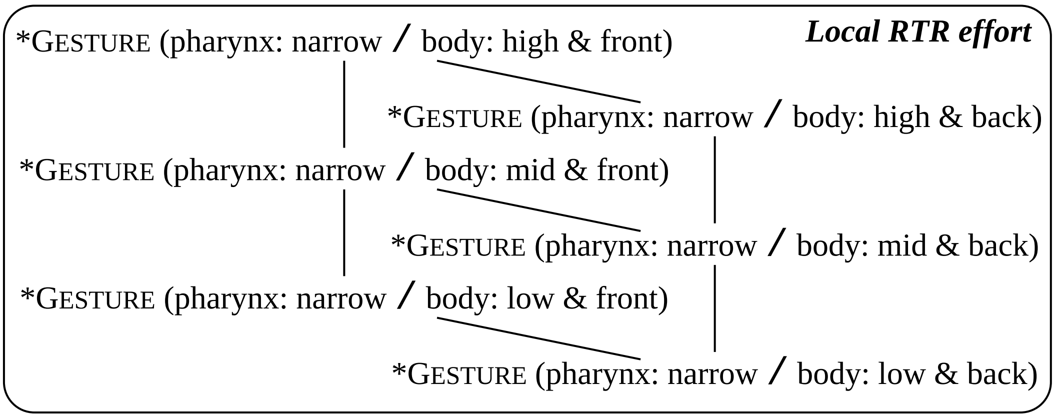
14.3.1 Universal ranking of articulatory constraints

From the functional point of view, the grounding constraints (14.40) could be articulatory constraints against the performance of tongue-root gestures, say *GESTURE (pharynx: *width* / body: *position*). These constraints have some universal rankings: it is easier to achieve a specified large pharynx width if the tongue body is pulled up or to the front, than if the tongue body is pulled down or backwards. These fixed rankings are shown with solid lines in the following figure:


*GESTURE (pharynx: wide / body: low & back)
Local ATR effort
*GESTURE (pharynx: wide / body: low & front)
*GESTURE (pharynx: wide / body: mid & back)
*GESTURE (pharynx: wide / body: mid & front)
*GESTURE (pharynx: wide / body: high & back)
*GESTURE (pharynx: wide / body: high & front)

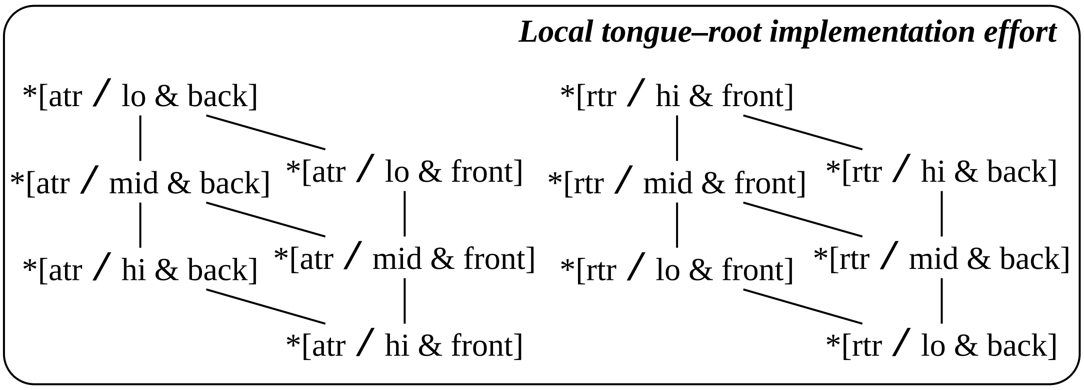
(14.41)

The universality of the rankings in this figure is subject to the local-ranking principle, which maintains that only pairs of minimally different (i.e., adjacent) constraints can ever be ranked in a universal manner, and that all other pairs (like the high-back versus mid-front pair) can be ranked freely in a language-specific way. Analogously to (14.41), the constraints against a narrowing of the pharynx are ranked as


*GESTURE (pharynx: narrow / body: high & front)
Local RTR effort
*GESTURE (pharynx: narrow / body: high & back)
*GESTURE (pharynx: narrow / body: mid & front)
*GESTURE (pharynx: narrow / body: mid & back)
*GESTURE (pharynx: narrow / body: low & front)
*GESTURE (pharynx: narrow / body: low & back)

(14.42)

If we assume, as a rather crude idealization, that the perceptual effects of tongue-root movement are a function of the realized pharynx width only (disregarding the interaction with the perceptual results of higher articulations), and that the perceptual feature values [front], [back], [high], [mid], and [low] correspond with horizontal and vertical tongue-body positions, we can write the articulatory constraint families (14.41) and (14.42) as implementation constraints for certain perceptual feature values. Respecting the local-ranking principle, we get



(14.43)

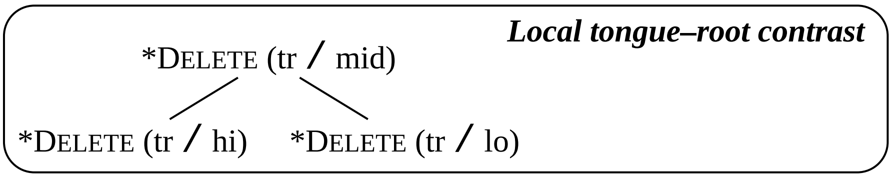
Note that all features in this picture are *perceptual* features, as opposed to those in (14.41) and (14.42), which are articulatory gestures. The equivalence between the two representations is a coincidence (and an idealization).

The local-ranking principle tells us that the [atr] hierarchy in (14.43) is not connected to the [rtr] hierarchy, because [atr] and [rtr] must be implemented by different muscle groups, if we assume that [atr] represents a tongue position forward from the neutral position, and that [rtr] represents a backward tongue movement.

14.3.2 Universal ranking of faithfulness constraints

The partial universal grammar (14.43) contains an idealization of ***perceptual invariance***: it shows how difficult it is to implement articulatorily the given perceptual feature values [atr] and [rtr]. Real languages will also show effects of ***articulatory invariance***: for the implementation of a given perceptual feature, the amount of effort considered worth spending will tend not to diverge much for the various possible environments. Thus, the feature value [atr], if implemented with the same genioglossus activity, will surface perceptually much more clearly for high front vowels than for low back vowels. For [rtr], the situation is the reverse: tongue-root retraction will be most clearly audible for high vowels.

If the relationship between the distance from the neutral position and the required articulatory effort is superlinear (e.g., $0.5^2 + 1.5^2$ for high and low vowels versus $1^2 + 1^2$ for mid vowels), this means that with invariance of articulatory effort, the perceptual contrast between the two tongue-root values will be largest for mid vowels, and smaller for high and low vowels. This can be translated into the following somewhat tentative fixed ranking of faithfulness constraints:

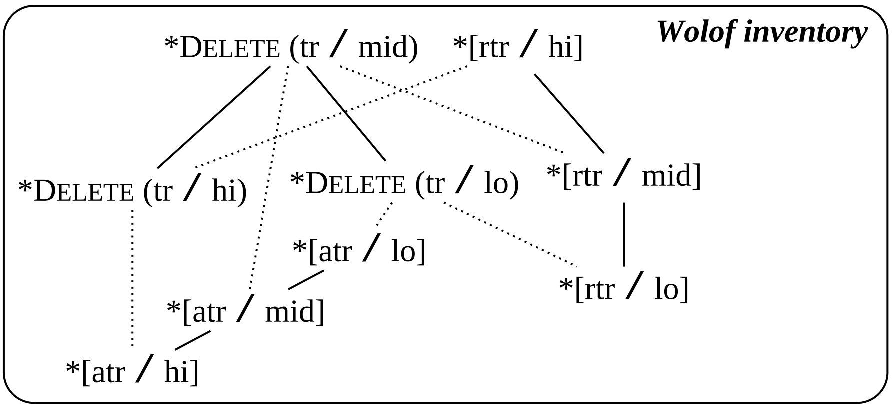


(14.44)

Real languages will trade some, but not all, perceptual invariance for articulatory invariance, so that they will combine (14.44) with a shrunk version of (14.43).

14.3.3 Typology of tongue-root systems

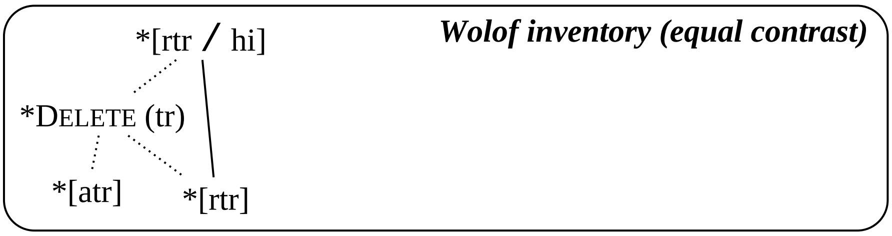
A combinatorial typology of possible tongue-root systems results from combining (14.43) and (14.44), subject to the local-ranking principle, which fixes the rankings that are represented with solid lines in these figures. Ignoring the back vowels, I will show two of the possible grammars. The system of Wolof (Pulleyblank 1996; Archangeli & Pulleyblank 1994: 225-239), which disallows the high RTR vowels /ɪ/ and /ʊ/ while the high ATR vowels are transparent to tongue-root harmony, and which does allow the low ATR vowel /ə/, will look like (not yet distinguishing ATR and RTR faithfulness):



(14.45)

The fixed rankings are drawn with solid lines, and the crucial Wolof-specific rankings with dotted lines. Most tongue-root-sensitive systems are defined by the ranking of *DELETE (tr / mid) above *[atr / mid] and *[rtr / mid], i.e. by the occurrence of an [e]-[ɛ] contrast.

A topology equivalent to (14.45) allows the generalization of some constraints, with a homogeneous *DELETE (tr):

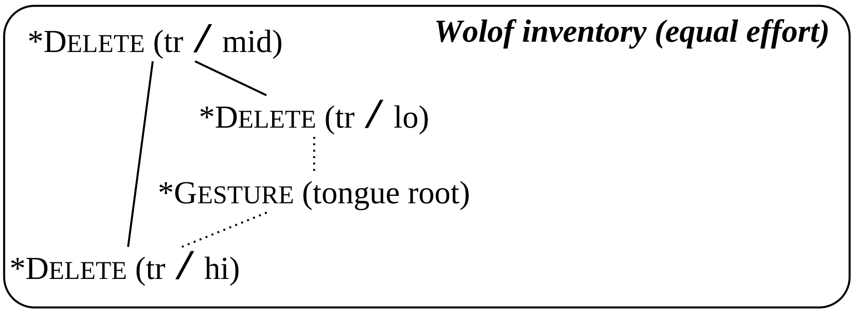


(14.46)

This grammar minimizes the number of constraints. Figure (14.46) also shows the technical possibility of an “elsewhere” formulation of *[rtr], licensed by its ranking below a more specific constraint.

Now we see the difference between positing phonetically motivated grounding constraints and the functional approach: though the constraint *[rtr / hi] (or HI/ATR) comes out on top, the constraint *[rtr / mid], which is universal in the sense that every language with a retracted mid vowel has to deal with it, “causes” other languages to have no tongue-root contrasts for mid vowels. The only restriction that the fixed ranking of these two constraints places upon possible grammars, is the fact that if a language licenses RTR in high vowels, it also licenses RTR in mid vowels.

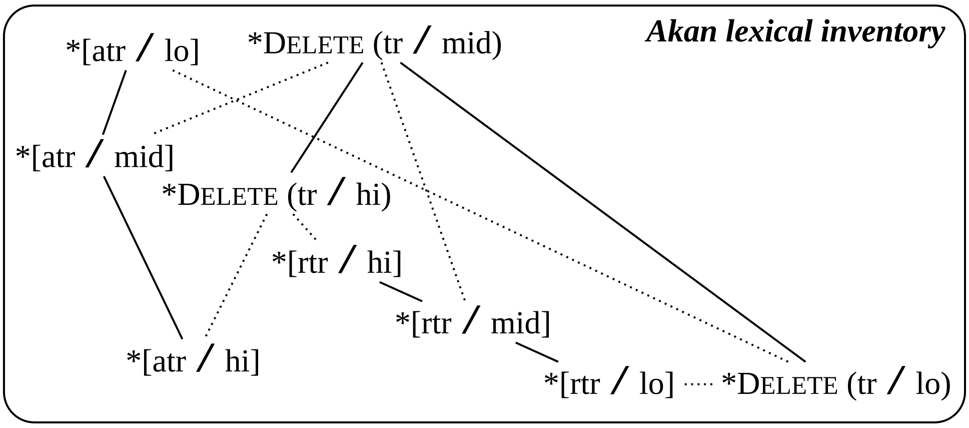
An empirically slightly different formulation of (14.45), with a reversal of the PARSE (tr / hi) and *[rtr / hi] constraints, would generalize all articulatory constraints:



(14.47)

This expresses the idea that if there is no tongue-root contrast for high vowels, the speaker will not bother to make either the advancing or the retracting gesture. In other words: if *DELETE (tr / hi) is ranked below *[rtr / hi], there is no tongue-root contrast for high vowels, and the contrast-dependency of the ranking of faithfulness will cause *DELETE (tr / hi) to fall even further, right to the bottom of the constraint-ranking continuum; specifically, below *GESTURE (tongue root: advance / high).

Other tongue-root systems vary as far as the freely rankable pairs are concerned, but keep the locally rankable pairs fixed. For instance, a pre-lexical stratum of Akan (Archangeli & Pulleyblank 1994: 212-225), which has no lexical tongue-root contrast for low vowels, can be analysed as



(14.48)

though constraint generalizations will result in something analogous to (14.46). The ranking of *[rtr / lo] versus *DELETE (tr / lo) is depicted as ambiguous, in order to allow both possible interpretations of Akan-like systems: if *[rtr / lo] is the higher of the two, the *GESTURE constraints can be regarded as homogeneous, and /a/ must be considered as having no articulatory specification for tongue-root movement; if *DELETE (tr / lo) is higher, the *DELETE constraints can be regarded as homogeneous, and /a/ contains [rtr]. An empirical difference between these two systems could be found in the spreading of the retracted-tongue-root gesture from /a/, which should be possible only in the latter case, under the assumption that only articulatory gestures can spread but perceptual feature values cannot.

From a combinatorial typology of tongue-root systems, we can derive two implicational universals, which are assumed by Pulleyblank & Turkel (1995, 1996):

1. If a tongue-root language with three vowel heights has [ɪ], it also has [ɛ] and [a].
2. If a tongue-root language with three vowel heights has [ə], it also has [e] and [i].

(14.49)

According to the local-ranking principle, these universals are independent from each other, i.e., the probability of a language having [ɪ] does not depend on whether it has [ə].

14.3.4 The learning process for continuous families

As I argued above, all learners start from the same empty grammar. As speakers, they start out with undominated *GESTURE constraints, because no articulatory speech gestures and coordinations have yet been learned. As listeners, they start out with undominated *CATEG constraints, because no perceptual categorizations have yet been established. As far as categorization is concerned, we can assume that every learner of a tongue-root language that includes /i/, /e/, /ɛ/, and /a/, learns to categorize the perceptual correlate of the constriction-based vowel-height dimension (perhaps with the help of the presence of harmony) into three classes (low, mid, and high), because these occur in the acoustic input; and that she learns to categorize the perceptual tongue-root dimension into two classes (ATR and RTR), because both of these values occur in the listener's input.

Initially, therefore, the learner is not capable of making either the tongue-root-advancing or the tongue-root-retracting gesture: they will still have to be learned.

For the learning process, therefore, we must consider a binary perceptual feature, with values [atr] and [rtr], and a ternary production feature: the advancing gesture, the retraction gesture, or no tongue-root gesture at all. If a vowel is pronounced without a tongue-root gesture (which I will denote by a dieresis diacritic), it must still be perceived as either [atr] or [rtr], with probabilities like those in the following table:

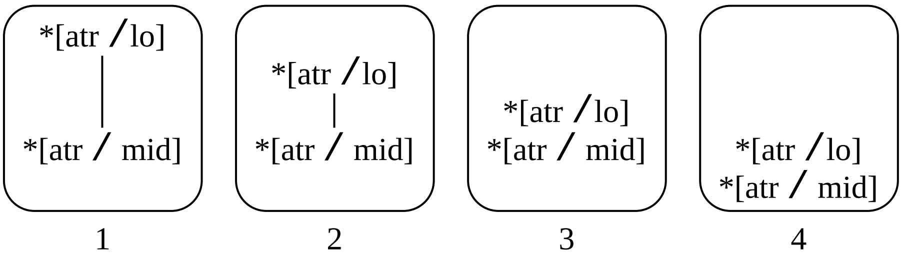
↓ produced perceived →	atr	rtr
[ä]	0.1	0.9
[ë]	0.4	0.6
[ï]	0.8	0.2

(14.50)

The fact these none of these probabilities is zero, will guarantee that the learner will at some time perceive some utterances without tongue-root gestures as unfaithful (cf. §14.2.9).

Initially, none of the constraints in (14.41) and (14.42) is dominated by a faithfulness constraint. This gives one of the typologically possible systems: one without any tongue-root contrasts.

The local-ranking principle asserts that *[atr / lo] is always ranked above *[atr / mid], so that if motor learning causes the demotion of *[atr / lo], the universally lower constraint *[atr / mid] (and the infinite number of constraints in between along the height dimension) will be pushed along down the hierarchy:



(14.51)

14.3.5 The learning of simplified Wolof

As an example, we will look at a language that is unlearnable with the greedy and conservative error-driven learning algorithm by Gibson and Wexler (1994), which I will discuss in §14.4.2. This language is a simplified version of Wolof, as used by Pulleyblank & Turkel (1995, 1996) in a three-parameter account of the typology of tongue-root harmony (discussed in §14.4.1). In §14.4.3, I will show that parameter-setting learners of this language can end up in an absorbant cycle of two hypotheses, both of which cannot account for the perfect Wolof sequence /ɛti/. By contrast, constraint-sorting algorithms, including the gradual algorithm advocated in this chapter, always end up in a hypothesis that accounts for /ɛti/.

Thus, we will follow the stages in the acquisition of a minimal Wolof-like language, all of whose utterances are V_1tV_2 sequences, where V_1 and V_2 are chosen from the set of six non-round vowels { a, ə, ɛ, e, ɪ, i }. The relevant surface forms of this language can be found below in column B of table (14.63). Wolof honours the Hi/Atr grounding condition, so that /ɪ/ does not occur on the surface, but it does not honour Lo/Rtr, so that /ə/ *is* a licit segment. Thus, possible VtV surface forms include the 13 harmonic utterances /iti/, /eti/, /ite/, /itə/, /əti/, /ete/, /ɛtɛ/, /etə/, /əte/, /ɛta/, /atɛ/, /ətə/, and /ata/, and do not include any of the 11 thinkable utterances with at least one /ɪ/.

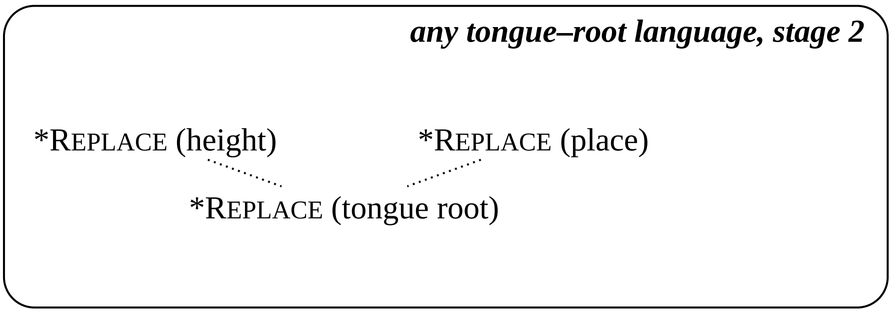
For the remaining 12 VtV words, we have to examine the fact that Wolof shows tongue-root ***harmony***, i.e. there is a structural constraint, say *[TR contour], which disallows an ATR and an RTR vowel to occur together in a word, and this constraint must outrank at least one faithfulness constraint. I will follow Pulleyblank and Turkel (1995, 1996) in their choice of faithfulness constraints (though with a perceptual interpretation), suppressing the dependence of *Delete on vowel height but letting it depend on the value of the tongue-root feature; thus, the constraints are *Delete (atr) (which for binary categorization is the same as *Replace (tongue root: advanced, retracted)), and *Delete (rtr). If the dominated constraint is *Delete (rtr), then the harmony constraint will force an underlying |etɛ| to become [ete]. However, Pulleyblank and Turkel state that Wolof is RTR-dominant, which means that *Delete (rtr) >> *Delete (atr), so that *Delete (atr) must be the dominated constraint and |etɛ| will surface as [ɛtɛ]. In either case, the eight disharmonic surface forms /ate/, /eta/, /ətɛ/, /ɛtə/, /ɛte/, /etɛ/, /atə/, and /əta/ will never occur.

However, Archangeli and Pulleyblank (1994) report no underlying disharmonic forms for actual Wolof, so RTR dominance must be assessed in a different way. Wolof allows the surface forms /ɛti/ and /ati/, which can be explained by the ranking *[rtr / hi] >> *DELETE (rtr) >> *[TR contour] >> *DELETE (atr): for underlying |ɛti|, the candidate [ɛtɪ] would violate *[rtr / hi], and /eti/ would violate *DELETE (rtr). Thus, because RTR faithfulness dominates harmony, our simplified Wolof allows /ɛti/, /itɛ/, /ati/, and /ita/ (for real Wolof, see §14.3.9). This concludes our description of the adult forms.

Stage 1. The initial state is the same for all languages, see figure (14.7). I assume that in the initial state of the acquisition of tongue-root contrasts, three abstract (constriction-based) vowel heights have already been learned. If the perceptual distances *within* the tongue-root pairs { e, ɛ } and { ə, a } are smaller than the distances *between* the pairs, this may simply result from similarity-based categorization. In stage 1 of the acquisition of a tongue-root inventory, the three possible non-back vowels are pronounced as [ä ë ï] (using the umlaut diacritic to denote tongue-root neutrality) and even the adult vowels are perceived as the equally undifferentiated set /ä ë ï/.

Step 1. In a language environment that makes extensive use of the tongue-root contrast, the learner will acquire a perceptual dimension not exploited in other languages: the perceptual [tr] feature. The chances that a separation of the two spectrum-based perceptual tiers for [height] and [tr] will occur, increases as the number of vowel qualities grows. The minimal categorization along the [tr] dimension is binary, and that is probably also the maximal one, lest confusion probabilities should get out of hand.

Stage 2. We may conjecture that the binary tongue-root contrast is perceptually less salient than a ternary height contrast or a binary place (front unrounded / back rounded) contrast (on the average; the local-ranking principle allows variation here). Therefore, the categorization step may well have led to a ranking like



(14.52)

This ranking will be similar for many tongue-root languages. The learner now perceives the adult set /a ə ɛ e i/, but still produces only the no-tongue-root vowels [ä ë ï].

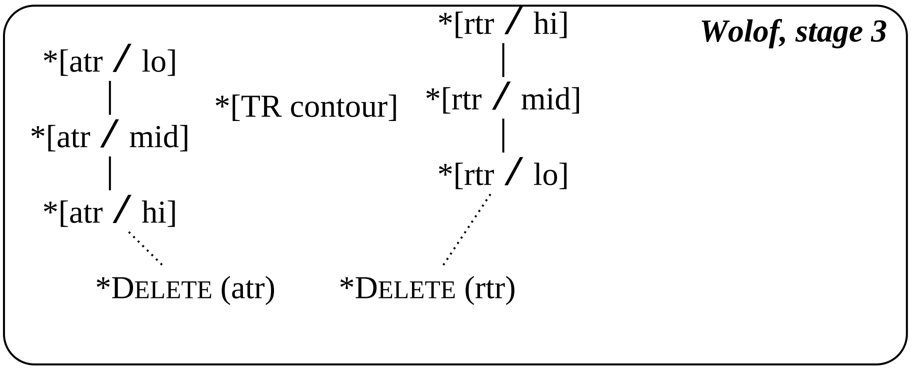
Step 2. Sensorimotor learning will push the necessary gestural constraints into the grammar from above. They will include the familiar *[atr] and *[rtr] constraints, plus, crucially for a tongue-root-harmony language, constraints against tongue-root gestures within an utterance or within a word: things like *GESTURE (tongue root: from advanced to retracted). I will collapse the various harmony constraints simply under the name of *[TR contour].

Also, the first Wolof-specific phenomenon will emerge: the language does not allow [ɪ], so that the learner will not practise the tongue-root retraction gesture for high vowels, and the constraint *[rtr / hi] will be undominated and can be left out of the grammar. One might think that when the "rich base" comes with an input that contains /ɪ/, we need *[rtr / hi] to keep it from surfacing.

Universality versus innateness 2: richness of the base. "*Richness of the base* is a concept useful for typological study, and can predict the behaviour of humans when they borrow words from another language, but cannot be proven to be psychologically real." (14.53)

This is somewhat harsh on the constraint *[rtr / hi], as that is just a member of the *[rtr] family. To show the difference with linguistically relevant but invisible constraints and those irrelevant constraints that are left out of the pictures by reason of space (like *[front] or *DELETE (front)), I shall include constraints like *[rtr / hi] in my figures, but they will be clinging to the ceiling or to the floor, as in (14.54).

Stage 3. Disregarding the height and place features, the constraints will be

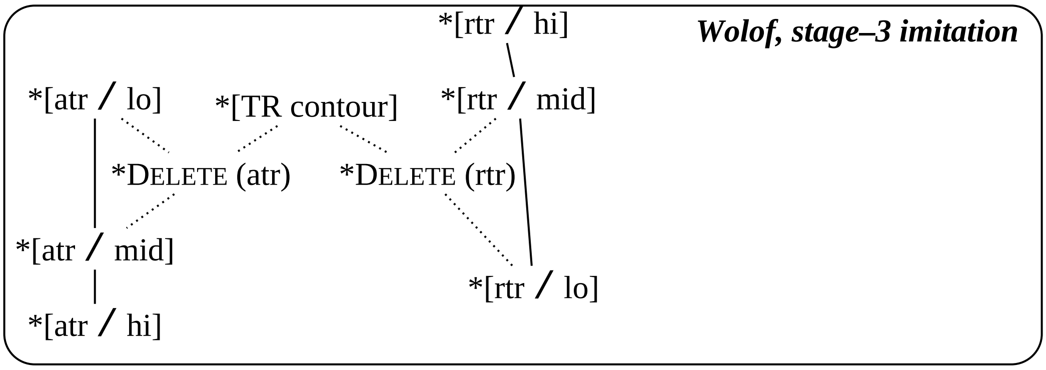


(14.54)

This figure shows the local rankings as solid lines, and the crucial rankings as dotted lines; the constraint *[rtr / hi] hangs from the ceiling, since it not a real part of the learner's inventory of constraints.

The output of grammar (14.54) will always be an utterance without any tongue-root gestures, surfacing only with the vowels [ä], [ë], and [ï] (which will be perceived according to table 14.50). The harmony constraint *[TR contour] is, therefore, automatically satisfied, so it is not yet crucially ranked with respect to any of the other constraints. As far as production is concerned, the articulatory tongue-root feature must at least be ternary; adult speakers, however, will never use the null gesture, because that would cause a large probability of confusion, according to table (14.50).

In ***imitation***, the *DELETE constraints are raised, and a lot of performances are possible, depending on the accidental relative heights of the gestural constraints. We can get an example by putting the *DELETE constraints on a horizontal line in (14.54) just above *[atr / mid]. The result is



(14.55)

In this example, a model [e] will be pronounced faithfully, because *DELETE (atr) dominates *[atr / mid]. At the same time, a model [ɛ] will be imitated as [ë], because *[rtr / mid] still outranks *DELETE (rtr), resulting in a faithful /ɛ/ perception only 60% of the time, according to table (14.50), and in an unfaithful /e/ 40% of the time; of course, this is still better than producing an [e] outright:

[ɛ] /ɛ/=|ɛ|	*[rtr / mid]	*DELETE (tr)	*[atr / mid]
[ɛ] /ɛ/	*!		
☞ [ë] 60% /ɛ/ 40% /e/		40% *	
[e] /e/		*	*!

(14.56)

Because of the high anti-contour constraint, the Wolof utterance [ita] would be rendered with vowel harmony:

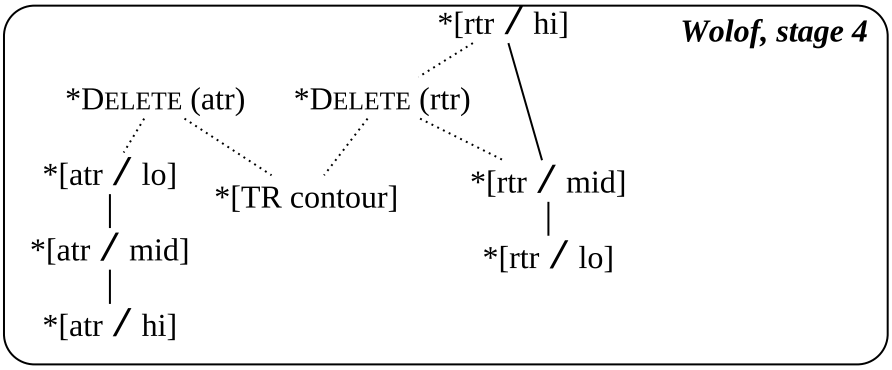
[ita] /ita/=|ita|	*[rtr / hi]	*[TR contour]	*DELETE (tr)
[ita] /ita/		*!	
☞ [itə] /itə/			*
[ɪta] /ɪta/	*!		*

(14.57)

Step 3. On hearing /ə/, the Wolof learner will have to demote *[atr / lo] and raise *DELETE (atr); because of local ranking, the constraint *[atr / mid] will be pushed along, causing the advanced pronunciation of the sound [e] suddenly to become licensed in the learner's grammar, without her needing any [e] data. For the demotion of the two *[rtr] constraints, the learner will need some [ɛ] data, and if these are presented to the learner, she will acquire the correct tongue-root gesture for [a] automatically. Finally, because Wolof allows the harmony violations /ɛti/ and /ati/, the anti-contour constraint has to be demoted below the *DELETE constraints. During the course of all these rerankings, the learner may go through a lot of different grammars, of which a grammar topologically equivalent to (14.55) is just one example.

Of all the possible articulatory tongue-root contrasts with neutral gestures, only [e ë] and [ɛ ë] come close to implementing a perceptual contrast (according to (14.50)). Because of local ranking, *[atr / mid] will always pass the *DELETE constraints before *[atr / lo] does, so that if the learner produces a faithful tongue-root contrast for low vowels, she will always be able to make some sort of contrast for mid vowels as well. Thus, the implicational universals (14.49) are satisfied at any moment during acquisition, and adult markedness relations reflect acquisition order.

Stage 4. The acquisition of the gestures comes to an end when all relevant constraints have fallen below the *DELETE constraints, which have risen:



(14.58)

The segment inventory has been learned correctly in this grammar of depth 2 (i.e. with at most two cascaded non-universal crucial rankings, denoted by the dotted lines). The two possible tongue-root contours, however, have been generalized, so that the learner would now allow in her grammar (= be able to pronounce) [ɛte], [ate], [atə], and [ɛtə]. To get rid of these, she needs evidence for the violability of a *DELETE constraint.

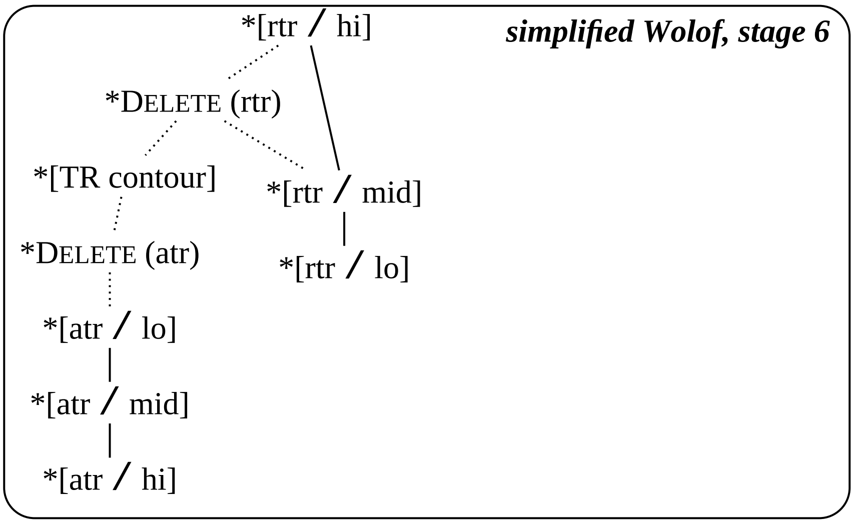
Step 5. Once the learner reconstructs by morphological analysis that a certain form is underlyingly |at+e| (no longer real Wolof), she will pronounce this faithfully as [ate], according to (14.58). However, when hearing that an adult actually pronounces this as /atɛ/, the learner is confronted with an output mismatch. She now has the negative evidence that she needed to rule out [ate]:

[atɛ] /atɛ/ |at+e|	*DELETE (rtr)	*DELETE (atr)	*[TR contour]
☞ [ate] /ate/			←*
[əte] /əte/	*!		
√ [atɛ] /atɛ/		*!→	

(14.59)

The learner discovers that it is not at all very important to pronounce an underlying |e| faithfully, or that the harmony is more important than she had in mind. Therefore, she will demote *DELETE (atr), and promote *[TR contour]. The constraint *DELETE (rtr) will stay where it is: apparently, this is an RTR-dominant language (the [rtr] specification of |a| dominates the [atr] specification of |e|). Note that the learner does not need to know that this has anything to do with RTR dominance nor with the interaction with a harmony constraint: it occurs automatically, and no innate parameters are needed. After some of these data, *DELETE (atr) will fall down past *[TR contour]; if *[TR contour] happens to rise above *DELETE (rtr) as a result of this procedure, disharmonic data like /ati/ will push *[TR contour] down and raise *DELETE (rtr). Eventually, a stable ranking *DELETE (rtr) >> *[TR contour] >> *DELETE (atr) will emerge.

While *DELETE (atr) is falling, it may come near *[atr / low]. In this case, the learner will probably experience some mismatches when comparing her output with adult /ə/ data, and she will demote the *[atr] family, with the idea of keeping [ə] pronounceable. When *DELETE (atr) finally drops below the anti-contour constraint, the situation is likely to be



(14.60)

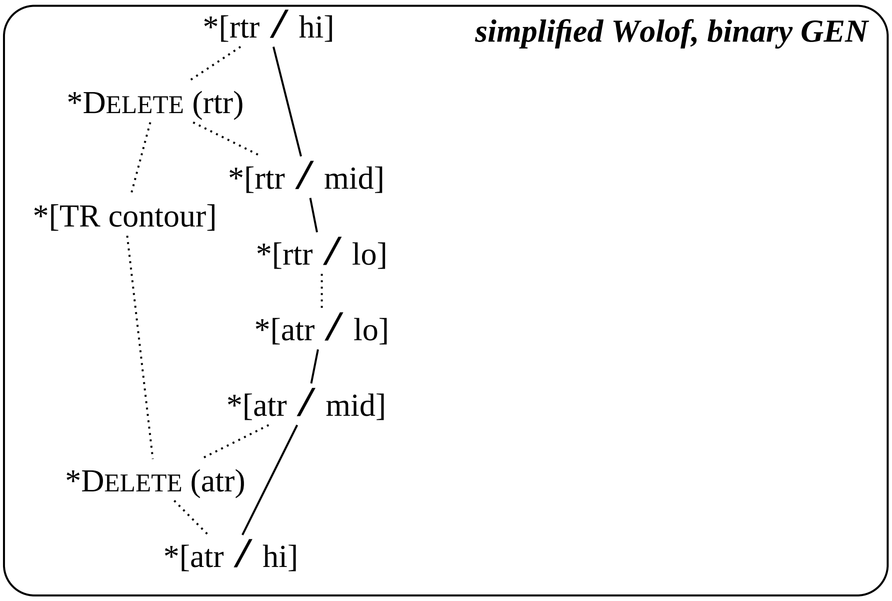
This is the correct grammar of our simplified “Wolof”. It is four crucial rankings deep: along the dotted path we see the grounding condition for high vowels (the top constraint), tongue-root harmony (the contour constraint dominating at least one faithfulness constraint), RTR dominance (the contour constraint sandwiched between two *DELETE constraints), and the availability of schwa (the domination of *[atr / lo]).

It is possible that the demotion of *DELETE (atr) below the harmony constraint actually precedes the demotion of *[atr / lo]. In that case, [ə] would temporarily become unpronounced (not unpronounceable) in some cases (see §14.3.7), but the output mismatches that result from it will eventually draw the *[atr] family to the bottom of the relevant hierarchy. In all thinkable cases, grammar (14.60) will result.

14.3.6 An alternative Wolof: articulatory versus perceptual candidates

Most OT work is performed within the generative tradition of hybrid phonological features. So let us try to restrict GEN to a binary tongue-root constraint, analogous to our perceptual feature, instead of to a ternary (in reality: continuous) gestural constraint. Wolof can then be described with the alternative grammar (14.61).

But (14.61) is the wrong grammar: [ə] surfaces not because *DELETE (atr) outranks *[atr / lo], but because *[rtr / lo] outranks *[atr / lo]. From the functional standpoint, this is outrageous: while not *perceiving* a feature as [atr] may mean that you perceive it as [rtr]

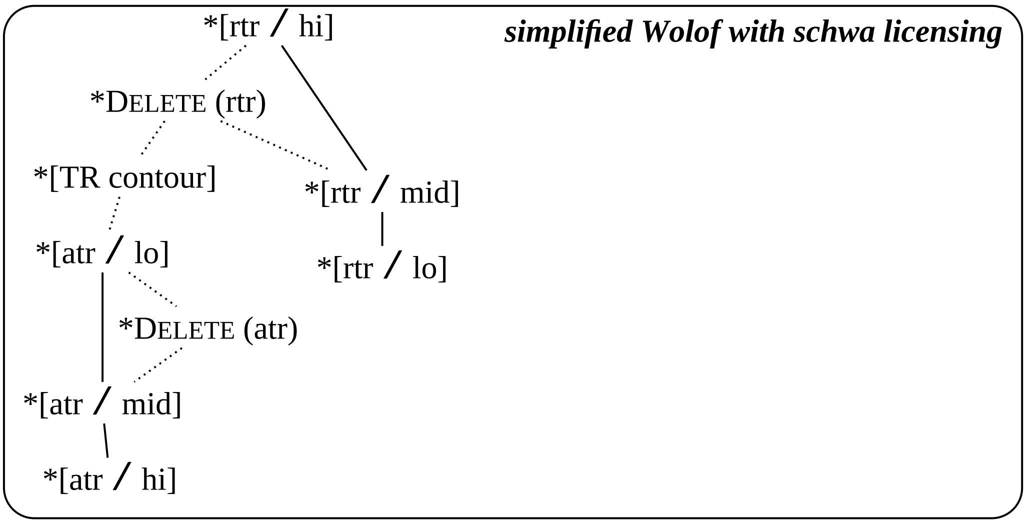


(14.61)

(because of the binary categorization), not *producing* a gesture should never mean that you have to make the opposite gesture: a prohibition on a gesture can never force another gesture; only *DELETE constraints can do that (from the generative standpoint with its hybrid features, there would be no problem, because there would be no *[rtr / lo] constraint). Instead of a forced choice between two gestures, there should always be the possibility of no gestures at all; if articulatory constraints are unviolated, the result should be *no gesture*, not the default gesture.

14.3.7 Wolof with schwa licensing

One of the possible grammars that are one step removed of converging onto simplified Wolof, has a depth of five:



(14.62)

This is a peculiar language indeed: it disallows an isolated *[ə], and it disallows *[ətə]. An underlying |ətə| will surface as [ata] because the grounding condition *[atr / lo] dominates *DELETE (atr). However, the other schwa words [əte] and [əti] are allowed, because *[TR contour] dominates *[atr / lo]. In other words, the ATR gesture of [e] and [i] licenses ATR in a low vowel. Note that this [ə] is not just a positional variant of |a|: underlying |ati| still surfaces faithfully because *DELETE (rtr) dominates the harmony constraint (and *[atɪ] is out because of *[rtr / hi]). Thus, |ə| is fully contrastive, though it requires an adjacent non-low ATR vowel to survive.

This example involves five crucial rankings in cascade. The free ranking allowed by the local-ranking principle (as well as the standard account with grounding constraints) would predict that this is a possible language.

14.3.8 Learning unnatural local rankings

The learning of certain combinations of gestures often involves the demotion of the relevant gestural constraints below “universally” easier gestures.

For instance, speakers of Dutch are used to implementing the /b/-/p/ contrast in such a way that both plosives require active gestures to make them voiced or voiceless; these same speakers have trouble pronouncing the English or German lenis voiceless plosive [b̥], though that sound would be easier than either Dutch plosive because it requires no active voicing or devoicing gesture. Likewise, speakers of tongue-root languages may learn to have trouble **not** performing any tongue-root gestures in vowels.

As a more dramatic example, consider the crosslinguistically abundant /i/-/u/ contrast. Most speakers of a language with exactly these two high vowels will have trouble pronouncing the unrounded high back vowel [ɯ], though that sound should be universally easier than [u] because it does not involve a lip gesture. The reason that /i/ is a spread front vowel and /u/ is a rounded back vowel, is that the perceptual contrast of “front” (high F_2) versus “back” (low F_2) is best implemented by varying the lip shape as well as the tongue-body position. It is highly unlikely that articulatory ease is involved in rounding back vowels: first, rounding the lips costs energy; second, there is no innate anatomical or functional relationship between rounding and backing; third, the prevalence of unrounded velar obstruents proves that even in speech, the relation is not automatic.

Thus, requirements of faithful voicing or F_2 contrasts lead to learning complex coordinative gestures, and the single gestures are unlearned. This is a normal procedure in human motor behaviour; one of its advantages is the reduction of cognitive load, i.e. the number of high-level neural commands. Still, we may suspect that in the early stages of acquisition the single gestures are still easier for the child than the complex gestures. Thus, the first plosive that the Dutch (or any other) child will learn, before trying to implement or even recognize the voicing contrast, is one without any active voicing or devoicing gestures: typically, a lax voiceless stop. Likewise, we would expect Turkish children, if they recognize a four-way contrast in the high vowels, to have less trouble with the pronunciation of [ɯ] than with [u].

There are also things that seem to go against the local-ranking principle. Adult speakers of Proto-Indo-European may have had trouble pronouncing [b] even though they had [d] and [g] (and [p]), and [b] is allegedly easier to voice. However, all Indo-European languages hurried to fill up the original gap at /b/ (§17.0), suggesting that learners may have considered it to be an accidental gap in their lexicon, not in their grammar. This suggests that the local-ranking principle may be valid into adulthood.

14.3.9 Real Wolof

The simplified Wolof described above was chosen for its known problems with parameter-setting learning algorithms. It differs from real Wolof (Archangeli & Pulleyblank 1994: 225–239) in a number of respects. I will now show that the differences do not require us to pull into question our functionalist approach.

First, Wolof has long and short vowels, and /ə/ is only allowed as a short vowel. Thus, the constraint *[atr / lo] must split, so that *[atr / long low] is unviolated. Whether the acquisition process involves constraint splitting (of *[atr / lo]) or constraint generalization (of e.g. *[atr / long mid] and *[atr / short mid]), or both, is a question that has no bearing on the nativist/environmentalist issue, since any OT account of Wolof will have to introduce a diacritic here; for instance, Pulleyblank (1996) summarizes the /ə/ facts with the ranking LO/ATR$_{\mu\mu}$ >> LO/ATR, where μ denotes a timing unit (the mora).

Secondly, Wolof tongue-root harmony is directional: it works only from left to right. Consider the form /doːraːtɛ/ ‘to hit usually’. The vowel in the medial syllable can never be /əː/, because of the high ranking of *[atr / long low]. Since the harmony constraint outranks *DELETE (atr), the form would have to be */dɔːraːtɛ/, i.e. every word containing an underlying RTR specification or a long vowel, and no high vowels, would have to be entirely RTR. As it is, the initial ATR /doː/ is allowed, and only the final syllable /tɛ/ must share its retraction with the preceding /aː/. Pulleyblank (1996) accounts for this phenomenon with a constraint whose simplest form could be written as ALIGN (RTR, right; Word, right); e.g. the form */doːraːte/ would violate this constraint by one syllable, since the right edge of the RTR span /raː/ is one syllable away from the right edge of the word. Since alignment constraints are highly language-specific (they are often morphologically conditioned), their specific forms cannot be innate anyway, so they must be learned; perhaps they are created automatically for every pair of learned features and/or morphological constituents (as suggested by Mark Ellison).

Opacity effects, like the opacity of /aː/ for rightward spreading of ATR, are expected for articulatory harmony constraints: opacity reduces the number of contours. The third distinguishing property of real Wolof, however, is that it also shows a transparency effect: Wolof allows forms like /tɛkːilɛːn/ ‘untie!’, but not */tɛkːileːn/. Apparently, RTR spreads to the right through the high vowel, which is not allowed to become RTR itself because of the high ranking of *[rtr / hi]. Instead of reducing contours, this kind of harmony maximizes the number of vowels that carry RTR. Functionally speaking, the RTR specification tries to express itself maximally, in order that it be heard optimally. We could call this constraint MAXIMUM (RTR); thus, Wolof seems to have articulatory harmony (opacity) as well as perceptual harmony (transparency). Pulleyblank (1996), however, uses the same alignment constraint as above, but alignment is not only to the Word, but also to the nearest RTR value. Thus, /tɛkːilɛːn/ violates it only once, because /tɛ/ is only one syllable away from the RTR sequence /lɛːn/, while */tɛkːileːn/ violates it twice, because /tɛ/ is two syllables away from the right edge of the word. Note that the two approaches are empirically different: with MAXIMUM (RTR), you would not expect a non-underlying RTR value (i.e. one that is forced by a long low vowel) to spread through a high vowel (/doːraːtɛbɔːbule/); with Pulleyblank’s ALIGN (RTR, right; RTR, left / Word, right), you would: /doːraːtɛbɔːbulɛ/. Of course, Pulleyblank’s prediction for this hypothetical sequence will be correct.

The fourth difference is that the word-initial forms /itɛ/ and /ita/ (with short /a/) are not allowed in real Wolof. Apparently, an underlying RTR must always be realized on the first syllable; if this is impossible because of high-vowel grounding, RTR must be deleted. Pulleyblank (1996) accounts for this with a ranking like HI/ATR >> ALIGN (RTR_{root}, left; Root, left) >> *DELETE (RTR), i.e. with an alignment constraint that refers to the underlyingness of its material.

Since constraint splitting, constraint merger, and alignment constraints, are devices that must be learned regardless of whether structural and faithfulness constraints are learned or innate, the differences between real Wolof and our simplified Wolof do not constitute any threat to the hypothesis that all constraints can be learned. The least expensive starting point, then, is that there are no innate phonological constraints.

14.4 Principles-and-parameters learning algorithms

The Gradual Learning Algorithms described above are ***error-driven*** (react only to output mismatches), ***incremental*** (small changes at a time) and ***greedy*** (only changes that are directly aimed at improving the grammar). It leads to the learnability of any tongue-root-harmony system: gestural constraints are lowered only if positive evidence (the occurrence in the adult utterance) forces the learner to start to practice the gestures. In the

language →	A	B	C	D	E	F	G
grammar → vocabulary ↓	A1: ATR A2: RTR	HI/ATR RTR	HI/ATR ATR	LO/RTR RTR	LO/RTR ATR	HI/ATR, LO/RTR RTR	HI/ATR, LO/RTR ATR
iti	1/18	1/17	1/13	1/13	1/17	1/12	1/12
itɪ, ɪti	0	0	0	0	0	0	0
ɪtɪ	1/18	0	0	1/13	1/17	0	0
ite, eti	2/18	2/17	2/13	2/13	2/17	2/12	2/12
itɛ, ɛti	0	2/17	0	0	0	2/12	0
ɪte, etɪ	0	0	0	0	0	0	0
ɪtɛ, ɛtɪ	2/18	0	0	2/13	2/17	0	0
itə, əti	2/18	2/17	2/13	0	0	0	0
ita, ati	0	2/17	0	0	2/17	2/12	2/12
ɪtə, ətɪ	0	0	0	0	0	0	0
ɪta, atɪ	2/18	0	0	2/13	2/17	0	0
ete	1/18	1/17	1/13	1/13	1/17	1/12	1/12
etɛ, ɛte	0	0	0	0	0	0	0
ɛtɛ	1/18	1/17	1/13	1/13	1/17	1/12	1/12
etə, əte	2/18	2/17	2/13	0	0	0	0
eta, ate	0	0	0	0	2/17	0	2/12
ɛtə, ətɛ	0	0	0	0	0	0	0
ɛta, atɛ	2/18	2/17	2/13	2/13	2/17	2/12	2/12
ətə	1/18	1/17	1/13	0	0	0	0
əta, atə	0	0	0	0	0	0	0
ata	1/18	1/17	1/13	1/13	1/17	1/12	1/12

(14.63)

coming sections, we will have a look at the problems that arise with several existing learning algorithms, and at the solutions that are provided by the GLA.

14.4.1 Seven possible tongue-root-harmony systems

Table (14.63) shows the seven possible tongue-root harmony languages used by Pulleyblank & Turkel (1995, 1996) in their learning algorithm, and the possible VtV words in those languages with their probabilities of occurrence.

The eight grammars stem from three binary ***parameters***. They are:

(1) LO/RTR: determines whether the grammar honours the LO/RTR grounding condition, i.e. whether (in our terms) *[atr / lo] dominates *DELETE (atr).
(2) HI/ATR: determines whether the grammar honours the HI/ATR grounding condition, i.e. whether *[rtr / hi] dominates *DELETE (rtr).
(3) ATR versus RTR: dominance of either feature value (harmony is assumed). In the case of the ATR setting of this parameter, the rankings ALIGNATR >> MAXRTR as well as MAXATR >> ALIGNRTR are valid; in our terms, this would be PARSE (atr) >> *[TR contour] >> PARSE (rtr)[7]. This ***packaging*** of constraints might turn out problematic for acquisition, as it entails ART >> FAITH and FAITH >> ART at the same time. Also, a language like (14.62) cannot be represented. (14.64)

For instance, language B is our familiar Wolof-like language: it honours the HI/ATR grounding condition and is RTR-dominant.

Language A really has two possible grammars: ATR-dominant (A1) and RTR-dominant (A2), as harmony is never violated. Language C is the intersection of A and B, and language D is the intersection of A and E. Language G is a proper subset of E, and F is a proper subset of B.

The numbers in table (14.63) denote the probability of occurrence of the 21 possible VtV patterns, under the assumption that every possible surface form is equally likely to occur in an utterance. The surface forms [iti], [ete], [ɛtɛ], and [ata] are always possible because they satisfy all harmony (alignment) and grounding (filter) constraints. Forms with [ɪ] are ruled out in languages that satisfy HI/ATR, and forms with [ə] are ruled out in languages that satisfy LO/RTR. The surface forms [itɪ], [etɛ], and [atə] can never occur because they respect neither the ranking ALIGNATR >> MAXRTR nor ALIGNRTR >> MAXATR; an underlying /etɛ/, for instance, would surface as [ete] in an ATR-dominant language, and as [ɛtɛ] in an RTR-dominant language.

The disharmonic forms [itɛ] and [ita] are licensed in RTR-dominant languages with a highly ranked HI/ATR grounding constraint (B and F): for underlying /itɛ/, RTR dominance would suggest the output [ɪtɛ], but HI/ATR grounding forbids [ɪ]; the candidate [ite] would satisfy ALIGNATR but violate MAXRTR; so [itɛ] surfaces unchanged. In the ATR-dominant language C, by contrast, underlying /itɛ/ gives [ite] because the output-oriented constraint ALIGNATR dominates MAXRTR. Likewise, the disharmonic forms [ate] and [ati] are licensed in ATR-dominant languages with a strong

[7] The difference between word-level alignment and contour prohibition is subtle. If there are vowels that are transparent to harmony, we might talk of alignment; if there are opaque vowels, we might talk of contours. Wolof, for instance, has transparent high vowels and an opaque /a/.

LO/RTR grounding constraint (E and G); in the RTR-dominant language D, by contrast, underlying /ate/ would give [atɛ].

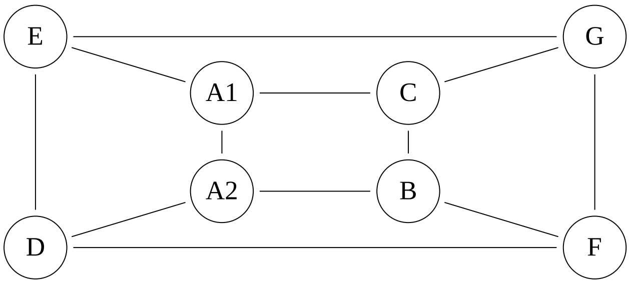
In absence of information about possible abstractness effects, we must assume that the probabilities of finding the various underlying forms are equal to the numbers shown in (14.63). For instance, the underlying form /itɪ/, though universally possible in OT, occurs with zero probability in each of the seven languages, since it would surface as either [iti] or [ɪtɪ], and the learner would analyse these surface forms as faithful reflexes of the underlying forms /iti/ and /ɪtɪ/.

14.4.2 The Triggering Learning Algorithm

Gibson & Wexler (1994) proposed their ***Triggering Learning Algorithm*** (TLA) in order to account for the acquisition of three syntactical parameters (subject location, object-verb order, verb-second). Pulleyblank & Turkel (1995) applied the TLA to the current three-parameter grammar space.

According to the TLA, the learner can, at any moment during acquisition, have as her current hypothesis any of the eight grammars A1 to G, and she may replace it with a different grammar only if incoming data conflicts with it; i.e., the algorithm is ***error-driven***. Suppose, for instance, that the learner is currently in grammar F, and that the language environment is A. A possible datum, now, is [ɪtɛ], with a probability of 1/18. This input conflicts with her current assumption of a highly ranked high-vowel grounding constraint, which forbids [ɪ]. The conflicting input is a ***trigger***: the learner will try a different grammar, which she chooses randomly from the set of grammars ***adjacent*** to F (i.e., the algorithm is ***conservative***), and only accept that new candidate grammar if it does allow [ɪtɛ] (the algorithm is ***greedy***).

According to the binary parameters (14.64), grammar F must be considered adjacent to B, D, and G. The adjacency of all eight grammars can be represented in a graph that connects every pair of adjacent grammars:

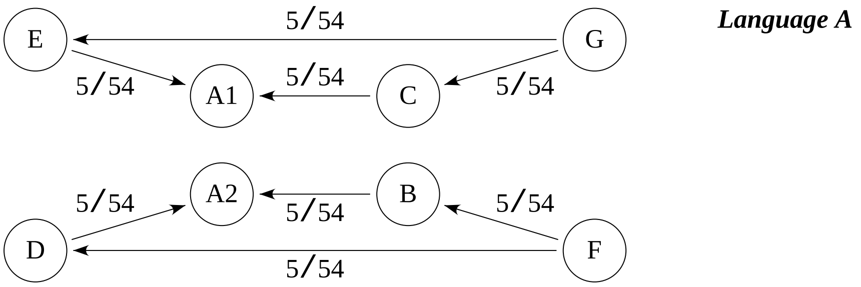


Adjacency

(14.65)

For instance, the path E-D-F-B involves three parameter flips.

So our language learner, confronted with [ɪtɛ] while her hypothesis is F, will try B, D, or G, all with probability 1/3. She will only change her grammar if the new grammar does license [ɪtɛ], i.e., if the grammar that she tries, happens to be A1, A2, D, or E; thus, she will only make the plunge if the new grammar is D. Since the data [ɛtɪ], [ɪta], [atɪ], and [ɪtɪ] trigger the same grammar change, the probability that the learner's grammar after the next datum (randomly taken from the environment A) is changed to D, is 5·1/18·1/3 = 5/54. The complete ***transition graph*** in a homogeneous A environment is



(14.66)

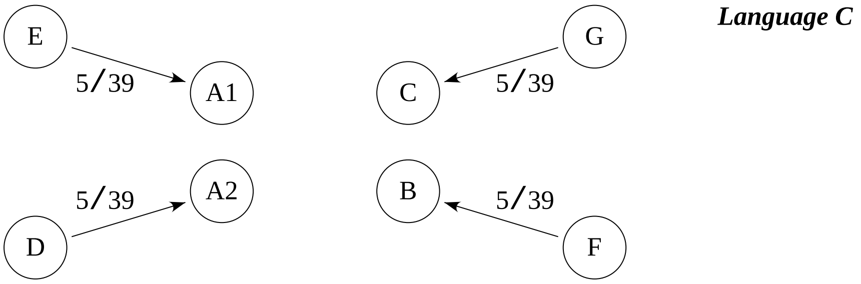
For simplicity, the self-loops are left out of (14.66): for instance, the probability that the learner, if her current hypothesis is F, will cling to F after the next random datum, is 44/54, i.e. 1 minus the sum of the probabilities of her going to B or D. If we compare figure (14.66) with (14.65), we see that the connection between the hypotheses D and E is broken, because the difference between the vocabularies of D and E consists solely of data that do not occur in A; the same goes for the pairs B-C, and F-G; the pair A1-A2 is a special case of this: in an error-driven learning scheme, the learner can never replace a grammar with a grammar that would generate the same language.

Pulleyblank & Turkel propose that "any of the set of possible languages is equally likely as the starting hypothesis of the TLA". The transition graph (14.66) shows that in the environment of language A, all grammars have a finite probability of being replaced with the grammar A after one or two steps, and that grammar A will never be abandoned once the learner has reached it. Therefore, all learners will eventually settle down in language A if the environment is A. Thus, language A will always be learned correctly. This is largely because it is the least restrictive of these languages: there is always positive data to wipe out the grounding constraints (but not the tacitly assumed harmony requirement).

14.4.3 The subset problem

According to Gold (1967), a language is ***learnable in the limit*** if the learner is guaranteed to find the target language if exposed to an infinite amount of data. Thus, language A is learnable in the limit, and I will now show that language C is not.

Suppose that a learner in a homogeneous C environment starts out in grammar A. Because language C is the intersection of the languages A and B, no data from the environment C can ever falsify the learner's hypothesis that the grammar of the language is A or B. Our learner, therefore, will never get out of the unrestrictive hypothesis A, because language C is a ***proper subset*** of language A. The complete transition graph for a homogeneous C environment is



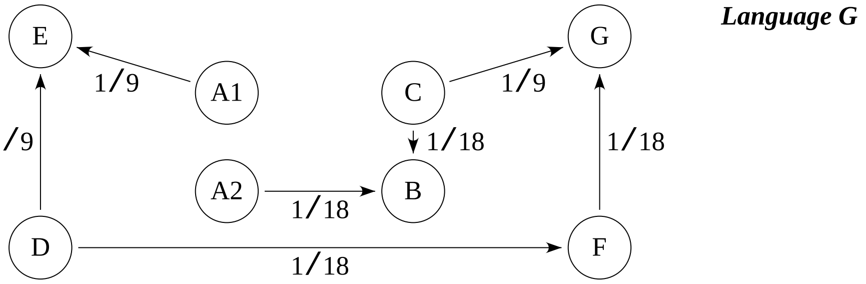
(14.67)

In the limit, 1/2 of the learners will end up in an A grammar, 1/4 in B, and only 1/4 in C. The problem with this procedure is that starting from the ATR-dominant hypothesis A, the filter grounding condition HI/ATR cannot be learned, and that starting from grammar B, the dominance of ATR cannot be learned. This non-convergence of a learning algorithm does not have to be a problem: it can be an explanation for language change (Clark & Roberts 1993). But if learning does proceed like this, we must predict that 1/2 of all learners will be willing to incorporate /ɪ/ in their words (perhaps in borrowings) because they have an A grammar. The language C, and its mirror image D, would thus be unstable, but D, according to Pulleyblank & Turkel (1995), is the Ijẹṣa dialect of Yoruba.

In an error-driven learning scheme, the non-occurrence of certain forms ("negative evidence") cannot be signalled by the learner. Starting from a uniform grammar distribution, this will lead to a high probability that the learner will end up in a too unrestrictive grammar. We can improve this number by starting with a wisely chosen initial distribution, e.g., by starting in the most restrictive grammar, i.e., a grammar with undominated grounding conditions; this would reflect the functional idea that a beginning language learner does not yet master marked feature combinations. Now, such a restrictive grammar will only be changed if positive evidence is found of the occurrence of ungrounded feature combinations. Thus, we could start in language F or G, with both grounding constraints HI/ATR and LO/RTR undominated.

If the distribution of initial states is not uniform, but a mixture of the F and G grammars, so that we have a probability of 1/2 of starting in G, the chance that we end up in the correct grammar C, is also 1/2. This is an improvement over the random initial state, but is not our ideal, because if the learner starts with hypothesis F, she will end up in grammar B, which again represents a superset of the C language.

Besides the graphs (14.66) and (14.67), we must consider the two remaining topologies. The learning graph for language G shows three sinks (B, E, and G):

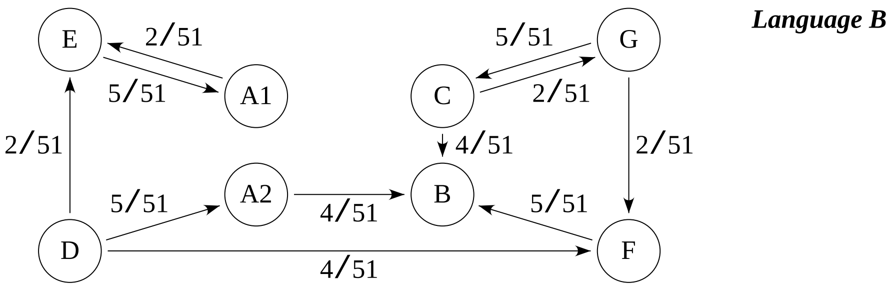


(14.68)

This is partly a subset problem again (G is a subset of E), but partly a problem of a locally optimal grammar (B, see §14.4.5).

If the distribution of initial states is uniform, the probabilities that the learner ends up in each of the eight grammars, is: 7/24 for B, 8/24 for E, and 9/24 for G. Graph (14.68) clearly shows, however, that if we allow only the most restricted initial states F and G, the learner will always end up in the correct grammar G.

The learning of the correct grammar for language B ("Wolof") is not guaranteed for a uniform initial distribution:



(14.69)

This graph also shows that a learner can go back and forth between two adjacent grammars: if she is in grammar C (so that an underlying /ita/ would have to surface as [itə]), the B-datum [ita] may cause her either to reject her hypothesis of ATR dominance (and go to B), or to hypothesize that harmony is dominated by LO/RTR (and go to G); conversely, if she is in G (which disallows [ə]) and is confronted with a B-datum that contains [ə], she will have to cancel the LO/RTR grounding condition by going to C. Graph (14.69) also shows an ***absorbing cycle***: there is no arrow pointing out from the cycle A1-E. If the learner starts in any of the restricted languages F or G, she will always end up in the correct grammar B; with a uniform initial distribution, her chances will be only be $(1+1+1+1+1+9/11)/8 = 8/11$.

The learning graphs for the three remaining languages E, D, and F, can be obtained from (14.69), (14.67), and (14.68), respectively, by replacing all instances of "B" with "E" and vice versa, and doing the same for the pairs C-D and F-G.

We can now compare the probabilities of convergence for all seven languages in the cases of the uniform and restricted initial distributions:

initial grammar ↓ target language →	A	B	C	D	E	F	G
A1A2BCDEFG	1	8/11	1/4	1/4	8/11	3/8	3/8
FG	1	1	1/2	1/2	1	1	1

(14.70)

Learnability is not perfect yet. There is no initial parameter setting that makes all languages A to G learnable. The source of this problem lies in the packaging of the tongue-root constraints. The parameter setting ATR stands (in our terms) for PARSE (atr) >> *[TR contour] >> PARSE (rtr). This is against the idea of the most restricted initial state, where the structural (harmony) constraint should dominate all faithfulness constraints (Smolensky 1996); the languages F and G, therefore, are not good candidates for the initial state.

So the initial state should have the harmony constraint on top. However, if the constraints are considered fixed, innate, and binary, they are inherently conflicting. For example, consider the underlying forms /ita/, /ɪta/, /itə/, and /ɪtə/, which, according to the principle of the "richness of the base" (Prince & Smolensky 1993), should all be possible inputs to a universal set of innate constraints. If the grounding constraints are honoured, the output in the initial state should be [ita], perhaps violating some PARSE constraints. However, [ita] violates the harmony constraint. Thus, the requirement of undominated structural constraints cannot be met in a system of binary feature values.

In contrast to the structural binary innate constraints, their three gestural correlates *[atr], *[rtr], and *[TR contour] can be satisfied all at once: just make no active tongue-root gestures at all. Underlying /ita/, /ɪta/, /itə/, and /ɪtə/ will surface as [ˈitä], and be perceived according to (14.50). With the GLA, all eight tongue-root-harmony languages are learnable. In phase 4, the grammars of languages B to G will look much like (14.58): one or two grounding constraints at the top, two PARSE constraints at the first level, and the harmony constraint and the remaining gestural constraints at the second level; language A will have the harmony constraint on top, and all grounding constraints at the second level. For all languages, step 5 will proceed in a way analogous to that described for “Wolof”.

14.4.4 Intermezzo: the correct P&P convergence criterion

Graph (14.68) illustrates a possible failure of the algorithm of Gibson & Wexler (1994), in the interpretation of Berwick & Niyogi (1996): in G&W’s algorithm (according to B&N), the initial state D would lead to the acquisition of the correct grammar G because there is a path from D to G; however, a learner starting in D could also end up in the wrong grammar E. Graph (14.69), however, shows the failure of the algorithm of Berwick & Niyogi (1996): according to their criterion, the initial state D would have to lead to the acquisition of the correct grammar B because the only absorbing state to which D is connected, is B; however, D is also connected to the absorbing *cycle* A1-E (B&N admit in a footnote that they “ignore cycles”). One of the correct ways to formalize a convergence criterion is with transition matrices, as was done by Pulleyblank & Turkel (1995) for the current problem.

Take Wolof as an example again. We can represent the general adjacency graph (14.65) with a symmetric ***adjacency matrix*** A_{ij}, and the Wolof transition graph (14.69) with a ***transition matrix*** $T_{ij}^{(B)}$:

$$A_{ij} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}; T_{ij}^{(B)} = \frac{1}{51}\begin{pmatrix} 49 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 47 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 51 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 45 & 0 & 0 & 0 & 2 \\ 0 & 5 & 0 & 0 & 40 & 2 & 4 & 0 \\ 5 & 0 & 0 & 0 & 0 & 46 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 46 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 2 & 44 \end{pmatrix} \qquad (14.71)$$

In both matrices, the rows as well as the columns enumerate the grammars A1 to G. In the adjacency matrix, adjacent pairs are represented with 1’s, non-adjacent pairs with zeroes; the diagonal is meaningless. The general formula for the off-diagonal elements of the transition matrix in (14.71) is

$$T_{ij}^{(k)} = \frac{A_{ij}}{N_i}\sum_m p_m^{(k)} \cdot \left(1 - l_m^{(i)}\right) \cdot l_m^{(j)} \qquad (14.72)$$

where

$T_{ij}^{(k)}$ = the average probability that the learner goes from grammar *i* to grammar *j* as a result of a randomly selected datum from language environment *k*.

$p_m^{(k)}$ = the probability of the perception of the datum *m* given the language environment *k*: the value in the *k*th column of the *m*th row in table (14.63). In the case of perception or production errors (Pulleyblank & Turkel 1995, 1996), this should be replaced with $\sum_n \overline{p}_n^{(k)} p_{nm}$, where $\overline{p}_n^{(k)}$ is the probability of the intended utterance *n*, and p_{nm} is the probability that the intended utterance *n* is perceived as the datum *m*.

$l_m^{(i)}$ = unity if grammar *i* allows the datum *m*, zero otherwise; can be rounded up from the value in the *i*th column of the *m*th row in table (14.63). In (14.72), the factor $\left(1-l_m^{(i)}\right)$ represents error-drivenness (it determines whether *m* is a trigger), and the factor $l_m^{(j)}$ represents greediness.

N_i = the number of grammars adjacent to grammar *i*; equal to $\sum_{j\neq i} A_{ij}$. In (14.72), the factor A_{ij}/N_i represents conservatism; A_{ij} is the single-value factor, and N_i is the single-try factor.

The elements on the diagonal are such that the elements of every row add up to unity:

$$T_{ii}^{(k)} = 1 - \sum_{j\neq i} T_{ij}^{(k)} \tag{14.73}$$

If the initial state is a randomly chosen grammar, the probability $P_i^{(k)}$ that the learner has the grammar *i* in the language environment *k*, is initially (for our eight grammars)

$$\mathbf{P}^{(k)}(0) = \begin{pmatrix} 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \end{pmatrix} \tag{14.74}$$

If the initial state must be one of the most restricted grammars F or G, its distribution is

$$\mathbf{P}^{(k)}(0) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix} \tag{14.75}$$

The probability of finding the learner in grammar *k* after a single exposure to a datum is

$$P_i^{(k)}(1) = \sum_j P_i^{(k)}(0)\, T_{ij}^{(k)} \qquad \text{or} \qquad \mathbf{P}^{(k)}(1) = \mathbf{P}^{(k)}(0)\cdot \mathbf{T}^{(k)} \tag{14.76}$$

The grammar distribution within the population of learners who have been exposed to *n* data, is given by

$$\mathbf{P}^{(k)}(n) = \mathbf{P}^{(k)}(0)\cdot \left(\mathbf{T}^{(k)}\right)^n \tag{14.77}$$

If the target language is B, the limit distribution in case of learnability should be

$$\mathbf{P}^{(B)}(\infty) = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{14.78}$$

If we start in a uniform distribution, we can only arrive at such a distribution if, as *n* goes to infinity, the *n*-th power of the transition matrix goes to a matrix with eight equal rows with ones in the third column and zeroes elsewhere. If we start in F or G, the requirement is only that the 7th and 8th rows end up in this way.

The convergence of the transition matrix, then, is one of the correct criteria for the learnability of a principles-and-parameters system. It will identify fewer target languages as learnable than either Gibson & Wexler’s or Berwick & Niyogi’s algorithms.

14.4.5 Local maxima

The current hypothesis of the learner may be a grammar from which she cannot get out because all adjacent grammars are worse. This is a problem of conservatism and greediness together: if you are at a hill top and want to reach the top of the higher mountain instead, you can choose between jumping the valley or climbing down; being conservative will disallow the former, and being greedy will disallow the latter.

First, we may note that the existence of local maxima does not have to be a problem for learning theories. Perhaps languages that allow learners to end up in local maxima actually exist; they would just be unstable, since a part of the next generation speaks a different language. If this new language is learnable without local maxima, the original language will die out in a few generations.

But language G, shown in graph (14.68), is like Pulaar, according to Pulleyblank & Turkel (1995). If it does not die out, it should be learnable. However, suppose that the learner's current hypothesis is B. According to table (14.63), she will only be urged to change her grammar if the incoming datum is [ate]. Conservatism tells us that she cannot go to G directly, because she would need to flip two parameters. Greediness, however, tells us not to go to the adjacent languages A2, C, or F, because these do not allow [ate].

14.4.6 Relaxing conservatism or greediness

To tackle local maxima, the validity of both the conservatism constraint and the greediness constraint was challenged by Berwick & Niyogi (1996):

> "if the learner drops either or both of the Greediness and Single Value Constraints, the resulting algorithm (...) converges faster (in the sense of requiring fewer examples) than the TLA" (p. 607)

They go on to show that this statement is true for Gibson & Wexler's (1994) example of a three-parameter word-order problem.

However, Berwick & Niyogi's criticism does not seem to hold for larger parameter spaces. We can see this if we compute the expected number of examples (data) the learner needs to arrive at the target grammar, in G&W's as well as in B&N's algorithm.

In B&N's algorithm, the conservatism and greediness factors are not used, i.e., the transition matrix (14.72) would reduce to

$$T_{ij}^{(k)} = \frac{1}{N-1} \sum_m p_m^{(k)} \cdot \left(1 - l_m^{(i)}\right) \qquad (14.79)$$

where N is the number of grammars. In a space of N_P binary parameters, N would be equal to 2^{N_P}. Assuming that the target language can only be described by one of these grammars, and taking into account the Markov (oblivious) character of the sequence of steps, the average number of triggers (conflicting data) required before reaching the target grammar (from any non-target grammar), is $2^{N_P} - 1$. For a three-parameter space, this is 7, but for a more realistic 30-parameter space, this is 1,073,741,823.

For Gibson & Wexler's algorithm, we can compute the convergence under the simplifying assumption that the parameters have independent influences on the language. In such a case, the worst initial state is one in which all N_P binary parameters have the wrong value. With every trigger, there is a probability of $1/N_P$ that the correct parameter change will be chosen. Thus, the target grammar will be reached after at most N_P^2 triggers. For a three-parameter space, this is 9, which may be worse than in Berwick & Niyogi's algorithm, but for a 30-parameter space, this is only 900, to which Berwick & Niyogi's alternative constitutes a drastic deterioration.

With respect to the number of data needed to reach the target grammar, the difference between the two algorithms is somewhat less dramatic, because in the TLA the probability that a datum is a trigger decreases as the hypothesis approaches the target grammar. Even if the number of data thus becomes proportional to N_P^3, a polynomial dependence of the learning time will always outperform an exponential dependence as soon as realistic degrees of freedom are involved.

We must conclude that without the conservatism and greediness constraints, the acquisition time of realistic grammar spaces would be prohibitively large.

14.4.7 Genetic algorithms

The local-maxima problem is smaller if the learner is allowed to consider multiple hypotheses at a time (Clark & Roberts 1993; Pulleyblank & Turkel, to appear). For example, suppose the learner arrives at two different local maxima X and Y in a 20-parameter space. Apart from the usual parameter swappings within the hypotheses X and Y ("mutations"), which will not help her out of the trap, she will have the option of creating a new hypothesis Z that copies, say, 11 parameter settings from X and the remaining nine from Y (a "recombination"). If this hypothesis is better than X and Y, the learner will have succeeded in getting out of a local maximum without sacrificing either conservativeness or greediness.

As everyone acknowledges, this approach still does not guarantee convergence onto the global maximum. Moreover, it places a large burden on the learner, who has to maintain several, possibly very distinct, hypotheses. By contrast, the Gradual Learning Algorithms are guaranteed to converge, even with a single hypothesis.

14.4.8 TLA versus GLA

The most obvious difference between the Triggering Learning Algorithm and our Gradual Learning Algorithm is convergence.

Another possible source of concern is the lack of robustness of the TLA: one erroneous input will change the setting of a parameter, and if we arrive in a superset language, we will not be able to get out. The algorithm could easily be made gradual, though, by replacing binariness with a continuous scale between, say –100 and +100, while every learning step would push the parameter value in the direction of the other side, by a small step; to resist errors, the learning process would have to supplement this with a slow opposite drift.

A fundamental problem with our GLA seems the following. To arrive at the correct grammar, the learner will need to know underlying forms, at least in step 6. This seems like a dirty trick. The GLA has this in common with the OT learning algorithms of Tesar & Smolensky (1993, 1996) and Tesar (1995), which have been criticised for this reason by Turkel (1994):

> "From the point of view of a model of language acquisition, the assumption of having the optimal parse available as part of the input is problematic." (Turkel 1994: 7)

In learning Wolof, for instance, a learner can only reintroduce the workings of the harmony constraint if she encounters underlying forms like |at+e| that should surface as /atɛ/ (the combination of these two forms is the "optimal parse").

We could reverse the argument. The question is how a P&P learning process of Wolof would handle this. If a P&P learner is in a superset language, like phase-4 Wolof (14.58) is as compared with adult Wolof, how could she ever learn that the surface form [ate] is forbidden? The answer is that she cannot. In the GLA, on the other hand, all learners of "Wolof" arrive in this superset language, and if there are alternations of the form described, they *will* adopt the more restrictive hypothesis.

14.5 Optimality-theoretic learning

If parameter grammars have learning algorithms that fail in local maxima, and we cannot do without conservatism and greediness, there is but one conclusion: grammars are not built around parameters.

The alternative, of course, is that grammars are built around ranked constraints. Tesar & Smolensky (1993) prove that OT grammars are learnable with a number of steps that goes as the square of the number of constraints, and without ever getting stuck in a local maximum. This is as fast as the grammars with independent parameters as described in §14.4.6.

In Tesar & Smolensky's original algorithms, the necessary losing candidate was randomly supplied by GEN. In later work, Tesar (1995) and Tesar & Smolensky (1996) propose that this loser is to be identified with the correct adult output form: Error-Driven Constraint Demotion (EDCD). An example of the workings of this algorithm is given in our footnote below tableau (14.34). EDCD is almost as simple as our Minimal algorithm, and it shows convergence, conservatism, and oblivion.

Some differences remain, though: EDCD is not very ***robust*** against errors: a single error may destroy the grammar in such a way that it can cost on the order of N^2 learning steps (N is the number of constraints) to climb out of the wrong grammar (though in practice, a typical number is N). In the Minimal GLA, the number of learning steps needed after an error is on the order of N (actually, approximately one half of the depth of the grammar). In the Maximal GLA, the number is exactly 1: after taking an incorrect adult uterance at face value, the constraints are shifted by a little amount, so that the probability of an error increases somewhat; once such an error occurs, the same constraints are shifted back to their original positions. Finally, EDCD does not show the flexibility or realism of an algorithm based on a continuous scale (§14.2.9).

14.5.1 The initial state in an Optimality-theoretic grammar

In Tesar & Smolensky's (1993) initial state, all constraints are born equal. But this raises some problems.

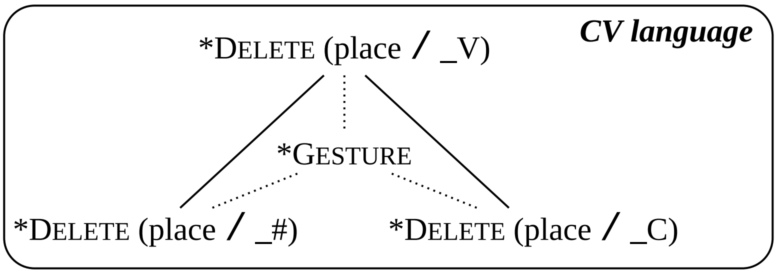
Consider T&S's example of a language that only allows CV syllables. If an underlying form /CVCVC/ surfaces as [CVCV], this is a sign that NOCODA is honoured. But presumably, some of those languages have no underlying codas in the lexicon. Still, according to the principle of ***richness of the base*** (Prince & Smolensky 1993; Smolensky 1996), NOCODA must be high-ranked in these languages, and some evidence for this is found in the adaptation of loan words, which will either lose their codas or be supplied by an epenthetic vowel. But how should anyone be able to learn this ranking? The only evidence that the learner is confronted with, is /CVCV/ → [CVCV]. In Tesar & Smolensky's algorithm, NOCODA will still vacuously come out on top, because it is not violated in any winner, and thus never demoted: the default position for a constraint in the hierarchy is at the top. Thus, invisible is undominated, just as with the gestural constraints described earlier.

But e.g. *DELETE (click) will come out on top, too, and still we would imagine that an underlying [!a] would surface as [ka] (i.e., a heard [!a] would be pronounced as [ka], even if the click were categorized as a click). Therefore, the default ranking for faithfulness constraints should *not* be at the top. To remedy this situation, Smolensky (1996) and Tesar & Smolensky (1996) propose that the initial state should have all structural constraints dominating all faithfulness constraints. This, as we saw, is the generic solution to the subset problem, and is reminiscent of the SUBSET PRINCIPLE (Berwick, 1985; Wexler and Manzini, 1987). But this solution had to be *posited*. By contrast, a functional division between constraints can derive it, as we have seen: the difficulty of an articulatory gesture decreases as it is practised, and the importance of a perceptual feature increases as it is practised. The real solution, therefore, is that constraints are learned, not innate.

14.5.2 Innateness

In Tesar & Smolensky's algorithm, it is crucial that NOCODA is a universally available constraint: even though (or because) the learner has never had to learn its ranking, she must know that it is at the top of the hierarchy, or it could not do its work in determining the surface shape of consonant-final loan words. Therefore, NOCODA must be ***innate***.

A functional theory of phonology can hardly accept the innateness, or even the existence, of a constraint like NOCODA: it must be an epiphenomenon of an interaction between gestural and faithfulness constraints. As the articulatory effort of an initial consonant cannot be much different from the effort of a final consonant, the asymmetry must lie in perceptual confusion. Because some place cues, for instance, depend on transient effects like release bursts, the average contrast between initial consonants is greater than the average contrast between final consonants. Together with the fact that the place cues of intervocalic consonants are better than those of consonants adjacent to other consonants, this leads to a preference of CV over VC syllables. The relevant constraint ranking, therefore, is something like:



(14.80)

The empirical consequences of this ranking are different from those of the single NOCODA constraint. It predicts, for instance, that coda-avoiding languages tend to have simple onsets, i.e., that languages with an apparent high-ranked NOCODA also appear to have a high-ranked *COMPLEXONSET.

The grammar of (14.80) can be learned by the usual demotion of *GESTURE from the top, given a local ranking of *DELETE (place / _V) >> *DELETE (place / _C). If this local ranking is valid, grammar (14.80) will have been arrived at in stage 4, without the need for underlying forms with codas. Only if a learner accidentally manages to get *GESTURE below all the *DELETE constraints, she will need the evidence of step 5 to learn that underlying codas do not appear in the output.

14.6 Algorithm

I will now show in pseudocode how you could simulate the handling of a single learning pair for the gradual learning algorithms if you already have a classical tableau-oriented evaluation algorithm. For tongue-root-harmony languages, for instance, you would only use the two grounding constraints, the two *DELETE constraints, and the harmony constraint.

We start with a hypothesized grammar *H*, consisting of an unordered constraint set $\{ C_i \}$, $i = 1...N$. Every constraint is assumed to have a ranking value.

(a) Generate an adult utterance. You could draw it randomly from a vocabulary list, like table (14.63), or compute it from a random input evaluated in the target grammar like:

```
adultInput := get_random_input                              // from richness of the base, for instance
adultOutput := get_winner (targetGrammar, adultInput)       // classical GEN, H-EVAL, etc.
```

(b) Compute the learner’s underlying form for this utterance.

```
if age ≥ UNDERLYING_FORM_START_AGE
    learnerInput := adultInput
else
    learnerInput := adultOutput
```

Instead of just copying the adult output, the young learner could try robust interpretive parsing (Smolensky 1996a; Tesar & Smolensky 1996), or possibly an iterative version of it (Tesar 1996, 1997, to appear), modified, of course, to include stochastic evaluation.

(c) Compute the disharmonies for all constraints:

```
for $i$ from 1 to $N$                       // $N$ is the number of constraints
    $C_i$.disharmony := $C_i$.ranking +
        + RANKING_SPREADING * z    // z is a Gaussian random deviate, with $\mu = 0$ and $\sigma = 1$
```

(d) Sort the constraints by disharmony from high to low.

(e) Compute the learner's output with your favourite OT implementation:

```
learnerOutput := get_winner ($H$, learnerInput)
```

(f) Find the offending mark and constraint (Minimal GLA):

```
if learnerOutput ≠ adultOutput     // error-driven
   i := 1                          // search for crucial mark; start at topmost constraint
   while number_of_marks ($H$, adultOutput, $C_i$) = number_of_marks ($H$, learnerOutput, $C_i$)
      i := i + 1                   // skip equal violations
   demote_constraint ($C_i$)
```

Or adjust the rankings of all constraints (Maximal GLA):

```
if learnerOutput ≠ adultOutput     // error-driven
    for $i$ from 1 to $N$
        adultMarks := number_of_marks ($H$, adultOutput, $C_i$)
        learnerMarks := number_of_marks ($H$, learnerOutput, $C_i$)
        if adultMarks > learnerMarks
            demote_constraint ($C_i$)
        else if learnerMarks > adultMarks
            promote_constraint ($C_i$)
```

The demotion procedure is recursive:

```
procedure demote_constraint ($C$)
    demotionStep := DAY_ONE_PLASTICITY
        * (1 + RELATIVE_PLASTICITY_SPREADING * z)        // z is Gaussian (0, 1)
        * (0.5 ** (age / PLASTICITY_HALF_TIME))
    $C$.ranking := $C$.ranking – demotionStep
    if age < LOCAL_RANKING_SUSPENSION_AGE
        for all $C_i$ that are locally ranked below $C$
            while $C_i$.ranking ≥ $C$.ranking
                demote_constraint ($C_i$)
```

The promotion procedure is analogous to this, with "–", "below", and "≥" replaced by "+", "above", and "≤".

(g) Sort the constraints by ranking value from high to low.

14.7 Proof of learnability

I will now prove the intimate relationship with between the Minimal Gradual Learning Algorithm and a grammar that consists of a set of crucially ranked pairs of constraints.

Definition. “Two ***grammars*** are ***equivalent*** if they give equal outputs for any thinkable phonological input.” (14.81)

Definition. “Two ***languages*** are the ***same*** if their grammars are equivalent.” (14.82)

This is a more restricted definition than the usual definition that they are the same if they have the same set of possible utterances; we need this definition because we are focusing on the learning of the grammar, and ignore the possibility of differences in the underlying forms.

Definition. “A ***crucially ranked pair*** is a pair that would give rise to a different language if their rankings were reversed.” (14.83)

Definition. “A ***crucial tie*** is a pair of constraints (with equal rankings) whose marks would be able to cancel each other.” (14.84)

Assertion. “Crucial ties are not allowed.” This is (14.20). (14.85)

Assertion. “Crucial ranking is ***transitive***.” If A >> B, and B >> C, then A >> C. (14.86)

Assumption. “Inputs that would result in the conclusion that two languages are not the same, are not withheld from the learner.” (Tesar & Smolensky 1996) (14.87)

Statement. “Every total ranking can be seen as a set of crucially ranked pairs.” (14.88)

Proof. The equivalent set is constructed as follows. Every pair of adjacent constraints constitutes a crucial ranking. For instance, the total ranking A >> B >> C >> D is equivalent to the set of crucially ranked pairs { A >> B, B >> C, C >> D }. Because of transitivity, pairs like A >> C can be vacuously included in the set.

Statement. “Every grammar hypothesized at any moment during the course of the Minimal Learning Algorithm, can be described with a set of crucially ranked pairs of constraints.” (14.89)

Proof. A set of constraints with different ranking values can always be seen as a total ranking without ties. For instance, the set { A, B, C, D } with respective ranking values of 40, 20, 80, and 60, can be seen as the total ranking C >> D >> A >> B. This is equivalent to the set of crucially ranked pairs { C >> D, D >> A, A >> B }. If two ranking values are equal, we can choose from two total rankings; these give the two possible interpretations of the “tie”. Note that this is very different from Tesar & Smolensky’s strata, which contain constraints that can cancel each other’s marks. Thus:

Corollary. “If all sets of crucially ranked pairs generate possible languages, the learner’s current hypothesis always describes a possible language.” (14.90)

In the rest of the proof, we consider a discretized version of the Minimal GLA: constraints are ranked along a discrete scale with integer values (e.g., 100, 99, 98, and so on), the demotion step is always 1, and there is no fuzzy ranking. As we are now going to prove the correct convergence of the Minimal GLA, the term “crucial ranking” always refers to the target (adult) grammar, and the terms “demotion”, “scale”, and “ranking value” refer to the learner’s grammar.

Statement. "If the algorithm converges (i.e. stops changing the grammar from a certain moment on), the resulting grammar will be a possible grammar of the target language." (14.91)

Proof. Suppose that the resulting grammar were not equivalent to the target grammar. Then there must exist an input for which the two grammars produce different outputs. According to assumption (14.87), the learner will in due time discover this input. The winner will then be different from the correct output, and one constraint will be demoted. Contradiction. So the algorithm has not converged yet. This leads to a remarkable result:

Corollary. "The algorithm will never become trapped in a local maximum." (14.92)

Statement. "Neither in the incorrect winner, nor in the correct output, are there any protesters (§14.2.1) with higher ranking values than the constraint that incurs the offending crucial mark (in the pair comparison)." (14.93)

Proof. If there were any of those protesters, the offending mark would not be crucial.

Statement. "An offending constraint must be crucially ranked (directly or by transitivity) below one or more constraints with lower ranking values." (14.94)

Proof. Suppose that the offending constraint were crucially ranked above all constraints with lower ranking values. In the target grammar, the incorrect winner must be less harmonic than the correct output, so the incorrect winner must endure some protesters, and according to (14.93), these must have lower ranking values than the offending constraint. In the target grammar, one of these protesters must incur the crucial mark on the incorrect winner of the learner's grammar, if that is compared with the correct output. To be able to incur the crucial mark, it must be ranked higher than the other protesters, including the original offending constraint. Contradiction. Thus:

Corollary. "A constraint can only be demoted if it is crucially ranked below one or more other constraints that currently have a lower ranking value." (14.95)

Definition. "A ***top constraint*** is a constraint that is not crucially ranked below any other constraint." (14.96)

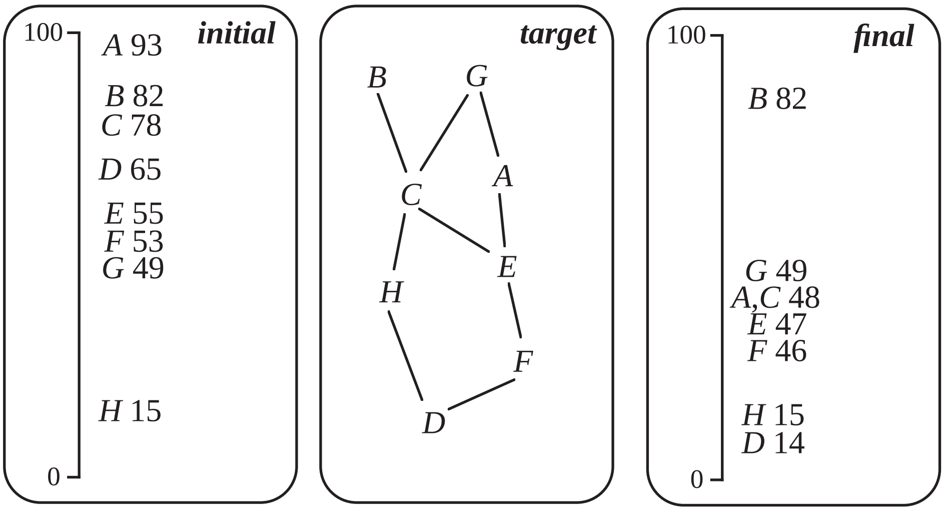
Statement. "A top constraint will never be demoted." (14.97)

Proof. Follows from (14.95) and (14.96).

Statement. "If all the dominators of a constraint have fixated ranking values (i.e., they will never be demoted again), and the ranking value of this constraint is below all the ranking values of these dominators, then this constraint will never be demoted again." (14.98)

Proof. Because of (14.95), the constraint cannot be demoted now. The fixation ensures that this situation will last forever.

We can tell beforehand where the algorithm will stop. The following is an example of a starting grammar, a target grammar, and the grammar on which the algorithm converges:



(14.99)

In the final grammar, the top constraints (B and G) still have their initial rankings 82 and 49. Constraint A, which started at 93, had to be demoted below its dominator G, and demotion stopped when it reached 48, which is one lower than the final ranking of its dominator. Constraint C, starting at 78, had to be demoted below both its dominators B (82) and G (49), and had, therefore, to end up at a ranking value of 48. Constraint H had to end up lower than C, but as C ended at 48 and H had already started at 15, no demotions had to take place. Constraint E had to end below the end rankings of A and C, so it had to fall from 55 to 47. Constraint D, finally, had to be demoted past both H and F, so it had to be come all the way down from 65 to 14. By whatever route, the number of errors needed to come from the initial to the final grammar, is (93–48) + (82–82) + (78–48) + (65–14) + (55–47) + (53–46) + (49–49) + (15–15) = 141.

The ranking of each constraint in the final grammar is the minimum of its initial ranking and one less than the minimum of the final rankings of its dominators.

Definition. “The ***depth*** of a constraint is the length of the longest path up from it to a top constraint.” Thus, in figure (14.99), B and G have a depth of 0, A and C have a depth of 1, E and H have a depth of 2, F has a depth of 3, and D has a depth of 4 (= the length of a route via E and F, not the shorter route via H). (14.100)

Corollary. “The depth of a constraint is one more than the maximum of the depths of its direct dominators.” For instance, the depth of D is 4, because the maximum of the depths of its direct dominators H (2) and F (3) is 3. (14.101)

Definition. “The depth of a grammar of crucial rankings is the maximum of the depths of its constraints.” The depth of (14.99) is 4. (14.102)

Statement. “A constraint at depth 1 can never be demoted once it is ranked below the top constraints.” (14.103)

Proof. The top constraints are fixated. A constraint at depth 1 has no other dominators than these top constraints, so must be fixated itself according to (14.98).

Corollary. “A constraint at depth 1 will never become lower ranked than the minimum of the rankings of its top constraints minus 1, unless it was already ranked lower in the initial grammar.” (14.104)

Coroll. “The number of possible demotions of constraints at depth one is finite.” (14.105)

Statement. “The number of possible demotions of constraints at any depth is finite.”(106)

Proof. This statement has been proven for a depth of 1. Now suppose it is valid for all depths from 1 to a certain depth *n*. After the finite number of demotions of constraints at depths 1 to *n* have been performed, the constraints at depth 1 to *n* are fixated (by definition). Now consider a constraint at depth *n* + 1. All its dominators are fixated. According to (14.98), once its ranking value is below that of all its dominators, the constraint will be fixated itself. In fact, it will never become lower ranked than the minimum of the ranking values of its fixated dominators minus 1, unless it was already ranked lower in the initial grammar. Therefore, if the statement is valid for all depths from 1 to n, it is also valid for depth *n* + 1. The rest follows by mathematical induction.

Corollary. “The number of possible demotions from the initial grammar is finite.” For (14.99), this number is 141. (14.107)

Statement. “For every language that can be described with a set of crucially ranked pairs of constraints, the Minimal Gradual Learning Algorithm converges.” For (14.99), the changes stop after 141 demotions. (14.108)

Proof. As the number of possible demotions is finite, demotion cannot go on forever. If it cannot go on forever, it will stop at some time in the future. In other words, the algorithm converges.

Statement. “For every language that can be described with a set of crucially ranked pairs of constraints, the Minimal Gradual Learning Algorithm, given those constraints, converges upon a possible grammar of that language.” (14.109)

Proof. Combine (14.108) with (14.91).

The correctedness of the Maximal Gradual Learning Algorithm is shown in §15.4.

14.8 Acquisition time

Suppose that all *N* constraints start out with the same ranking, and the target grammar is totally ranked. To reach its target ranking, the average constraint will have to travel up or down by a distance of $\frac{1}{4}N$ multiplied by the safety margin (the minimal stable distance between two adjacent crucially ranked constraints), divided by the plasticity. For instance, if the plasticity is about five percent of the ranking spreading, it will be about one percent of the resulting safety margin. With *N* constraints, the minimum number of constraint rerankings is $\frac{1}{4}N \cdot N \cdot 100$.

During acquisition, however, the number of non-triggers increases. When all rerankings but one have been performed, only one constraint pair out of the total of $\frac{1}{2}N(N-1)$ pairs is out of rank, and the probability of finding it on the next learning pair may well be as small as 1 part in $\frac{1}{2}N(N-1)$ (from the schwa-licensing example of

§14.3.7, we see that the last step of learning simplified Wolof may involve a probability of 1/17 of encountering the disambiguating /ətə/). The acquisition time, therefore, scales as the fourth power of *N*. For 100 constraints, the minimum number of required constraint evaluations is on the order of $10N^4$ = 1,000,000,000. Since constraints are often reranked in the wrong direction, the real number of constraint evaluations will be higher than the minimum by a factor that our simulations show to be consistently around 3. However, the average learning step will rerank three constraints, so the expected number of data needed to convergence upon a 100-constraint totally-ranked target grammar is about 10^9.

For a grammar of crucial rankings, the situation greatly improves. The acquisition time scales as the fourth power of the ***depth***. If a grammar with 100 constraints has a depth of 5, the ***width*** of the grammar is approximately $N/(depth+1)=100/6$, and the expected number of required data before convergence is on the order of $width \cdot 10 \cdot (depth+1)^4 = 216{,}000$, a marked improvement over the total-ranking case.

If the demotion procedure honours the local-ranking principle, the effective depth of the grammar decreases. If it becomes 4 instead of 5, the number of required data is on the order of $216{,}000 \cdot (5/6)^4 \approx 104{,}000$ (the width did not change). Thus, local ranking may reduce the acquisition time with a factor of 2 or so.

At 36 pieces of data a day, the required 104,000 data are provided in eight years; after one half of that time, the grammar will on the average have been acquired up to a depth of 4, and the learner has the remaining four years to acquire the deepest level.

Apparently, large segmental grammars can be learned even with a modest plasticity and a low degree of exposure to language data.

14.9 Conclusion

Making a principled distinction between articulatory and perceptual constraints within Functional Phonology leads to a straightforward learning process in which all articulatory constraints enter at the top of the hierarchy, and all faithfulness constraints enter at the bottom. The procedure moves on by promoting faithfulness constraints in the process of the acquisition of perceptual categorization, and demoting gestural constraints in a process of motor learning, aided by the bootstrapping power of play and temporary variation of constraint ranking. Continuous constraint families are handled with the help of the local-ranking principle, which ensures that falling constraints must push along their locally easier or less contrastive neighbours. In this way, universal markedness relations in adult phonology come to reflect the child's acquisition order. Under an error-driven learning scheme, all segmental phonological grammars are learnable. Thus, the Gradual Learning Algorithm is the first constraint-sorting algorithm that can be connected to the actual acquisition process.

This chapter showed that any segmental constraint set can be learned, without assuming any set of innate constraints, and that any segmental phonology can be learned unambiguously. Universal Grammar appears to contain no substance; the main innate things in phonology seem to be the desire and the ability to learn articulatory and perceptual features, the propensity to organize functional principles into an Optimality-theoretic grammar, and the large plasticity in the acquisition of this grammar.

15 How we learn variation, optionality, and probability[1]

Abstract. Variation is controlled by the grammar, though indirectly: it follows automatically from the robustness requirement of learning. If every constraint in an Optimality-Theoretic grammar has a ranking value along a continuous scale, and the disharmony of a constraint at evaluation time is randomly distributed about this value, the phenomenon of optionality in determining the winning candidate follows from the finiteness of the difference between the ranking values of the relevant constraints; the *degree* of optionality is a descending function of this difference.

In the *production grammar*, a symmetrized maximal gradual learning algorithm will cause the learner to copy the degrees of optionality from the language environment. In the *perception grammar*, even the slightest degree of noise in constraint evaluation will cause the learner to become a probability-matching listener, whose categorization distributions match the production distributions of the language environment. Evidence suggests that natural learners follow a symmetric demotion-and-promotion strategy, not a demotion-only strategy.

A typical example of optionality in speech production is place assimilation of nasals at the sentence level, i.e. a word underlyingly ending in |an| and a word starting with |pa| may, when concatenated, be pronounced either as [anpa] or as [ampa]. This poses a problem for a theory with fixed relative constraint rankings, like the original version of Optimality Theory (Prince & Smolensky 1993).

Let’s say that the relevant constraints for our example are *GESTURE (tongue tip: close & open) and *REPLACE (place: coronal, labial / nasal / _C), i.e., the choice between [anpa] and [ampa] is the outcome of a struggle between the importance of not performing a tongue-tip opening-and-closing gesture and the importance of honouring an underlying specification for the value [coronal] on the perceptual place tier as conditioned by a nasal environment before a consonant (chapter 11). The candidate [anpa] would win if the ranking were *REPLACE (cor) >> *GESTURE (tip):

|an+pa|	*REPLACE (cor)	*GESTURE (tip)
☞ [anpa] /anpa/		*
[ampa] /ampa/	*!	

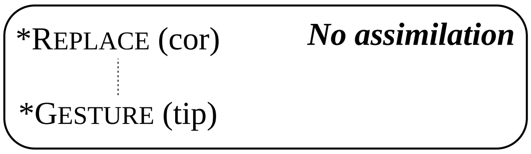
(15.1)

A short explanation of the notation may be appropriate. According to the ideas of Functional Phonology, the gestural constraint evaluates the articulatory candidate [anpa], and the faithfulness constraint evaluates the difference between the underlying perceptual specification |an+pa| and the output /anpa/, which is the acoustic result of [anpa] as perceived by the listener; the similarities between these forms are deceptive: the brackets

[1] This chapter appeared as Boersma (1997b).

contain shorthand notations for articulatory events, the slashes contain shorthands for perceptual features.

If (15.1) were the only possible outcome, we could describe it with the following grammar (the dotted line depicts a language-specific crucial ranking):



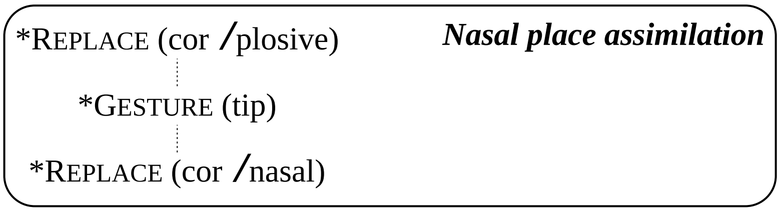
(15.2)

With the reverse ranking, [ampa] would win:

|an+pa|	*GESTURE (tip)	*REPLACE (cor)
[anpa] /anpa/	*!	
☞ [ampa] /ampa/		*

(15.3)

With the following grammar, nasals would assimilate, but plosives would not:



(15.4)

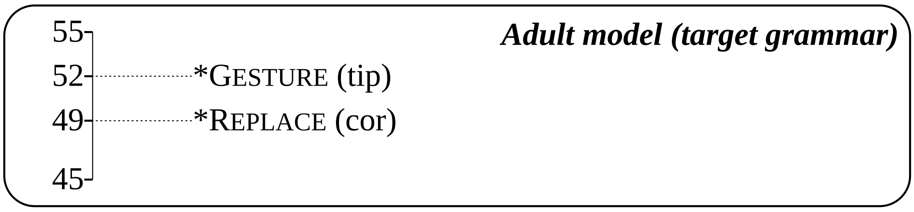
If place assimilation is optional, and if we cannot have both grammars (15.2) and (15.4) at the same time, then we have a problem.

One possibility would be to rank *REPLACE (cor) and *GESTURE (tip) equally high. We should then not follow the suggestion by Tesar & Smolensky (1993), who interpret equal ranking in such a way that the violation marks incurred by the two constraints are capable of cancelling each other. Rather, we should interpret equal ranking in a probabilistic manner: if the two constraints are in conflict, either of them could win at evaluation time, both with a probability of 50% (Anttila 1995). However, real life learns us that optionality is often gradient, e.g., one form may occur in 80%, the other in 20% of the cases, and these numbers differ between neighbouring dialects. The proposal of the current paper shows that a continuously ranking OT grammar can maintain any degree of optionality, that speakers will learn to reproduce the degree of optionality of their language environment, and that listeners will learn to match the degree of optionality of their language environment in their categorization systems.

15.1 Continuous ranking scale and stochastic disharmony

Our optionality problem is solved by a random stochastic element in constraint evaluation (in §15.2.4, we will see that this random element is independently needed to implement the robustness requirement of a natural language user's learning strategy).

If place assimilation occurs more often than not, we say that *GESTURE (tip) is ranked higher than *REPLACE (cor) along a continuous scale (whose physiological correlate could be synaptic strength), with a real number attached to each constraint:



(15.5)

In this example, the ranking value of *REPLACE (cor) is 49, and the ranking value of *GESTURE (tip) is 52. In the absence of stochastic evaluation these values would determine the order of the constraints in an evaluation tableau, in which case this ranking would be equivalent to grammar (15.4). However, with stochastic evaluation (whose physiological correlate could be the noise in the amount of locally available neurotransmitter), this order is determined by the ***disharmonies*** ("effective" rankings) of the constraints, which are determined at evaluation time from the ranking value and a random variable:

$$disharmony = ranking + rankingSpreading \cdot \mathbf{z} \qquad (15.6)$$

where **z** is a Gaussian random variable with mean 0 and standard deviation 1. For instance, a simulation of ten implementations of |an+pa| with a *rankingSpreading* of 2 yielded the following disharmonies:

trial	1	2	3	4	5	6	7	8	9	10
*GESTURE disharmony	50.5	51.2	50.2	49.1	52.9	52.9	52.7	53.8	55.4	54.3
*REPLACE disharmony	50.8	48.3	50.7	51.2	48.9	48.8	48.2	50.3	48.1	48.7
outcome	np	mp	np	np	mp	mp	mp	mp	mp	mp

(15.7)

We see that in most replications, *GESTURE (tip) was evaluated as higher than *REPLACE (cor), but that *REPLACE (cor) was higher in three of the ten cases. Thus, our simulated speaker would have said [ampa] seven times, and [anpa] three times. The distribution of the disharmony difference between two constraints C_1 and C_2 with rankings r_1 and r_2 is given by

$$disharmony_1 - disharmony_2 = r_1 - r_2 + rankingSpreading \cdot (\mathbf{z}_1 - \mathbf{z}_2) \qquad (15.8)$$

Now if both $\mathbf{z}_1$ and $\mathbf{z}_2$ are Gaussian distributions with standard deviations of 1, their difference $\mathbf{z}_1 - \mathbf{z}_2$ is also Gaussian, with a standard deviation of $\sqrt{2}$, so that the probability that C_1 is evaluated higher than C_2 is

$$P(disharmony_1 > disharmony_2) = \tfrac{1}{2} \cdot \left(1 - \mathrm{erf}\left(\tfrac{1}{2}\sqrt{2} \cdot \frac{r_1 - r_2}{rankingSpreading \cdot \sqrt{2}}\right)\right) \qquad (15.9)$$

which for a ranking spreading of 2 can be tabulated as

$r_1 – r_2$	0	1	2	3	4	5	6	7	8	9	10	11	12
P	1/2	36%	24%	14%	7.9%	3.9%	1.7%	0.7%	0.2%	$7\cdot10^{-4}$	$2\cdot10^{-4}$	$5\cdot10^{-5}$	$1\cdot10^{-5}$

(15.10)

So our speaker will say [anpa] 14% of the times. If the ranking difference is less than 10 (or so), we may talk of ***optionality***; if it is greater, of ***obligation***. The optionality may still divide into ***variation*** (for distances below, say, 7) and ***error***, though these subjective labels will generally be assigned with more criteria than rate of occurrence alone.

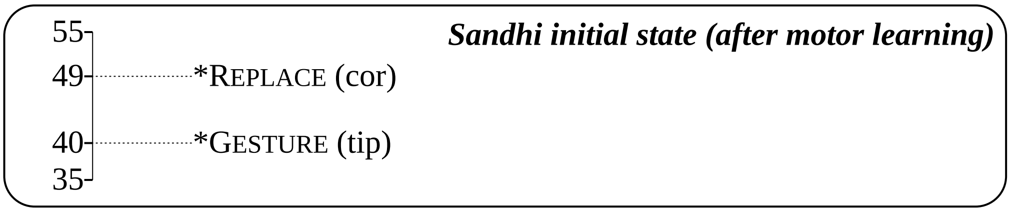
In chapter 14, I have shown that the continuous ranking scale allows a very simple and robust gradual learning algorithms, and that the current idea of optionality leads to a realistic learning curve. In this chapter, I will show that optionality in the production grammar can be learned and that the listener's categorization system automatically adapts to asymmetries in the distributions of variations in production. Finally, I present a shallow proof of the correctness of the "maximal gradual" algorithm for learning stochastic grammars.

15.2 Learning optionality in production

In this section, I will show that if adults exhibit place assimilation of nasals in 86% of all cases, like with grammar (15.5), then their children will copy this degree of optionality in their production grammars.

15.2.1 Learning that faithfulness can be violated in an adult grammar

At four years of age, Dutch children tend to pronounce |an+pa| faithfully as [anpa], though their parents probably say [ampa] most of the time. This is a natural stage in phonological development: the underlying form ends in |-an|, which the learner can easily deduct from the form as spoken in isolation. Because the child perceives her own form [anpa] as /anpa/, no faithfulness constraint is violated. In fact, earlier stages in learning have centred around acquiring all the gestures necessary for implementing the perceptual contrasts of the language, and the adult form, as perceived by the learner, has always been taken to be the underlying specification, with respect to which she evaluates the faithfulness constraints. Thus, the child's grammar is something like



(15.11)

The next step in phonological development is to learn that faithfulness constraints can be violated: the separation between perceived and underlying forms can begin. The learner will notice that the says /anpa/, but that adults sometimes say /ampa/. The discrepancy within this *learning pair* is shown in the following tableau (cf. 14.34):

[ampa] /ampa/ |an#pa|	*Replace (place: cor)	*Replace (place / nas)	*Replace (place / _C)	*Gesture (lip)	*Gesture (tip)
☞ [anpa] /anpa/				*	*
√ [ampa] /ampa/	*!	*	*	*	

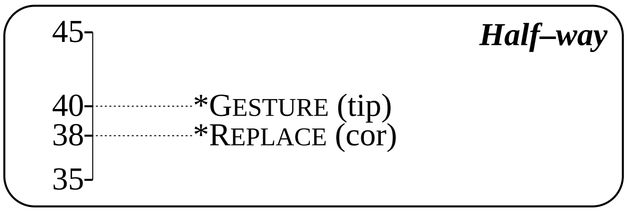
(15.12)

In this tableau, the top left shows the adult production [ampa] and the child's perception of it: /ampa/. Her own production is [anpa], which she perceives as /anpa/. This is the winner of the evaluation, as shown by the pointing finger (☞). However, the learner knows that /ampa/ should have been the winner, and she has already learned in an earlier stage that she can implement that by saying [ampa]. Therefore, the row with the check mark ("√") shows the correct, but losing candidate. Something will have to be done. The learner will take a *learning step.*

15.2.2 The minimal gradual learning step: demoting the offender

The offending incorrectly ranked constraint is the one with the crucial violation (the exclamation mark) in the evaluation of the correct candidate [ampa] (in the row with the check mark). This offending constraint is *Replace (cor). A simple strategy that will eventually prevent the mistake from reoccurring after a number of errors, is to **demote the offender**, i.e., to lower the ranking of *Replace (cor) by a small amount (e.g. a step of 0.01) along the continuous ranking scale. In chapter 14, I showed that with this strategy (the Minimal Gradual Learning Algorithm) any target constraint ranking can be learned within a reasonable time, independently of the initial rankings of the constraints.

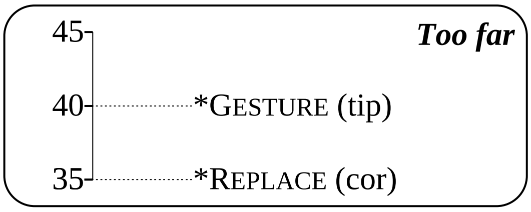
Demotion will proceed until *Replace (cor) ranks below *Gesture (tip). But suppose that at a certain time, the ranking is already as follows:



(15.13)

According to table (15.10), the probability that a subsequent learning pair will contain an adult model /ampa/ and a learner's utterance /anpa/, is still 86%·24% = 21%, and such a case will lead to a further demotion of *Replace (cor); the probability that the adult model is /anpa/ and the learner's utterance is /ampa/, is 14%·76% = 11%, and such a case would lead to demotion of *Gesture (tip). Thus, even now that faithfulness has fallen below the gestural constraint, there will still be more demotions of *Replace than of *Gesture, and this will raise the difference between the ranking values even further.

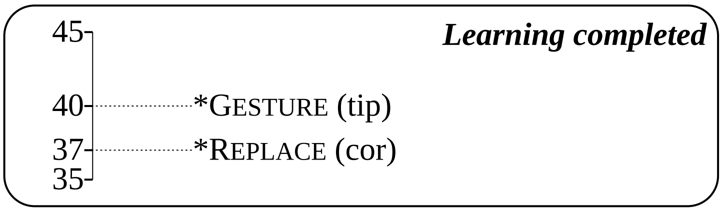
However, if the ranking difference becomes large, there will be more demotions of *GESTURE than of *REPLACE:



(15.14)

In this case, a demotion of *GESTURE will occur in only 86% · 3.9% = 3.3% of the cases, and a demotion of *REPLACE in 14% · 96.1% = 14% of the cases. The net result is that the two constraints will get closer.

A state of stable equilibrium will be reached when the ranking difference has become such that the demotion probabilities of *GESTURE and *REPLACE are equal, i.e., when they are 86%·14% and 14%·86%, respectively. This, of course, occurs when the ranking difference is 3, as in the adult grammar:



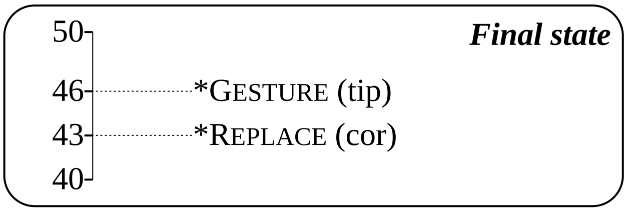
(15.15)

Thus, stochastically evaluating learners acquire not only the adult ranking order, but also the adult ranking differences and, therefore, the adult degree of optionality in production. In §15.3, we will see that for a demotion-only learner, this result is valid only if there are only two interacting constraints.

15.2.3 A remedy for downdrift: symmetric demotion and promotion

Optionality causes a problem for demotion-only learning. Considered as a whole, the grammar is not very stable, because the finite error probabilities that come with optionality cause the relevant constraint pair to keep on falling down the constraint hierarchy: in (15.14), learning may be completed but demotion of both constraints will continue (§14.2.12). In general, such a movement will push along any constraint that is crucially ranked lower than this pair in the target (adult) hierarchy, and it will drag down any constraint that is ranked higher and has an optionality relationship with one of the members of the pair. For instance, if place assimilation for plosives has a probability of 2%, the constraint *REPLACE (cor / plosive) will be dragged along at a distance of 6 above *GESTURE (tip) (in first approximation, but see §15.3.7).

Several stabilizing scenarios can be thought of, and one local scenario involves a symmetric combination of demotion of one of the members of the pair, and promotion of the other: when *REPLACE falls by 0.01, *GESTURE will rise by 0.01. More precisely, we should look at the evaluation of the incorrect winner (the row with the pointing finger) and find the highest violated constraint that is not violated by the correct (but losing) candidate. If our constraint set is correct, we know that such an uncancelled mark must exist in the winner, because the winner is obviously not the optimal candidate in the target (adult) grammar. In (15.12), this constraint is *GESTURE (tip). We now promote this constraint by a small step along the continuous ranking scale. With an original ranking as in (15.11), the two constraints will end up in the following grammar:



(15.16)

We see that the centre of the two constraint rankings is still at 44.5, as in the initial state (15.11). We are justified in calling (15.16) the "final state" because the rankings will stay in the vicinity of where they are in (15.16), without joining in a wholesale demotion race.

This combined demotion-promotion scheme seems to be as convergent and robust as the Mimimal GLA, though it is not as "minimal" and local: to implement it, we will have to consider one of the violation marks in the "grey cells" of the tableau (15.12).

In §15.3, we will see that the matching of the degree of optionality found in §15.2.2 for a single constraint pair, extends to larger sets of constraints only if the learner follows the combined demotion/promotion strategy described here.

15.2.4 Stochastic evaluation independently needed

We did not introduce random constraint evaluation with the intent of accounting for variation. Rather, this random evaluation is independently needed to guarantee a fundamental property of the natural language user's learning behaviour: robustness. If a minority of errors in the input is to have no dramatic consequences in our grammar, the learner must be allowed to adjust constraint rankings only by an amount that is much smaller than the difference between the rankings of relevant constraints. To prevent a modest number of errors from turning the grammar upside down, a ***safety margin*** (safe ranking difference) must be maintained. In an error-driven learning scheme, this can only be achieved by stochastic evaluation: only by making a few mistakes herself can the learner refresh a safety margin that has been shrunk by an error. Thus, optionality follows directly from the robustness requirement of learnability.

15.3 Learning a perception grammar

Consider perceptual categorization along a continuous auditory dimension with values from [0] to [100]. Suppose that a language has the three contrastive categories /30/, /50/, and /70/ along this dimension.

15.3.1 An OT grammar for perceptual categorization

In the listener's perception grammar, the relative fitness of the various categories, given an acoustic input value *x*, is described by a family of *WARP constraints for each category *y*:

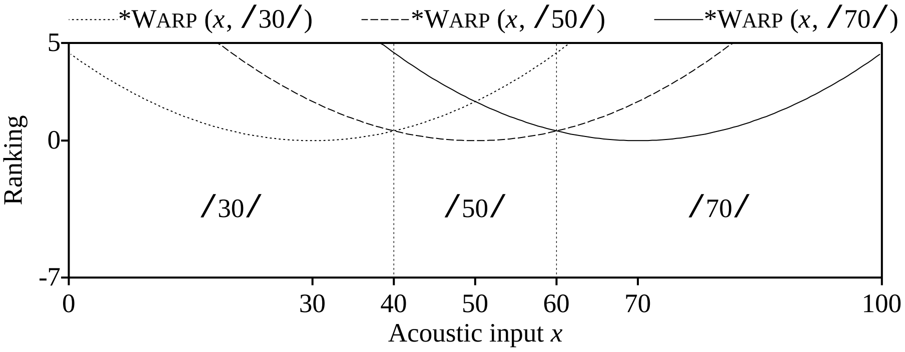
Def. *WARP (*f*: *x*, *y*)

"An acoustic value *x* on a perceptual tier *f* is not categorized into the category whose centre is at *y*." (15.17)

This formulation is slightly different from (8.3) because of its dependence on *y*, so that *WARP is now analogous to the *REPLACE family of the production grammar. Now, a less distorted recognition is preferred over a more distorted recognition, so the *WARP constraints are locally ranked according to

$$\text{*WARP } (\textit{feature}: x_1, y) \gg \text{*WARP } (\textit{feature}: x_2, y) \Leftrightarrow |y - x_1| > |y - x_2| \qquad (15.18)$$

provided that x_1 and x_2 are on the same side of the category centre *y*. The continuous *WARP families for our three categories could thus be depicted as



(15.19)

To see how these constraints interact in the listener's categorization system, consider what happens to the datum [44]. The listener has three candidate categories, and the perception grammar determines the winner:

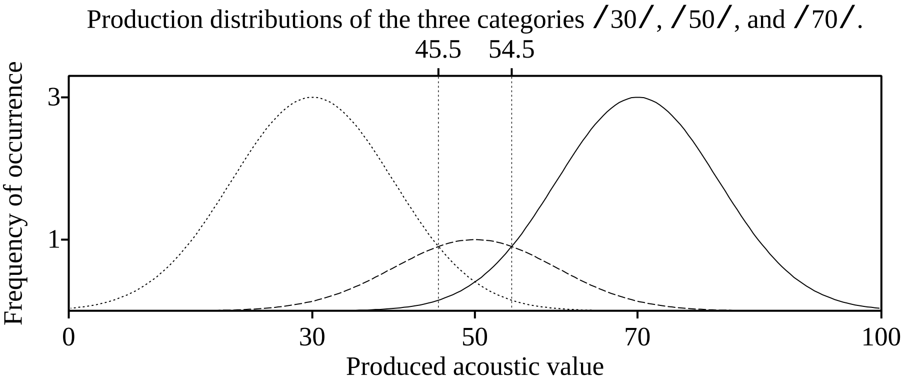
[44]	*WARP ([44], /70/)	*WARP ([44], /30/)	*WARP ([44], /50/)
/30/		*!	
☞ /50/			*
/70/	*!		

(15.20)

The ranking of the three relevant *WARP constraints can be read from the dotted line at [44] in figure (15.19): in going from the bottom up, it first cuts the *WARP (*x*, /50/) curve, then the *WARP (*x*, /30/) curve, and finally the *WARP (*x*, /70/) curve. Because the *WARP (*x*, /50/) curve is the lowest of these curves for *x* = 44, the listener categorizes the acoustic input into the /50/ class. Given the three equally shaped and equally high curves in (15.19), the discrimination criteria are obviously at [40] and [60], and if evaluation is not stochastic, these criteria are hard: every input above [60] is classified as /70/, every input below [40] as /30/, and every other input as /50/.

15.3.2 Production distributions and the optimal listener

Variations within and between speakers will lead to random distributions of the acoustic input to the listener's ear. Suppose that a language has three categories with midpoints at [30], [50], and [70] along a perceptual dimension, and a problematic three-way contrast: the middle category is weaker than the others (e.g. has fewer lexical occurrences). The speaker's productions, which are the inputs to the listener's perception grammar, are then distributed as follows:



(15.21)

The listener will make the fewest mistakes in initial categorization if she uses the criterion of *maximum likelihood*, i.e., if given the acoustic input *x* she chooses the perceptual category *y* that maximizes the a posteriori probability

$$P(prod = y \mid ac = x) = \frac{P(ac = x \mid prod = y) \cdot P(prod = y)}{P(ac = x)} \tag{15.22}$$

For instance, if the acoustic input is [44], an optimal listener will choose the /30/ category because the curve of the distribution of the production of /30/ in figure (15.21) is above the curve associated with the production of the category /50/, although the value [44] is nearer to the midpoint of the /50/ category than to the midpoint of /30/. Therefore, she will initially categorize all inputs below the criterion [45.5] into the class /30/, all the values between [45.5] and the second criterion [54.5] into the class /50/, and all values above [54.5] into the class /70/. I will now show how an OT listener establishes these criteria.

15.3.3 The initial state and its inadequacy

Figure (15.19) shows a possible initial state with unbiased categorization. Given the language environment, the listener will more often have to recognize a [44] input into the /30/ class than into the /50/ class, though she will prefer the /50/ class herself. Therefore, (15.19) is not an optimal grammar.

15.3.4 Learning from categorization errors

The categorization according to (15.19) is independent from what the speaker's intended category was, but if the listener gets to know (in the recognition phase, after lexical access etc.) which category the speaker had meant to produce, she may take a *learning step*. Suppose that the speaker had intended the /30/ category. Tableau (15.20) can then be enriched in a way analogous to (but somewhat simpler than) the learning tableau for production grammars (15.12):

/30/ [44]	*WARP ([44], /70/)	*WARP ([44], /30/)	*WARP ([44], /50/)
√ /30/		*!	
☞ /50/			*
/70/	*!		

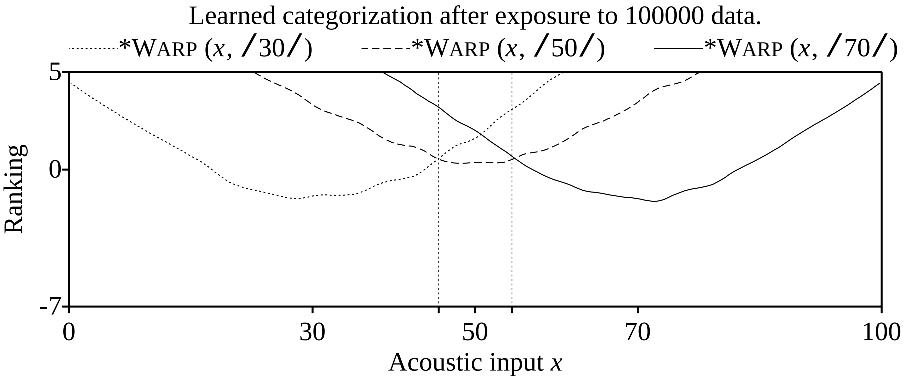
(15.23)

The listener now "knows" that she has made a categorization error. The offending constraint is the one with the crucial violation (the exclamation mark) in the evaluation of the intended category /30/ (in the row with the check mark). This offending constraint is *WARP ([44], /30/). A simple learning strategy (§15.2.2) is to demote the offender, i.e., to lower the ranking of *WARP ([44], /30/) by a small amount (say 0.01) along the continuous ranking scale, which runs from –7 to 5 in figure (15.19).[2]

15.3.5 Stochastic categorization and the optimal criterion

A crucial ingredient for the model is the stochastic constraint evaluation of §15.1: the ranking of each categorization constraint at evaluation time is drawn from a Gaussian distribution about its ranking in figure (15.19), again with a spreading of 2. This means that an acoustic input of [44] has a chance of being initially categorized as /30/, /50/, or even /70/, with probabilities that depend on the differences between the heights of the three *WARP ([44], *y*) curves. Even after *WARP ([44], /30/) has fallen below *WARP ([44], /50/), there is still a chance that a [44] datum will be initially perceived as /50/. This optionality will lead to safety margins between the curves: *WARP ([44], /30/) will be demoted below *WARP ([44], /50/) until the error probabilities, given the production distributions and the categorization noise, are the same for both classes. After exposure to 100,000 data, the perception grammar of a demotion/promotion learner will look like

[2] Because the constraint family is continuous, I used a Gaussian *demotion window* in the simulations, i.e., the nearest neigbours (say, [39] through [49]) were also demoted, according to a Gaussian window with a spreading of 1.58 acoustic units.



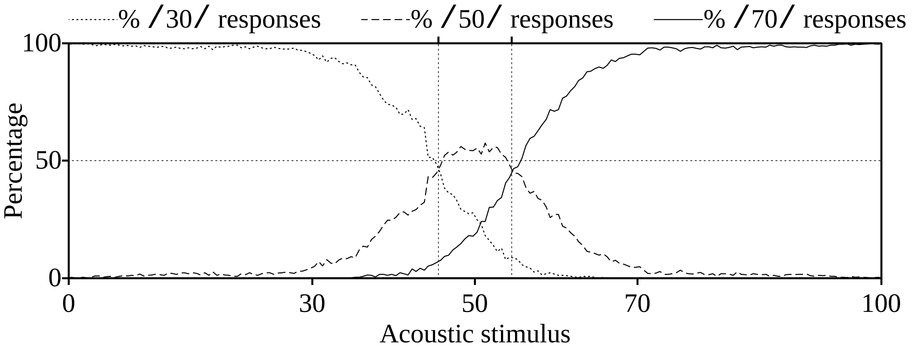
(15.24)

In the simulation that led from the initial unbiased grammar (15.19) to the adult grammar (15.24), the perceptual range was divided into 200 steps of 0.5, the error-driven demotion step (*plasticity*) was 0.01 (also stochastic, with a relative spreading of 0.1), and the categorization spreading was 2, and the local-ranking principle was not enforced[3]. We see that the minimal gradual learning algorithm causes the two criteria between the middle category and its neighbours (the cutting points in the figure) to shift in the direction of the middle category, until they fall together with the optimal criteria identified in §15.3.2. Thus, **the minimal gradual OT learner will automatically learn to set the criteria in a way that a maximum-likelihood listener would**. Note how the local learning strategy of demoting a single incorrectly invalidating constraint implements the global functional principle of maximizing ***the ease of comprehension***, i.e. minimizing the number of initial perception errors, thus minimizing the number of cases that the initial categorization will have to be repaired by the “higher” parts of the recognition system.

15.3.6 Probability matching

But our learner does not become a perfect maximum-likelihood listener. This is because the learned criteria are ‘soft’: because of the stochastic categorization, there will be regions in the acoustic space where more than one category can be initially perceived: even though the acoustic input [44] is most likely to come from an intended /30/ production, there is still a large probability that it is initially perceived as /50/. From grammar (24), we can determine the perception-probability curves for the three categories, by the following simulation. We present 1000 acoustic replications of each of the 200 acoustic stimuli 0.25, 0.75, 1.25, ..., 99.75 to the (simulated, patient) listener who is defined by the grammar (24). We will ask her what she hears and force her to choose from the categories /30/, /50/, and /70/; we will assume that her grammar is fixated, i.e. that she will not adapt her criteria to the uniform distribution of the stimuli (only computerized listeners can be frozen in this way). The 200,000 stimuli gave the following three curves for the percentages of the responses of each of the three categories, as functions of the controlled acoustic stimulus:

[3] The Praat script that performs these simulations and produces the figures (15.19), (15.21), (15.24), (15.25), (15.29), (15.34), (15.35), and (15.36), is available from http://fonsg3.hum.uva.nl/paul/.



(15.25)

These curves are very similar to the categorization curves for controlled acoustic stimuli, as have been measured for several ternary categorizations: voice-onset time (the [b]-[p]-[p^h] continuum) in Thai (Lisker & Abramson 1967); vowel height (the [ɪ]-[ɛ]-[æ] continuum) in English (Fry, Abramson, Eimas & Liberman 1962); and place "of articulation" (the perceptual [b]-[d]-[g] continuum) in English (Liberman, Harris, Hoffman & Griffith 1957).

So, the listener does not maintain an accurate maximum-likelihood strategy. We can compute the categorization probabilities from the production probabilities, if we realize that in an equilibrium situation, the demotion frequencies of the two competing categories will be equal. For instance, the acoustic input [40] represents an intended /30/ category in 74% of all cases, and the /50/ category in 25% of all cases. Equilibrium has been achieved (for a demotion/promotion learner, who shows no "downdrift") when the probability of the error of classifying an intended /30/, realized as [40], into the /50/ category, is equal to the probability of the error of classifying an intended /50/, also realized as [40], into the /30/ category:

$$P(prod = 30 \wedge perc = 50 / ac = 40) = P(prod = 50 \wedge perc = 30 / ac = 40) \qquad (15.26)$$

Under the assumption that the initially perceived category does not depend on the speaker's intended category, but only on the acoustic input, we can rewrite the combined probabilities as

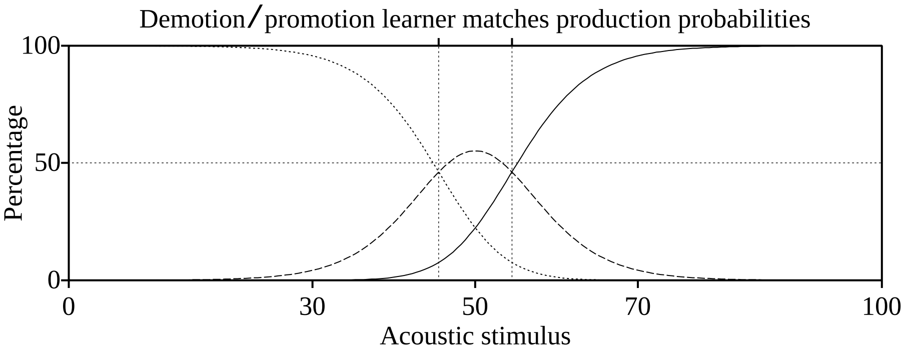
$$\begin{aligned} P(prod = 30 \wedge perc = 50 / ac = 40) &= P(prod = 30 / ac = 40) \cdot P(perc = 50 / ac = 40) \\ P(prod = 50 \wedge perc = 30 / ac = 40) &= P(prod = 50 / ac = 40) \cdot P(perc = 30 / ac = 40) \end{aligned}$$

(15.27)

Combining (15.26) and (15.27), we get

$$\frac{P(perc = 30 / ac = 40)}{P(perc = 50 / ac = 40)} = \frac{P(prod = 30 / ac = 40)}{P(prod = 50 / ac = 40)} \qquad (15.28)$$

Thus, our learner becomes a ***probability-matching listener***: her perception bias is going to equal the production bias: she will categorize the input [40] into the /30/ class in 74% of all cases, and into the /50/ class in 24% of the cases. We may note the similarity between (15.25) and a graph of the posterior production probabilities given any acoustic input, which can be derived easily by dividing the three values for each acoustic value in (15.21) by the sum of these three values:



(15.29)

The probability-matching strategy automatically results from OT learning with stochastic evaluation, **no matter how weak the random part of it is, as long as it is greater than the plasticity**.

Note that this strategy does not minimize the global number of perception errors, though it may aid in the recovery from initial errors if the acoustic signal is still in short-term memory.

15.3.7 Poor categorization performance of a demotion-only learner

The results of §15.3.6 are valid for demotion-only learners in learning a single constraint pair, and for combined demotion/promotion learners in general. We will now see how a demotion-only learner would mess up the *three* constraint families that are relevant for our categorization problem.

For a given acoustic input, say [40], an equilibrium is reached when all three *WARP constraints are demoted equally often, i.e., when the listener makes an equal amount of "mistakes" in classifying an intended /30/, /50/, or /70/ production. Thus, suppressing the condition clause, (26) expands to

$$\begin{aligned} &P(prod = 30 \wedge perc = 50) + P(prod = 30 \wedge perc = 70) = \\ &P(prod = 50 \wedge perc = 30) + P(prod = 50 \wedge perc = 70) = \\ &P(prod = 70 \wedge perc = 30) + P(prod = 70 \wedge perc = 50) \end{aligned} \qquad (15.30)$$

Again under the assumption of independent categorization, this becomes

$$\begin{aligned} &P(prod = 30) \cdot \big(P(perc = 50) + P(perc = 70)\big) = \\ &P(prod = 50) \cdot \big(P(perc = 30) + P(perc = 70)\big) = \\ &P(prod = 70) \cdot \big(P(perc = 30) + P(perc = 50)\big) \end{aligned} \qquad (15.31)$$

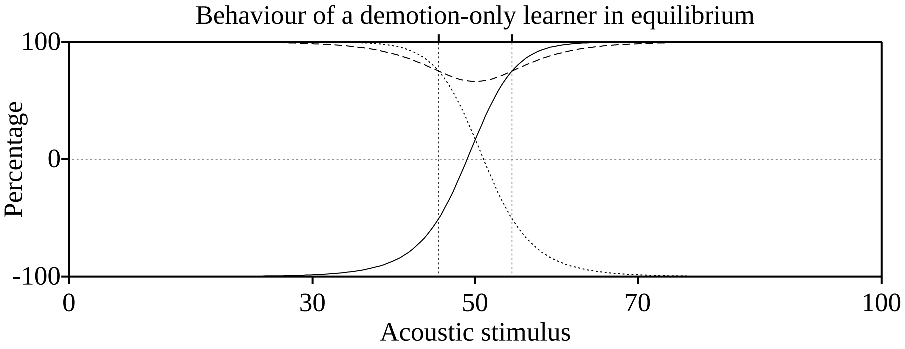
Remembering that

$$P(perc = 30) + P(perc = 50) + P(perc = 70) = 1 \qquad (15.32)$$

we can compute the three unknown perception probabilities by solving the three linear equations (31) and (32). Instead of the probability-matching formula (28), we get (with a notation adapted to the width of the page):

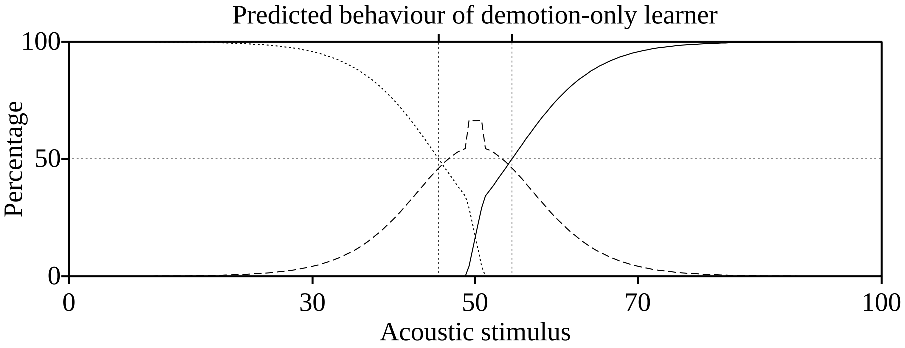
$$P(perc = 30) = 1 - 2\frac{P_{prod}(50)\cdot P_{prod}(70)}{P_{prod}(30)\cdot P_{prod}(50) + P_{prod}(30)\cdot P_{prod}(70) + P_{prod}(50)\cdot P_{prod}(70)} \tag{15.33}$$

This predicts the following categorization probabilities for each acoustic input:



(15.34)

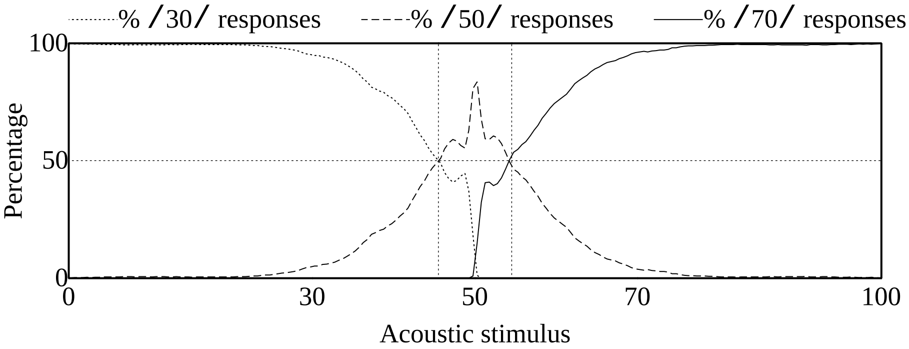
The situation is clearly pathological: we see negative probabilities except in a small range of acoustic values around [50]. This just means that outside this domain there is no concerted downdrift of the three constraints: at [60], for instance, *WARP (x, /50/) and *WARP (x, /70/) will be drifting down the ranking scale, but *WARP (x, /30/) will be left behind, driving the probability that the listener classifies an acoustic input [60] as /30/ to zero. In the limit, therefore, the listener's perception will seem to follow a two-constraint probability-matching strategy outside the small acoustic domain in the centre:



(15.35)

A simulated demotion-only learner confirmed this when asked to classify the whole acoustic range after a million learning data[4]:

[4] The small differences between (15.35) and (15.36) arise from using the Gaussian demotion window (fn. 2).



(15.36)

To my knowledge, the discontinuities and exact zeroes exhibited by (15.35) and (15.36) have not been found in categorization experiments. To the extent that the response distributions (15.25) and (15.29) are more realistic, we must conclude that a symmetric demotion/promotion learning model better represents reality than a demotion-only model. This, added to the solution of the grammatical downdrift problem, leads us into questioning the validity of demotion-only learning schemes, be they gradual (chapter 14) or not (Tesar & Smolensky 1993, 1996).

15.4 The correct maximal algorithm for learning a stochastic grammar[5]

Contrary to what §15.3.6 suggested, the symmetric version of the ***minimal*** gradual learning algorithm does not lead to probability matching. Instead, the correct algorithm must demote *all* violated constraints in the adult's utterance, and promote *all* violated constraints in the learner's utterance. In the example of (15.23), there would be no difference between this ***maximal*** algorithm and the minimal algorithm, but in a grammar with a larger number of constraints, there would.

Suppose that there are K candidates, each of which has a probability P_k^{L} (k = 1...K) of being chosen by the learner, and a probability of P_k^{A} of being chosen by the adult. Suppose that the grammar contains N constraints with rankings r_n (n = 1...N). As a result of the demotion of all the adult's violated constraints, the ranking of constraint n will increase upon the next learning pair by a negative amount Δr_n, whose expectation value is

$$\mathrm{E}^{\mathrm{A}}\left[\Delta r_n\right] = -p \cdot \sum_{k=1}^{K} P_k^{\mathrm{A}} m_{kn} \qquad (15.37)$$

where p is the plasticity constant, and m_{kn} is 1 if candidate k violates constraint n and 0 otherwise (for now, we consider only constraints that can be violated only once). Likewise, the promotion of all the learner's violated constraints will lead to an expected positive ranking increase of

[5] This section did not occur in the ROA version of this chapter. The maximal algorithm evolved after a computer simulation of the learning of the extensive optionality data from Hayes & MacEachern (to appear).

$$\mathrm{E}^{\mathrm{L}}\left[\Delta r_n\right] = +p \cdot \sum_{k=1}^{K} P_k^{\mathrm{L}} m_{kn} \tag{15.38}$$

The total expected change in the ranking is

$$\mathrm{E}\left[\Delta r_n\right] = p \cdot \sum_{k=1}^{K} \left(P_k^{\mathrm{L}} - P_k^{\mathrm{A}}\right) m_{kn} \tag{15.39}$$

We can see that if a candidate occurs with greater probability in the speaker than in the adult, its violated constraints will rise on average, so that the probability of this candidate in the speaker will decrease. Thus, the expected ranking change seems to decrease the gap between the two grammars. Now, we will have to find a more formal proof.

We can see immediately that if the learner's grammar equals the adult's grammar, i.e. if P_k^{L} equals P_k^{A} for all *k*, the expected ranking change of every constraint *n* is zero, i.e. the expected change in the learner's grammar is zero. To prove learnability, however, we have to show the reverse, namely the convergence of the learner's grammar upon the adult's grammar. An important part of the proof involves showing that the learner cannot end up in a different grammar from the adult. Suppose the learner does end up in such a *local maximum*, i.e. $\mathrm{E}\left[\Delta r_n\right]$ is zero for every constraint *n*. We can write this situation in vector-matrix notation:

$$\mathbf{m}^T\left(\mathbf{P}^{\mathrm{L}} - \mathbf{P}^{\mathrm{A}}\right) = 0 \tag{15.40}$$

Given a violation matrix **m**, the learner can end up in any grammar $\mathbf{P}^{\mathrm{L}}$ that satisfies (15.40). As we know from linear algebra, however, the vector $\mathbf{P}^{\mathrm{L}} - \mathbf{P}^{\mathrm{A}}$ must be zero if the matrix **m** behaves well. We can distinguish the following cases of ill-behaved violation matrices:

1. There are two candidates *k* and *l* who violate the same set of constraints. Equation (15.40) is then valid for any **P** for which there is an *a* so that $P_k^{\mathrm{L}} = P_k^{\mathrm{A}} - a$ and $P_l^{\mathrm{L}} = P_l^{\mathrm{A}} + a$. However, under our evaluation regime, these candidates are equally harmonic in every respect, so they must have equal probabilities in the learner's grammar $\left(P_k^{\mathrm{L}} = P_l^{\mathrm{L}}\right)$ as well as in the adult's grammar $\left(P_k^{\mathrm{A}} = P_l^{\mathrm{A}}\right)$. Combining the four equations, we see that *a* must be zero.
2. There is a candidate *k* that violates all constraints violated by candidate *l* as well as those violated by candidate *m*. Equation (15.40) is then valid for any **P** for which there is an *a* so that $P_k^{\mathrm{L}} = P_k^{\mathrm{A}} - a$, $P_l^{\mathrm{L}} = P_l^{\mathrm{A}} + a$, and $P_m^{\mathrm{L}} = P_m^{\mathrm{A}} + a$. However, if candidate *k* violates a proper superset of the constraints violated by another candidate, it should always be judged less harmonic than that other candidate in the pairwise evaluation, regardless of the constraint ranking. Therefore, $P_k^{\mathrm{L}} = P_k^{\mathrm{A}} = 0$, so that *a* must be zero.
3. Candidate *k* violates constraints A and C, *l* violates B and D, *m* violates A and D, and *n* violates B and C. Equation (15.40) is then valid for any **P** for which there is an *a* so that $P_k^{\mathrm{L}} = P_k^{\mathrm{A}} - a$, $P_l^{\mathrm{L}} = P_l^{\mathrm{A}} - a$, $P_m^{\mathrm{L}} = P_m^{\mathrm{A}} + a$, and $P_n^{\mathrm{L}} = P_n^{\mathrm{A}} + a$. This is a genuine case of degeneracy: the constant *a* will be adjusted so that $P_k^{\mathrm{L}} / P_m^{\mathrm{L}}$ ends up near $P_n^{\mathrm{L}} / P_l^{\mathrm{L}}$, **irrespectively of the initial constraint rankings**; for instance, if the adult

has $P_k^{\mathrm{A}} = 0.1$, $P_l^{\mathrm{A}} = 0.2$, $P_m^{\mathrm{A}} = 0.27$, and $P_n^{\mathrm{A}} = 0.43$, the learner will arrive near $P_k^{\mathrm{L}} = 0.2$, $P_l^{\mathrm{L}} = 0.3$, $P_m^{\mathrm{L}} = 0.17$, and $P_n^{\mathrm{L}} = 0.33$, and she will never reach the adult distribution. But! This adult distribution could never have been derived from a stochastically evaluating OT grammar: there is no constraint ranking that produces it. In fact, with the given candidates and violations, any grammar must satisfy the condition that if $P_k < P_m$ (i.e., C dominates D), then also $P_n < P_l$. This is one of the empirical predictions of our hypothesis of stochastic evaluation: some distributions are impossible.

4. Any more complicated dependencies between the violations of the candidates. Generally, if there are many more candidates than constraints (which is true under most interpretations of the candidate generator in OT), and if these candidates cover the range of possible sets of violations, (15.40) must lead to the conclusion that $\mathbf{P}^{\mathrm{L}} = \mathbf{P}^{\mathrm{A}}$, i.e. that the algorithm converges upon the adult grammar.

We have made plausible, though not yet rigorously proved, that the maximal symmetrized gradual learning algorithm is capable of learning any stochastically evaluating OT grammar.

15.5 Conclusion

Optionality follows directly from the robustness requirement of learnability: a demotion/promotion learner will show the same error rate herself as she hears in her environment. To be resistant against 5% errors, you must make 5% errors yourself; 30% variation in your environment will make you produce 30% variation yourself; and if a certain acoustic input has a 30% probability of stemming from an intended category *x*, your perception grammar will make you classify this acoustic input into the category *x* 30% of the times.

These results are exact only for a symmetrized and maximal version of the Gradual Learning Algorithm, i.e., a version in which the learning step involves demotion of all constraints with uncancelled marks in the correct (but losing) candidate, and simultaneous promotion of all constraints with uncancelled marks in the incorrectly winning candidate. There is some evidence that this combined demotion/promotion learning scheme is a better model of learning than the demotion-only scheme: apart from the grammar-internal downdrift problem, the observable quantities of categorization show unrealistic behaviour with the demotion-only scheme.

The account of optionality presented here naturally encapsulates pragmatics-based reranking. For instance, if you want to speak more clearly, you may raise all your faithfulness constraints by, say, 5 along the continuous ranking scale. In this way, an 80%-20% preference *for* place assimilation will turn into a 18%-82% preference *against*. Depending on whether the faithfulness constraint is ranked above or below its rival, slight variation may turn into obligations or the reverse. If the ranking difference is large to begin with, however, nothing happens; so we see that discrete properties of surface rerankability are compatible with, and may well follow from, a general continuous rerankability of all constraints.

Our account of optionality may well extend to other parts of the grammar, including the problem of constituent ordering in syntactical theory, which is a field where optionality is very common. Our account may well explain how the “interacting and possibly competing principles and preferences” of Functional Grammar (Dik 1989: 337) determine the choice between, say, surface SVO and OVS orders in a V2 language: one part of the answer will be pragmatical reranking of the relevant functional principles (like “subject first”, “human first”), and another part will be the random variation that occurs at evaluation time; in an obligatory SVO language, one of the constraints is just ranked so far above the other that the degree of variation is essentially zero.

16 Inventories[1]

Abstract. We can describe as well as explain the *symmetries* as well as the *gaps* in inventories of vowels and consonants. Symmetries are the language-specific results of general human limitations on the acquisition of perceptual categorization and motor skills. Gaps are the results of local hierarchies of articulatory effort, perceptual contrast, and perceptual confusion. There is no need to posit a dedicated inventory grammar: inventories are the automatic result of the constraints and their rankings in the production grammar.

Consider the short-vowel system of Frisian:

i y u

e ø o

ɛ ɔ

a (16.1)

Inventory (16.1) shows two common properties of inventories:

(a) **Symmetry**. The eight non-low vowels occur in only three heights and three “places”; they are not scattered randomly throughout the space of possible vowels. In §16.1.1, I will show that this symmetry is real. ‘Phonetic’ approaches to sound systems like (16.1) have not taken into account the symmetrizing principles of perceptual categorization and motor learning, though these are general phenomena of human behaviour, not specific to phonology.

(b) **Gaps**. The lower-mid-short-vowel system has a *gap* at /œ/. This has must have something to do with the fact, stressed by the trapezoidal shape of (16.1), that the perceptual front-back contrast is smaller for lower vowels than for higher vowels. ‘Phonological’ approaches have ignored the explanatory power of the communicatively functional principles that segments will tend to be well contrasting and easy to articulate, although these principles can now easily be expressed as near-universal rankings in a constraint grammar

In Functional Phonology, the effects of perceptual categorization, motor learning, perceptual contrast, and articulatory effort are expressed directly in the grammar. In the current chapter, I will show that this approach is capable of representing the symmetries as well as the gaps of segment inventories.

16.1 Phonological and phonetic approaches to inventories

Since this chapter is intended for a mixed audience of phonologists and phoneticians, it seems worthwhile to discuss the approaches to the modelling of inventories that have

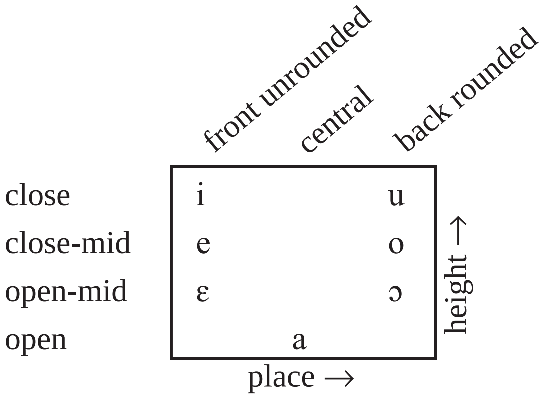
[1] This chapter appeared as Boersma (1997c).

been proposed from each side, and to point out the strengths and weaknesses of both with respect to explanatory power and empirical adequacy. In later sections, I will show that the functional phonological account of inventories combines the strengths of the two sides, without, I hope, copying any of the weaknesses.

As an example, we will make a comparison of how the two approaches have answered the question: “if a language has seven vowels (without differences in length, tone, or phonation), what will they be?”

16.1.1 The reality of symmetry

The following seven-vowel system is very common (Crothers 1978, Maddieson 1984):



(16.2)

The first thing that meets the eye of the phonologist, is the ***symmetry*** within this inventory. There are three front unrounded vowels, two close vowels, and so on. Groups like these are called ***natural classes***, and they are the most likely candidates to co-occur in synchronic ***phonological rules*** or historical ***sound changes***.

Phonological rules. In northern standard Italian, which has exactly the portrayed vowel structure in stressed syllables (ignoring the diphthongs, and admitting that [a] is rather front), the open-mid vowels merge with the close-mid vowels when these syllables become unstressed as a result of a morphological operation. So /spˈɛndo/ ‘I spend’ next to /vˈendo/ ‘I sell’, but /spendiˈamo/ ‘we spend’ next to /vendiˈamo/ ‘we sell’; likewise /pˈɔrgo/ ‘I present’ and /sˈorgo/ ‘I arise’, but /pordʒˈamo/ and /sordʒˈamo/. The important thing here is that the back vowel [ɔ] behaves exactly like the front vowel [ɛ], which is the rationale behind suggesting that [ɛ] and [ɔ] constitute a natural class, and the reason why we can talk generalizingly about “open-mid” vowels at all.

Sound change. In some parts of Italy (let’s call them area A), /ɛ/ does *not* contrast with /e/ (there is only one mid front vowel). And in area B, /ɔ/ does not contrast with /o/. Now, the areas A and B *are the same*. This means that front and back vowels must have had a common property at the time when the late sound change (merger of lower and higher mid vowels) occurred.

A more striking example of this front/back concerto is found with the sound changes that converted the Latin vowel system into system (16.2) when the Latin length correlations were lost. Latin long /iː/ and /uː/ became Italian /i/ and /u/ (viːnum → vino ‘wine’, luːna → luna ‘moon’), while Latin short /i/ and /u/ were both lowered to

Italian /e/ and /o/ (fidem → fede ‘belief’, supraː → sopra ‘above’). Latin long /eː/ and /oː/ became Italian /e/ and /o/ (feːtʃiː → fetʃi ‘I did’, doːnum → dono ‘gift’)[2], whereas both short /e/ and /o/ were lowered (and often diphthongized) to /(i)ɛ/ and /(u)ɔ/ (pedem → piɛde ‘foot’, rotam → ruɔta ‘wheel’). In all these cases, the height contour of a front vowel was changed in the same way as that of its corresponding back vowel; the symmetry was preserved[3].

Phonetics. There is nothing mysterious about the common behaviour of vowels at the same height. The symmetry suggests that there must be something similar in vowels of the same height. As this cannot be found in the muscles used for these sounds (genioglossus, lower longitudinals, and risorius for unrounded front vowels; styloglossus and orbicularis oris for rounded back vowels), it must be a ***perceptual*** similarity. As argued by e.g. Lindau (1975), this perceptual similarity between vowels at equal height is the ***first formant*** (F_1), the location of the peak in the excitation of the basilar membrane in the human inner ear furthest from the oval window. So, vowels at the same height have equal values for the first formant. Leoni, Cutugno & Savy (1995) measured the acoustic F_1 values (acoustically, formants are measured as peaks in the frequency spectrum) of the Italian seven-stressed-vowel system, and of the five-unstressed-vowel system. /ɛ/ and /ɔ/ have the same F_1, so have /e/ and /o/, and so have /i/ and /u/[4]. Thus, the changes from Latin to Italian can be described as changes in the ***F_1 contours*** of the vowels or diphthongs.

The degree of symmetry seems to depend on inventory size. While most four-vowel systems (Maddieson 1984) are asymmetric (/a ɛ i u/ instead of /ɛ ɔ i u/), large systems like the 18 long vowels and diphthongs of Geleen Limburgian are very symmetric:

i	y	u	iː	yː	uː		
			iæ	yɶ/øɶ	oɐ		
e	ø	o	eː	øː	oː		
			ɛi	œy	ɔu	↓ F_1	
ɛ̞	œ̞	ɔ̞	ɛː	œː	ɔː		
æ		ɑ	æi		ɑu		
				aː			(16.3)

First note that (apart from a possible emerging split in the opening diphthongs), the 18 long vowels are distributed over only seven distinct F_1 contours.

[2] Any length in the Italian reflexes is related to Italian stress, not directly to length in Latin.

[3] If the account of the Latin vowel system can be questioned on the ground that we do not know for sure that, say, /iː/ and /i/ had the same quality, the historical relations can be replaced by a comparison of Italian with Sardinian which simply merged Latin /iː/ with /i/, /uː/ with /u/, /eː/ with /e/, and /oː/ with /o/.

[4] Apart from an apparently cross-linguistically fixed and unexplained 30-Hz difference between front and back vowels, with back vowels having the slightly higher F_1.

Distribution. Sounds of equal length and height form natural classes in Limburgian phonemic distributions:

- [ɛi œy ɔu], [æi ɑu], and [i y u] do not occur before /ʀ/ in the same morph.
- [iæ yɶ/øɶ oɐ] always carry the acute accent.
- Of the short vowels, only [i u] can occur at the end of a word.

The point, again, is that the front unrounded, front rounded, and back rounded vowels act in the same way.

Phonological rules. Sounds of equal length and height also form natural classes in a pervasive phonological rule:

- The ***umlaut*** rule, which is used in the formation of diminutives and in the formation of many plurals, makes the following vowel changes: /u/ → /y/, /o/ → /ø/, /ɔ̝/ → /œ̝/, /ɑ/ → /æ/, /uː/ → /yː/, /oɐ/ → /øɶ/, /oː/ → /øː/, /ɔu/ → /œy/, and /ɔː/ → /œː/ (also /aː/ → /ɛː/). So, this is an alternation that uses very different tongue-body movements, but keeps vowel height intact.

Sound change. More proof of the organizational power of vowel height can be seen in sound changes and regional variation. The following examples take the Geleen dialect, which has a very conservative vowel system, as a reference:

- In the Sittard dialect (Dols 1944 [1953]), underlying acute /éː ǿː óː/ (/béːʀ/ 'beer', /zǿːkə/ 'search', /ɣóːt/ 'good') became /ɛ́i œ́y ɔ́u/ (/bɛ́iəʀ/, /zœ́ykə/, /ɣɔ́ut/) and [iæ yɶ/øɶ ʊɐ] (/kiæs/ 'cheese', /ɦyɶʀə/ 'hear', /ɣʀoɐt/ 'great') became [éː ǿː óː] (/kéːs/, /ɦǿːʀə/, /ɣʀóːt/).
- In the Roermond dialect (Kats 1985), [iæ yɶ/øɶ ʊɐ] merges with and into [éː ǿː óː], so that /ɣóːt/ 'good' rhymes with /ɣʀóːt/ 'great'.
- In the Venlo dialect (Peeters 1951), which does not contrast [æ] with [ɛ], [ɛ̝ œ̝ ɔ̝] are much lower ([ɛ œ ɔ]), [ɛː œː ɔː] are much higher ([ɛ̝ː œ̝ː ɔ̝ː]), and [iæ yɶ/øɶ oɐ] are [iə yə uə] (with accent contrasts).
- In the Maastricht dialect (Tans 1938), [ɛː œː ɔː] became [eː øː oː], surface acute [íː ýː úː] became [ɛ́i œ́y ɔ́u], and [iæ yɶ/øɶ oɐ] became [i y u] (/kis/ 'cheese', /py/ 'paws', /ɣʀut/ 'great').

In all these cases, the three vowel places act in the same way.

The examples discussed above can be multiplied at will for all kinds of languages and features. The symmetry is real, phonetically as well as phonologically. If a front vowel changes its height contour, the corresponding back vowel, if it exists, follows suit in the far majority of cases.

We are in search of a theory that both accounts for symmetry and explains it, i.e. a theory that has both descriptive adequacy and explanatory power.

16.1.2 The phonological approach to symmetry in rules

To account for symmetry, phonologists take the solution of describing each segment as a bundle of ***features***, preferably in such a way that they can describe *both* the inventory *and* the phonological processes in terms of these same features. These features traditionally take on no more than two values, so let us first see how these ***binary*** features describe the Italian mid-vowel merger rule.

First, vowel height, with its four possible phonetic values in Italian, has to be split up into at least two binary features. Several ways of doing this have been proposed, but let us work here with the labels in the left-hand side of the figure. For instance, we may use the features [open] and [mid], with the values [+open], [–open], [+mid], and [–mid]. As ***place*** features, we could use [back] and [round]. So, /e/ would be [–open, +mid, –back, –round]. The mid-vowel merger rule can now be stated as a ***feature-changing rule***:

$$\begin{bmatrix}+\text{open}\\+\text{mid}\end{bmatrix}\rightarrow\begin{bmatrix}-\text{open}\\+\text{mid}\end{bmatrix}\Big/\,[-\text{stress}] \tag{16.4}$$

which says that if an unstressed segment is [+open] and [+mid], it is *changed* to [–open] and [+mid], without changing any other feature. We could OT-ize this with an equivalent constraint ranking, with faithfulness constraints for each relevant underlying feature value and an ad-hoc phenomenological structural output constraint against open-mid vowels in unstressed position ("*" means "should not occur in the output"):

Italian mid-vowel merger

PARSE (+mid)
⋮
*[+open & +mid / –stress]
⋮
PARSE (+open)

(16.5)

This would correctly account for the surface forms of underlying open-mid vowels:

Input: /ɛ/ unstressed	PARSE (+mid)	*[+open & +mid / –stress]	PARSE (+open)
[ɛ]		*!	
☞ [e]			*
[a]	*!		

(16.6)

The output candidate [ɛ] honours all faithfulness constraints, but violates the constraint against open-mid vowels. In the winning candidate [e], the structural constraint is satisfied at the cost of violating the lower-ranked PARSE constraint that calls for the surfacing of the underlying [+open] specification. The structural constraint could also be satisfied by unparsing the [+mid] specification, giving as our third candidate the open non-mid vowel [a], but that would violate the higher-ranked PARSE constraint that says that underlying mid vowels should surface as mid vowels.

The important asset from formulation (16.4) or (16.5) is that it merges the two rules /ɛ/ → [e] and /ɔ/ → [o] into one; it accomplishes this by generalizing over all values for the features [back] and [round].

We can now be more precise about what defines a natural class in a phonological rule. The Italian mid-vowel example gives us five natural classes, defined in various ways:

- The group of segments that undergo the rule must have something in common; this is the ***structural description*** of the rule: the left-hand side plus the environment clause (unstressed) in (16.4), or that what is forbidden by the structural constraint in (16.5): the class of open-mid vowels, consisting of /ɛ/ and /ɔ/.
- The group of segments that are the result of applying the rule must have something in common; this is the right-hand side in (16.4): the class of close-mid vowels: /e/ and /o/.
- The segment that undergoes the rule must have something in common with the result of the rule, namely, the features not mentioned in (16.4) or (16.5). These common properties also define a natural class. So /ɛ/ and /e/ belong to the class of front unrounded vowels. A natural class found in this way may contain more segments than just these two (in this case, /i/).
- Likewise, /ɔ/ and /o/ belong to the class of back rounded vowels, with /u/.
- Combining the arguments, the four segments /e/, /ɛ/, /o/, and /ɔ/ together must form a subset of yet another natural class, the mid vowels. The rule involves ***neutralization*** of the class [+mid], as shown by the position of this class at the top of the sandwich in (16.5) and by the fact that the ***merger*** of /ɛ/ and /e/ into [e] allows us to rewrite (16.4) as $[+\text{mid}] \rightarrow [-\text{open}]$. Rule (16.4) is said to apply ***vacuously*** to /e/ and /o/: though these segments meet the structural description of the rule, they are not changed by it; in (16.5), this is reflected by the satisfaction of faithfulness for these underlying segments.

It should be noted that although the proposed account adequately describes the generalizations in the process of mid-vowel merger, it provides no explanation. This is quite apparent in the explicitly arbitrary formulation (16.4), but the OT version (16.5) fares no better: while the two PARSE constraints might be seen as natural (but where do the features come from?), the formulation of the structural constraint hides any relations that could explain the behaviour of open mid vowels, such as their relations with their neighbours, or the reason for the dependence of its ranking on stress: though we could imagine a not very active constraint *[+open & +mid / +stress] ranked below PARSE (+open), there is no explanation for why it should rank below *[+open & +mid / –stress].

16.1.3 The phonological approach to symmetry in inventories

When looking at the Italian inventory (16.2), we see that the four binary features generate together 16 possible combinations, whereas Italian uses only seven of them, and not a random subset. Now, in Italian, either of the features [back] or [round] is redundant, in the sense that every [+back] vowel is also [+round], and every [+round] vowel is also [+back]. So, as far as the inventory is concerned, we could do without the feature [round] (or [back]) at all. The remaining three features give eight possible combinations, which is enough for Italian. However, if we look at the French inventory:

	front unrounded	front rounded	back rounded
close	i	y	u
close-mid	e	ø	o
open-mid	ɛ	œ	ɔ
open	a		

height ↑

place → (16.7)

we see that for French we do need all four features [open], [mid], [back] and [round] in the following ***complete specification***:

	i	y	u	e	ø	o	ɛ	œ	ɔ	a
[open]	–	–	–	–	–	–	+	+	+	+
[mid]	–	–	–	+	+	+	+	+	+	–
[round]	–	+	+	–	+	+	–	+	+	–
[back]	–	–	+	–	–	+	–	–	+	–

(16.8)

Now we take from the theory of ***contrastive underspecification*** (Clements 1987, Steriade 1987) the idea that the inventory can be described by its ***redundancies***:

(a) All back vowels are rounded; and, logically, all unrounded vowels are front.

(b) Low (i.e., open non-mid) vowels are unrounded, so that rounded open vowels must be mid, and rounded non-mid vowels must be non-open.

Instead of the original derivational formulation, it is easier to describe these facts as language-specific ***constraints*** or ***well-formedness conditions*** on possible segments, expressible as OT-able output constraints analogous to the structural constraint in (16.5):

*[+back & –round] ; *[+open & –mid & +round] (16.9a;b)

Instead of this bottom-up approach (from segments to constraints), we could also perform the top-down procedure, borrowed from the theory of ***radical underspecification*** (Archangeli 1984, 1988), of starting with a language-specific feature set, deriving a maximum inventory, and limiting this with universal and/or language-specific constraints. Thus, the four binary features [open], [mid], [round], and [back] yield 2^4 = 16 possible sounds; of these, the four back unrounded vowels [ɯ], [ɤ], [ʌ], and [ɑ] are ruled out by (16.9a), and the two low rounded vowels [ɶ] and [ɒ] by (16.9b). This leaves exactly the ten attested vowels. So the redundancy constraints form not only ***necessary***, but also ***sufficient*** conditions on possible feature combinations.

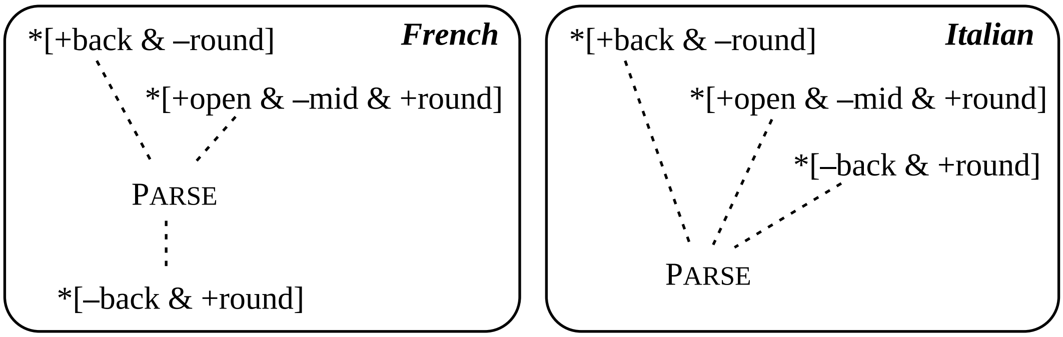
According to the principle of the ***richness of the base*** (Prince & Smolensky 1993), the limitations of each language are caused not by limitations in the lexicon, but by the workings of the constraint system. For inventories this means that the underlying form could contain any universally possible utterance, specified in the usual universal (hybrid) phonological features, and that the grammar filters this into a well-formed utterance. For

French, the inventory should *follow* from the set of four features together with the two constraints (16.9a;b) dominating faithfulness. For instance, a hypothetical underlying /ʌ/ is filtered into a different vowel, most likely [ɔ] or [ɛ], depending on the exact ranking of the faithfulness constraints:

Input: /ʌ/	*[+back & –round]	PARSE (+back)	PARSE (–round)
[ʌ]	*!		
☞ [ɔ]			*
[ɛ]		*!	

(16.10)

Collapsing the PARSE constraints, the French vowel system can be described with the grammar (at the left)



(16.11)

We can now describe sound inventories in a generalizing manner: given only four features and two constraints, we can derive a system of ten vowels. Most languages seem to have this kind of economically representable grammars, a phenomenon that we identified earlier as the preference for symmetry. With a basic tenet of radical underspecification theory, it would be good if the constraints proved universal, i.e., if all languages draw from the same set of non-conflicting constraints. The two constraints of the French vowel structure, at least, are widely found throughout the world. In fact, both the Italian and the French system can be described with the same constraint set, but with different rankings, as (16.11) shows. A simpler representation of Italian, however, without the feature [round], would only have the single dominating constraint *[+open & –mid & +back], if /a/ is considered a front vowel.

Note that the structural filters can still be explained only very indirectly.

16.1.4 Evaluation of the phonological approach

The most important requirement in the phonological approach is ***empirical adequacy***: we aim at a theory that predicts what is possible in language and what is not. The strengths of this approach are reflected in two main points:

(a) **Symmetry**. If the distinctive features for a language have been identified, the principle of ***maximum use of available features*** guarantees symmetry in inventories.

(b) **No autonomous inventories**. In most phonological theories, inventories are not posited but follow from the grammar. In Radical Underspecification theory (Archangeli 1988), the segments follow from the features, which are used maximally, and from the constraints, which restrict the use of feature combinations. For instance, the default rule [+back] → [+round] was meant to fill in the values of unspecified features, and at the same time this rule was an implicational generalization on segment structure. The optimality-theoretic correlate of this position is the richness of the base, combined with structural constraints.

The supposed connection between inventories and production grammars makes empirical predictions: non-contrastive or redundant feature values are thought to be phonologically inert. Contrastive Underspecification theory (Steriade 1987), for instance, holds that only contrastive features can be transparent to spreading.

The downside of the phonological approach is its lack of ***explanatory power***, which the filter constraints of the previous section still share with Chomsky & Halle's (1968) ***markedness conventions***, of which they were mere reformulations. The constraints *follow* from the language data, and can, therefore, not *explain* these data; they do not tell us why we should maintain certain features, why these features should be binary, and why we should maintain certain constraints and not others.

Features. The term *distinctive features* suggests that features are chosen on the basis of their ability to implement perceptual distinctions. So, vowel height would correspond to the acoustic and perceptual feature of first formant, and the front-back distinction to the second formant. The terms [back] and [round], however, sound as if they refer to articulatory gestures. To quote Hammarström (1973: 161): "[Using] articulatory terms to describe auditory facts (...) may be acceptable for the purpose of many descriptions (as long as one knows what one is doing)". The danger is that if the next generation is not told what they are doing, they will take the articulatory terms at face value. From the functional standpoint, distinctive features can only be perceptual (i.e., auditory and visual) categories, because proprioceptive categories cannot be communicated directly.

Binarity. In (16.8), the really natural class [+open & –mid] has a more complicated representation than the alleged class [–mid], which would contain /i/ and /a/. This strange situation is a direct result of the obligatory binarity, which breaks up phonetically continuous dimensions.

The real solution to the quantization problem is to let go of binarity as an organizational necessity, and regard vowel height as multi-valued (Ladefoged 1971). Communicatively, the notion of an originally continuous vowel-height feature is not problematic at all: because of innate capacities of human perception, the learner will divide it into a finite number of categories. This number is language-specific; the fact that many features are binary is caused by nothing more than the inability of the listener to distinguish faithfully more than two values of those perceptual features.

Constraints. The largest problem is how to restrict the output of the grammar. In (16.11), we could indirectly detect two articulatory constraints (against lip rounding and against jaw widening), and two perceptual constraints (favouring maximal distinctivity by requiring back vowels to be round and by preferring that all vowels are either low or high). Radical underspecification theory (Archangeli 1988) tells us that the ***default rules*** [] → [–round] ("by default, vowels are not rounded") and [+back] → [+round] ("back vowels are rounded") are universal and innate. However, these indirectly stated rules express what we would expect to result from functional considerations: the former is "we'd rather not perform a lip rounding gesture", and the latter is "to implement perceptual backness (low F_2) in a contrast with front vowels, we have to make a tongue-backing gesture as well as a lip-rounding gesture". We would like, therefore, to disentangle these explanations into directly expressed functional constraints. Further, Archangeli admits that language-specific constraints are also needed, but we will see that these can be expressed directly as general functional principles as well.

The ***innateness*** requirement seems to be connected to the general lack of explicability, although a learnability issue has also often been advanced (for OT: Tesar & Smolensky 1993). From the functional standpoint, we can explain constraints and show that they can be learned (chapter 14), so we need not assume their innateness. Gaps in sound systems, expressed here as arbitrary filters, will be seen to be caused in fact by asymmetries in the human speech production and perception systems.

16.1.5 The "phonetic" approach to sound inventories

Phonetic attempts to explain sound inventories have used only a few functional principles.

Kawasaki (1982) restricted her explanations to the two perceptual principles of maximization of distinction and salience.

Stevens (1989) tried to explain the commonness of some sounds as the minimization of precision and the simultaneous maximization of acoustical reproducibility.

Liljencrants & Lindblom (1972) investigated what vowel systems would look like if they were built according of the principle of maximum perceptual contrast in a multi-dimensional formant space. They searched for the optimal 7-vowel system by maximizing within a fixed two-dimensional perceptual space the perceptual contrast, which they defined as the sum of inverse-squared distances between all pairs; they based the distance between two vowels on the difference in F_1 and F_2 expressed in Mels. The results were not satisfactory: because they gave equal weight to F_1 and F_2 differences, the simulated systems showed too many place contrasts relative to the number of height contrasts.

Lindblom (1986) did the same by comparing all subsets with seven vowels taken from a fixed set of 19 'possible' vowels, and choosing the subset that has the largest internal perceptual contrast, based on the distance between two vowels in terms of the difference between the excitation patterns that the vowels would give rise to in the inner ear of a listener. This did not solve the F_2 problem.

Ten Bosch (1991) explained vowel systems on the basis of maximal distinctions within an articulatory space bounded by an effort limit based on the distance from the neutral vocal-tract shape. He decided to ***fit*** the parameter that determines the relative importance of the front-back distinction with respect to the importance of the height distinction, to the data of the languages of the world, assigning a value of 0.3 to the relative importance of the second-formant distance with respect to the first-formant distance.

A similar approach is found in Boë, Perrier, Guérin & Schwartz (1989), Schwartz, Boë, Perrier, Guérin & Escudier (1989), Vallée (1994), Boë, Schwartz & Vallée (1994), Schwartz, Boë & Vallée (1995), and Schwartz, Boë, Vallée & Abry (1997). Their simulations pointed to a value of 0.25 for the weighting of the F_2 distance.

In an attempt to derive, instead of fit, the relative unimportance of place distinctions with respect to height distinctions, Lindblom (1990) suggested that for determining the contrast between two vowels, proprioceptive contrasts in the speaker (jaw height can be felt more accurately than tongue-body place) are equally important as auditory contrasts in the listener. His predicted ‘optimal’ 7-vowel system was

i	ʉ		u	
		ɤ		
ɛ				
a		ɑ		(16.12)

which he considered to be in “extremely close agreement” (p. 79) with the most common 7-vowel systems found in Crothers (1978), which are

i	ɨ	u		i		u	
e	ə	o		e		o	
	a			ɛ		ɔ	
					a		(16.13a;b)

16.1.6 Evaluation of the “phonetic” approach

The main problem with a result like (16.12) is that it is descriptively totally inadequate: it shows no symmetry, no features, no organization. None of these approaches derives the symmetry that is visible in (16.1). Schwartz, Boë, Vallée & Abry (1997: 261) admit that symmetry “does not always emerge from the intrinsic principles of the theory”. Indeed, each of their four proposed six-vowel systems is less symmetrical than any of the four most common six-vowel systems in Maddieson’s (1984) database. Basically, the cause of the problem is that the distance function will actually favour an asymmetry of height between front and back vowels, because a difference in F_1 will always contribute positively to the perceptual distance between a pair of vowels.

Also, Lindblom takes ***finiteness*** for granted, as witnessed by his use of a finite inventory of phonemes. Schwartz et al. (1997: 265) state that “the problem of the finiteness of the number of speech sounds, important from a theoretical point of view, is in fact impossible to address in a technically satisfying way”. The Lindblom school appears to consider tone, duration, and voice quality to be independent features, as witnessed by his neglect of these dimensions. Apparently, these three features are tacitly considered “suprasegmental”, or better: independent from the other (here: spectral) features; we can call this ***autosegmental***. But for large vowel inventories, F_1 is an autosegmental feature like the others; we can see that when we realize that it is an acoustically distinct aspect of vowels, ready to be divided up into a number of perceptual categories by the language learner.

Lindblom himself (1990) tries to tackle the symmetry problem, and boasts of having found self-organization in a hypothetical language consisting of nine CV utterances only. The nine utterances that emerged most often in repeated simulations were rather symmetric together, but these were not simulated as a group. Nevertheless, let’s concede that Lindblom’s optimization criterion would yield the following very symmetric set of non-low vowels:

i	y	u
e	ø	o
ɛ	œ	ɔ

(16.14)

Even then, the symmetry would break down if we asked Lindblom’s optimization criterion to give us eight instead of nine utterances. Without performing the actual simulation, we can predict that Lindblom’s strategy will yield something like:

i	y	u
e	ø̞	o
ɛ		ɔ

(16.15)

because the perceptual space gets narrower as vowel height decreases. In reality, however, we find things like the Frisian short-vowel system (16.1) without a lowered /ø/, thus retaining four vowel heights. Obviously, it is the features, not the segments, that structure sound systems.

It seems thus impossible to build an algorithm for generating possible sound systems without symmetrizing principles.

So, the phonetic approaches do not perform well on describing symmetry, which we identified in §16.1.4 as one of the strong points of the phonological approach. The other point was the connection between inventories and the grammar; in all the phonetic approaches, the modelling of inventories is a goal in its own right, and the grammar (natural classes, output constraints) is not even considered.

What, then, could be the strong points of the phonetic approaches?

(c) **Predicting dispersion**. Phonetic principles could explain some of the constraints on the basis of perceptual contrast: if back vowels are round, they are more unlike front vowels than if they are not round; maximizing the perceptual contrast helps the listener to recognize the speaker’s message. Further, the vowel bucket is narrower for the low vowels than it is for higher vowels, i.e., the distance between the F_2’s of [a] and [ɑ] is much smaller than the distance between the F_2’s of [i] and [u]; this answers the question which of the eight possible feature combinations in Italian should be the most likely candidate for not being found (namely, a low vowel, be it back or front).

(d) **Predicting the gap**. Given three height and place features, the maximal inventory of non-low vowels is (16.14). The phonetic approaches can answer the question: if this system has one gap, where will it be? The answer is that the gap will be at /œ/, because the front-back distance is smaller there than at the other heights. This simple contrast-based account seems more natural than the awkward feature-cooccurrence constraints of the phonological approach.

(e) **Predicting the arity**. If height and place are multi-valued features, how many values will they have? Specifically, what is the relation between the average number of heights and the average number of places? Unfortunately, the phonetic approaches have not been able to derive this relation, although the general conviction is that it would be possible if we knew enough about the perception of frequency spectra.

Unfortunately, these approaches have not yet been able to measure any phonetic spaces; a problem with one degree of freedom can always be "solved" by fixing one parameter. Unless we accept Lindblom's (1990) proposal for taking into account the speaker's proprioceptive height and place distinctions, the relative importance of the first formant must be sought in its greater loudness with respect to the second formant: the second spectral peak has a larger chance of drowning in the background noise. In chapter 4, I computed the distances between the basilar excitation patterns of [a], [i], and [u] in units of just-noticeable differences (jnd), and found that the distance between [i] and [u] was 12 jnd, and the distance between [a] and each high vowel was 18 jnd. This means that a system with four heights is equally well dispersed as a system with three places, namely, with 6 jnd between each pair of neighbours. This would predict that (16.13a) and (16.13b) would be equally common inventories, and this seems to be the case.

16.1.7 "Integrated" approach 1: enhancement

It seems that we will need to combine phonological and phonetic principles if we want to describe and explain inventories at the same time. The example of the rounding of back vowels will make this clear.

In the vowel systems of the languages of the world, most back vowels are round and most rounded vowels are back. The "phonological" approach has not given any explanations for this fact: the correlation between [round] and [back] was viewed as a part of Universal Grammar, hard-wired into the human language faculty. In phonetic terms, however, the explanation of the correlation between [round] and [back] is straightforward. For a maximal perceptual contrast between two places of articulation, a language should have unrounded front vowels (maximum F_2) and rounded back vowels (minimum F_2).

Even in phonetics, however, the necessary distinction between perception and production seems not always to be made. Stevens, Keyser & Kawasaki (1986) speak of the ***enhancement*** by lip rounding of the perceptual contrast between vowels with high and vowel with low F_2. With a proper division of labour between perception and production, the statement should be altered to: "a maximal F_2 contrast is *implemented* by having a group of vowels with front tongue position and lip spreading, and a group with back tongue position and lip rounding". Rounding, therefore, does not enhance a contrast, but helps to implement it. For why should styloglossus be the ***agonist***, and orbicularis oris the ***synergis****t*? The asymmetric interpretation by Stevens et al. of this phenomenon as the enhancement of backness by rounding smacks of a confusion of the phonological feature [back], which can be used as an arbitrary label for a certain perceptual contrast, with the articulatory gesture of backward tongue-body movement. Apparently acknowledging this problem, Stevens & Keyser (1989) explicitly divide phonological features into primary and secondary features. While this move was in itself data-driven, because partly based on commonness in speech, the notion that frequency of occurrence has a strong correlation with perceptual distinctivity, is indubitable.

16.1.8 “Integrated” approach 2: inventory constraints

A functionally-oriented Optimality-Theoretic account was given by Flemming (1995), who handles inventories as the result of the interactions between the functional principles of maximizing the number of contrasts and maximizing the auditory distinctiveness of contrasts. These two principles correspond to Passy’s (1891) assertion that speakers will try to get their messages across as *quickly* and *clearly* as possible (respectively).

These principles lead to fixed rankings, e.g. for high vowels along a fixed F_2 axis { i y ɨ ɯ u }. First (for the maximization of the rate of information flow), it is more important to maintain *two* contrasts than it is to maintain *three* contrasts:

MAINTAIN 1 F_2 contrast >> MAINTAIN 2 F_2 contrasts >>
>> MAINTAIN 3 F_2 contrasts (16.16)

Secondly (for the minimization of confusion), it is less bad to have two vowels at an “auditory distance” of *three* steps along the discretized F_2 axis { i y ɨ ɯ u } than it is to have them at a distance of *two* steps:

MINDIST$_{F2}$ = 1 >> MINDIST$_{F2}$ = 2 >> MINDIST$_{F2}$ = 3 >> MINDIST$_{F2}$ = 4
(16.17)

Interleaving these two constraint families in a dedicated inventory grammar (i.e. a grammar that evaluates inventories directly), we can choose a grammar that gives { i ɨ u } as the best inventory:

	MAINTAIN 1 contrast	MINDIST = 1	MAINTAIN 2 contrasts	MINDIST = 2	MINDIST = 3	MAINTAIN 3 contrasts
i-u			*!			*
☞ i-ɨ-u					*	*
i-ɨ-ɯ-u				*!	*	*
i-y-u				*!		
i-y-ɯ				*!		

(16.18)

MINDIST is a constraint formulated as the OT optimization criterion itself: “minimize the maximum problem”; therefore, it is probably the surface result of a more primitive constraint system, e.g. with constraints like “distance ≤ 2”. In contrast with (16.18), such a system would rank the inventory { i y u } above { i y ɯ }, because all three pairs would be evaluated, not just the closest pair; this is a desirable property.

The system { i ɨ u } turns out to be better than the Frisian system { i y u }, for every possible ranking of the constraints, as long as the rankings (16) and (16.17) are kept fixed. The Frisian preference for { i y u } over { i ɨ u } probably has to do with the choice of the gestures that should implement the central F_2 value: either with frontal tongue-body raising and lip rounding, or with central tongue-body raising. To account for this, constraints against performing the relevant articulatory gestures should be added to the inventory grammar, and Flemming does so in several cases.

But there is a problem with Flemming's approach, namely that (16.18) does not represent a production grammar, i.e., it is not a model of how a speaker converts underlying to surface forms: it evaluates inventories instead of output candidates. Flemming gives up a requirement still honoured by the underspecification approaches of §16.1.3, namely, that inventories are built on the same principles as the grammar. In an OT production grammar, the connection with the inventory can be upheld by the principle of richness of the base; inventory grammars like (16.18), however, do not explain how a random input is filtered into a well-formed utterance. Thus, while Flemming's approach is more advanced than any of the phonetic approaches discussed earlier, as it combines the notion of sufficient contrast while taking symmetry for granted, and though the notion of the interaction between articulatory effort and perceptual contrast is correct, Flemming's global inventory evaluation procedure is not a model of grammar; it just shows that inventories can be described with strict ranking of principles, just like so many real-life weighings of pros and cons. If, by contrast, the functional principles could be expressed directly in a local production grammar, and this production grammar could derive inventories from richness of the base, a separate global inventory grammar would be superfluous; I will show below how this can be achieved.

16.1.9 "Integrated" approach 3: local expression of functional principles

The faithfulness and structural constraints of (16.5) and (16.11) have direct analogues in functional principles. Structural constraints limit the possible surface structures; the functional principle of minimization of articulatory effort can be expressed in articulatory constraints against the performance of certain gestures. Faithfulness constraints punish any differences between underlying and surface forms; if the two forms are equal, an underlying contrast is still heard on the surface; thus, faithfulness constraints can implement the functional principle of minimization of perceptual confusion in a local manner, without having to compare any forms with all possibly contrasting forms as in Flemming's inventory evaluation procedure.

The implementation of gestural and faithfulness constraints in a theory of grammar requires a principled distinction between articulatory and perceptual features, so we have no hope of translating (16.11) directly into functional grammars for French and Italian. Instead, we should start a bottom-up procedure from first principles. This will be performed rigorously in the next sections. The resulting theory will combine all the desirable properties that we found in the phonological and phonetic approaches discussed above:

(a) **Symmetry**. Follows from the finiteness of the number of learned perceptual categories and articulatory gestures.

(b) **No autonomous inventories**. Inventory structure follows directly from the constraints in the production and perception grammars, not from a dedicated inventory grammar.

(c) **Predicting dispersion**. Sufficient contrasts emerge from the fact that a listener is also a speaker: local minimization of confusion demands enhancement of contrasts in phonetic implementation.
(d) **Predicting the gap**. The locations of the gaps follow from asymmetries in articulatory effort and perceptual contrast, as these are reflected in the local rankings of gestural, faithfulness, and categorization constraints.

16.2 Functional Phonology

Functional Phonology makes a principled distinction between articulatory and perceptual representations, features, and constraints.

16.2.1 Representations and grammars

As illustrated in figure (6.1), the speaker's ***production grammar*** handles the evaluation of the perceptual specification, the articulatory implementation, and the perceptual result, and their relations; the listener's ***perception grammar*** maps the acoustic features of an utterance onto language-specific numbers of values along language-specific perceptual dimensions. It is used by the listener for the initial categorization of acoustic speech events, and by the speaker to monitor her own output. In this book, I make the simplifying assumption that for the listener, perceptual categorization is *followed* by the recognition process.

16.2.2 Gestures and features

From (16.1), we see that Frisian speakers must have acquired the articulatory gestures of rounding and spreading of the lips, fronting, backing, and lowering of the tongue body, and lowering of the jaw. Most of the five gestures of the Frisian vowel system exist in various degrees of distance moved away from the neutral position (to stress the activity of the movement, the list represents each gesture with one of the main muscles involved):

(a) The back of the tongue is raised further for [o] and [ø] (by styloglossus activity) than for [ɔ], and even more so for [u] and [y].
(b) The front of the tongue is raised further for [e] and [ø] (by genioglossus activity) than for [ɛ], and even more so for [i] and [y]
(c) The lips are rounded more strongly for [ø] and [o] (by orbicularis oris activity) than for [ɔ], and even more so for [y] or [u].
(d) The lips are spread more strongly for [e] (by risorius activity) than for [ɛ], and even more so for [i].
(e) The jaw is lowered further (by mylohyoid activity) for [a] than for [ɛ] and [ɔ].
(16.19)

These things are shown schematically in the top half of table (16.20) (the actual numbers are meaningless; they just enumerate locations along a continuous scale):

		i	y	u	e	ø	o	ɛ	ɔ	a
art.	body: back up	0	0	3	0	0	2	0	1	0
	body: front up	3	3	0	2	2	0	1	0	0
	lips: round	0	3	3	0	2	2	0	1	0
	lips: spread	3	0	0	2	0	0	1	0	0
	jaw: down	0	0	0	0	0	0	1	1	2
perc.	height	4	4	4	3	3	3	2	2	1
	place	fr	ce	ba	fr	ce	ba	fr	ba	ce
	(round)	–	+	+	–	+	+	–	+	–

(16.20)

From (16.1), we see that Frisian listeners must have acquired four vowel heights. Thus, Frisians have learnt to distinguish four different F_1 “contours” for short vowels. The front-back dimension can be associated with the second main spectral peak: we have the values “maximum F_2, given the value of F_1” (implemented by the acquired gestures of tongue-body fronting and lip spreading), “minimum F_2, given the value of F_1” (implemented by tongue-body backing and lip rounding), and a value in between (implemented by tongue-body fronting and lip rounding). The perceptual features of the nine Frisian short vowels are shown schematically in the bottom half of table (16.20); “fr”, “ce”, and ”ba” stand for [front], [central], and [back], but could also have been named 1, 2, and 3, or [high F_2], [mid F_2], and [low F_2].

Articulatorily or perceptually related segments can form natural classes in phonological processes. Thus, we can talk of the articulatorily defined class of rounded vowels, or of the perceptually defined class of higher mid vowels.

16.3 Finiteness

The most important thing to be learnt from (16.1) is the fact that only nine short vowels occur in tens of thousands of words; and this is also the main fact that has to be explained.

16.3.1 Articulatory constraints and finiteness

The typical articulatory constraint that occurs in the speaker’s production grammar is (§7.2):

*GESTURE (*g*): “an articulatory gesture *g* is not performed.” (16.21)

The acquisition of motor skills has supplied every speaker with only a finite number of gestures that she can perform. The only “real” gestural constraints that are visible at all in a speaker-oriented grammar and have any claim to psychological reality, are the constraints against the acquired gestures: each of these must be dominated by at least one other constraint, typically a specification-to-perception faithfulness constraint like PARSE, which says that a specified perceptual feature value shall be implemented by *any* gesture.

From the universal descriptive linguistic standpoint, however, there would exist a constraint against every thinkable gesture that humans could learn to perform. Now, most of these universal constraints are undominated and play no role at all; these "virtual" constraints are merely a descriptive device for communication between linguists: they can describe aspects of the learning process and the production of loan words. For instance, the absence in the Dutch speaker's brain of any structures referring to gestures that implement implosives, can be described by an undominated *GESTURE (hyoid: lower) constraint.

Thus, citing (8.1), "low-ranked *GESTURE and *COORD constraints determine the language-specific finite set of allowed articulatory features and feature combinations". Therefore, Frisian grammars must simply contain the following dominated constraints: *GESTURE (lips: rounded), *GESTURE (lips: spread), *GESTURE (body: front up), *GESTURE (body: back up), *GESTURE (body: low), and *GESTURE (jaw: low).

16.3.2 Perceptual constraints and finiteness

The relevant perceptual constraint that occurs in the listener's perception grammar is (§8.3):

*CATEG (*f*: *v*): "the perceptual feature *f* is not categorized as the value *v*." (16.22)

From the linguistic standpoint, there exists a constraint against every thinkable category that humans could learn to perceive. In this sense, the set of categorization constraints is universal, and these constraints are innate in the sense that every normal human child can learn to perceive any category.

However, the acquisition of perceptual classification has supplied every listener with only a finite number of categories that she can perceive. In the grammar of every listener, therefore, most of the universal categorization constraints are undominated and play no role at all; again, these "virtual" constraints are merely a descriptive device for communication between linguists: they can describe aspects of the learning process and the perception of loan words. The only "real" categorization constraints that are visible at all in the listener's perception grammar, are the constraints against the acquired categories: these constraints must be dominated by at least one other constraint, typically the peripheral acoustics-to-perception correspondence constraint PERCEIVE, which says that it is important that an acoustically available feature shall be classified into *some* category.

Thus, citing (8.14), "low-ranked *CATEG constraints determine the finite set of allowed perceptual feature values". Therefore, Frisian perception grammars must simply contain the following dominated constraints: *CATEG (height: open) (= *CATEG (F_1: maximum)), *CATEG (height: open-mid), *CATEG (height: close-mid), *CATEG (height: close), *CATEG (place: front) (= *CATEG (F_2: maximum)), *CATEG (place: centre), *CATEG (place: back).

Now we know the causes of the finiteness of segment inventories. Citing (8.15), "the continuous articulatory and perceptual phonetic spaces are universal, and so are the constraints that are defined on them; the discrete phonological feature values, however, are language-specific, and follow from the selective constraint lowering that is characteristic of the acquisition of coordination and categorization.” An exhaustive use of four vowel heights and three places would lead to a system of twelve vowels, which is three more than Frisian actually has.

The dependence of symmetry on inventory size can be explained with a general property of categorization: the number of perceptual dimensions increases with the number of classes. Speakers of a four-vowel system may recognize the four different excitation patterns associated with /a ɛ i u/; whereas speakers of a 18-vowel system cannot recognize 18 unrelated percepts, but divide up the perceptual space along at least two dimensions: “place” and “height”.

16.3.3 Faithfulness constraints and finiteness

The categorization constraints are not expressed directly in the production grammar. In the production grammar, the categorization is reflected by faithfulness constraints.

An important principle of effective communication is the requirement that specified features are received by the listener. Because the speaker is a listener, too, the correspondence constraint TRANSMIT (§9.2) requires that a specified value (category) of a perceptual feature is heard by the listener as *some* category on that same perceptual tier, and the constraint *REPLACE (§9.2) forbids the two corresponding values to be different. For features with few categories (in this chapter, even vowel height will be taken to be such a feature), we can collapse the correspondence and similarity requirements into a single constraint *DELETE or PARSE (§9.8):

PARSE (*f*: *v*): “an underlyingly specified value *v* of a perceptual feature *f* appears (is heard) in the surface form”. (16.23)

In Frisian, therefore, we have PARSE constraints for all perceptual categories. These constraints can be abbreviated as PARSE (open), PARSE (open-mid), PARSE (close-mid), PARSE (close), PARSE (front), PARSE (centre), and PARSE (back).

16.3.4 Gaps and richness of the base

From the functional standpoint, the input to the grammar must be specified in perceptual feature values, i.e. categorizable values of perceptual dimensions specific to the language. For Frisian, this would mean that the input may contain 12 different short vowels, if the categorization of place is independent of height (which is open to doubt, see §16.5). So:

Richness of the base (functional version):

> “the input may contain any combination of categorizable perceptual features; the combinations that do not occur on the surface are filtered out by the constraint system.” (16.24)

For Frisian, this means that the constraint system will have to explain the gap in the open-mid-vowel system, and the two gaps in the open-vowel system.

16.4 Local ranking

According to the ***local-ranking principle*** (chapter 11), gestural and faithfulness constraints can be locally ranked with the functional principles of minimization of articulatory effort and perceptual contrast.

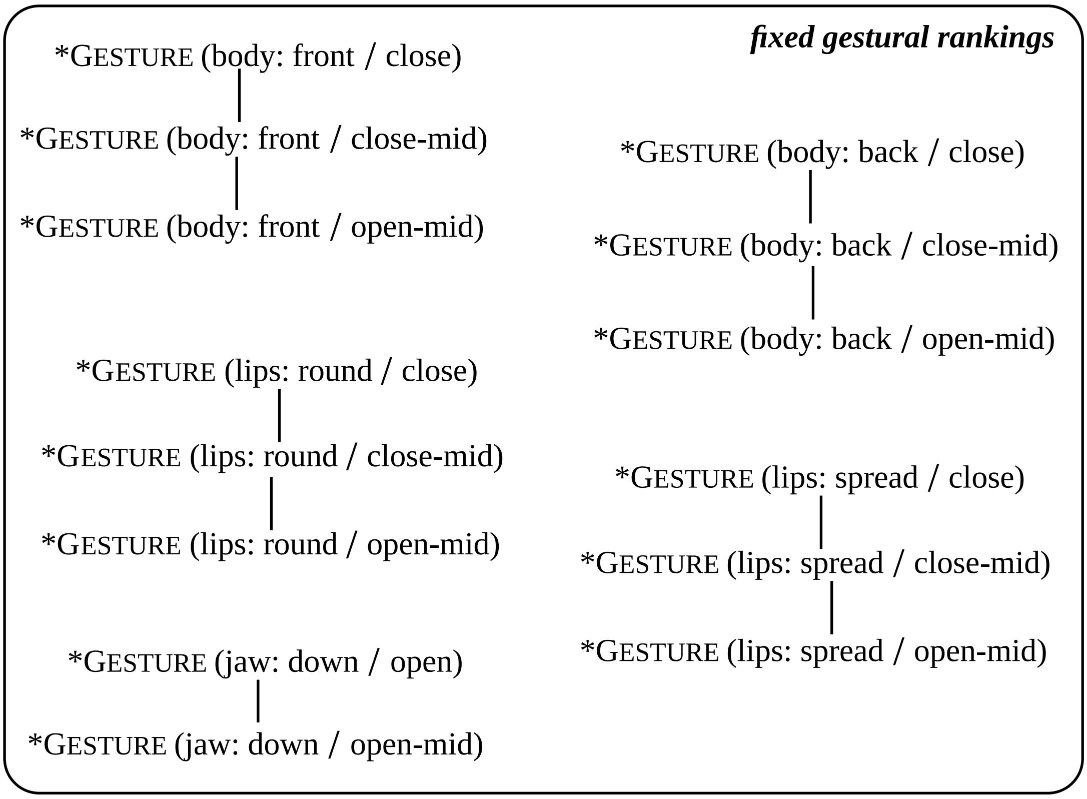
16.4.1 Local ranking of gestural constraints

The ranking of a gestural constraint may depend on a number of circumstances. These dependences effectively split each *GESTURE constraint into a multidimensionally continuous family (7.34):

*GESTURE (*a*: *g* / *d*, *v*, *p*, *t*): “the articulator *a* does not perform the gesture *g* along a certain distance *d* (away from the rest position), and with a certain speed *v*, reaching a position *p* for a duration *t* .” (16.25)

Basically, articulatory constraints are ranked by ***effort***: constraints against gestures that require more effort are universally ranked higher than constraints against easier gestures, but only in the following case: the *same* gesture is more difficult if its distance, speed, duration, *or* precision is greater, *and everything else is kept equal*; this can lead to a fixed ranking of gestural constraints.

With (16.19), this yields the following fixed distance-based rankings, given the Frisian gesture system:


fixed gestural rankings
*GESTURE (body: front / close)
*GESTURE (body: front / close-mid)
*GESTURE (body: front / open-mid)
*GESTURE (body: back / close)
*GESTURE (body: back / close-mid)
*GESTURE (body: back / open-mid)
*GESTURE (lips: round / close)
*GESTURE (lips: round / close-mid)
*GESTURE (lips: round / open-mid)
*GESTURE (lips: spread / close)
*GESTURE (lips: spread / close-mid)
*GESTURE (lips: spread / open-mid)
*GESTURE (jaw: down / open)
*GESTURE (jaw: down / open-mid)

(16.26)

Since this height-dependent differentiation of the vowel constraints seems to be small once the gestures are mastered, it will be ignored in the rest of this paper. Clearer examples may be found in obstruent voicing and tongue-root inventories.

Many languages with voicing contrasts in obstruents still lack a segment /g/ in their inventory of plosives, i.e., the symmetry is broken by a *gap* at /g/:

(p)	t	k	
b	d	(g)	(16.27)

It is more difficult to maintain a voicing contrast in plosives with a closure close to the larynx, than it is at other places. One of the preconditions for phonation is the presence of a stream of air through the glottis. During the closing interval of plosives, both the nasal and oral pathways are closed, and the flow through the glottis will eventually stop. One of the things that influence the maintenance of the flow is the amount to which the supralaryngeal air will be allowed to expand. For the back closure of [g], the cavities above the glottis are filled earlier with air than in [b] and so voicing will stop earlier in [g] than in [b] because of the more rapid drop in transglottal pressure (§5.12; Ohala & Riordan 1979).

Thus, a specified degree of voicing is more difficult to maintain for a dorsal plosive than for a labial or coronal plosive. Likewise, a specified degree of voicelessness is more difficult to implement for a labial plosive than for a coronal or dorsal plosive. This leads to a fixed hierarchy of implementation constraints for voiced and voiceless plosives:

equal-contrast obstruent voicing

*[+voiced / plosive / dorsal]	*[–voiced / plosive / labial]
|	|
*[+voiced / plosive / coronal]	*[–voiced / plosive / coronal]
|	|
*[+voiced / plosive / labial]	*[–voiced / plosive / dorsal]

(16.28)

Because the degrees of voicing and voicelessness were taken constant, we can assume a homogeneous PARSE constraint for the plosive voicing feature values. According to our version of richness of the base, the constraint system should remove an underlying /g/ in a language that lacks [g] at the surface. This will indeed be the outcome if PARSE (±voice) is sandwiched between the coronal and dorsal voicing constraints:

/g/	*[+voice / dorsal]	PARSE (±voice)	*[+voice / coronal]
[g]	*!		
☞ [k]		*	

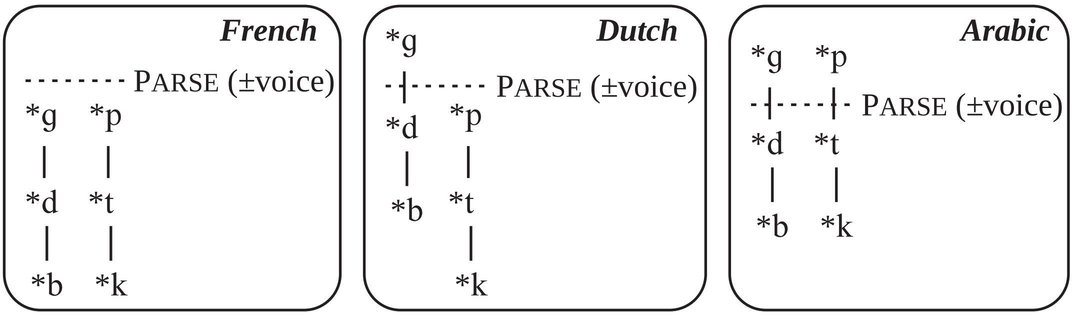
(16.29)

Note that with the same hierarchy, coronal voiced plosives surface faithfully:

/d/	*[+voice / dorsal]	PARSE (±voice)	*[+voice / coronal]
☞ [d]			*
[t]		*!	

(16.30)

As the hierarchy for [+voice] is independent from the hierarchy of [–voice] (they use different types of gestures), the three following grammars are some of the possibilities (the gestural constraints are maximally abbreviated; the homogeneous PARSE constraint is shown by a dotted line):



(16.31)

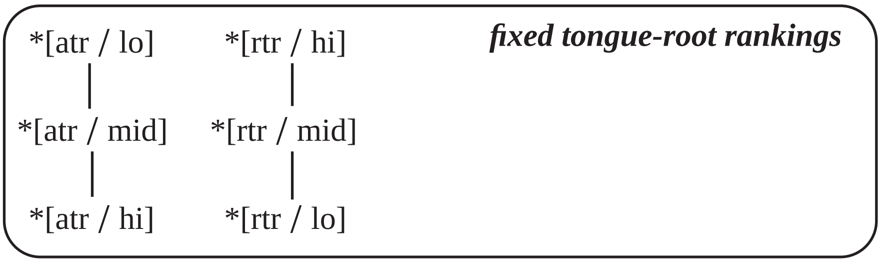
Thus, French shows no gaps, Dutch lacks [g] and Arabic lacks both [p] and [g].

In the realm of vowel inventories, we find analogous rankings in tongue-root systems. If the short-vowel system becomes much larger than the Frisian example of (16.1), it is probable that speakers construct a third dimension. This is a general property of categorization. If a language has two vowel places (front and back) and more than four segments should be distinguished, the language has the option of dividing the F_1-based height dimension into two new dimensions, say the perceptual correlates of tongue-body (oral) constriction and tongue-root (pharyngeal) constriction, which we shall call [height] and [tr], respectively. Most tongue-root languages (Archangeli & Pulleyblank 1994), have three categories along the height dimension (low, mid, high), and two along the tongue-root dimension (atr, rtr). As is explained in more detail chapter 14, the following rankings of articulatory effort can be posited:

(a) The [atr] value is more difficult to implement for lower than for higher vowels.
(b) The [rtr] value is more difficult to implement for higher than for lower vowels.

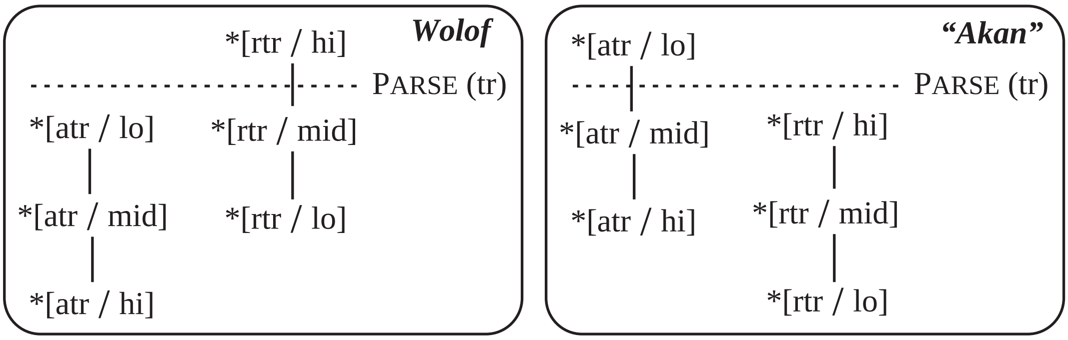
(16.32)

With the most common categorization, this leads to the following fixed hierarchies of implementation constraints:



(16.33)

These are larger sets of constraints than Pulleyblank's (1994) two ***grounding constraints*** LO/RTR "if a vowel is low, then it has a retracted tongue root" and HI/ATR "if a vowel is high, it has an advanced tongue root", whose actions are comparable to those of *[atr / lo] and *[rtr / hi], respectively. From the functional standpoint, we should derive, not posit, which of the many possible constraints tend to be strong and which tend to be weak: *[atr / mid], for instance, also exists although it may be universally lower ranked than *[atr / lo]. Again, we can assume a homogeneous PARSE (tongue root) constraint, because our use of implementation constraints supposes equal tongue-root contrasts for all three heights. Two of the possible grammars are (Archangeli & Pulleyblank 1994):



(16.34)

From the set of categorizable front vowels { i, ɪ, e, ɛ, ə, a }, Wolof lacks [ɪ] (if PARSE (height) is ranked high, a hypothetical underlying /ɪ/ would become [i]), and (a hypothetical lexical stratum of) Akan lacks /ə/.

16.4.2 Local ranking of faithfulness constraints

The ranking of a faithfulness constraint for a particular perceptual feature may depend on the simultaneous presence of other features and on the perceptual events preceding and following that feature:

PARSE (*f*: *v* / *condition* / *environment*): "the value *v* on the perceptual tier *f* in the input is present in the output under a certain *condition* and in a certain *environment*." (16.35)

Basically, faithfulness constraints are ranked by perceptual ***contrast***: constraints that require the faithfulness of strongly distinctive features are ranked higher than constraints for weakly distinctive features, but only in the following case: the *same* replacement is more offensive if the difference between the members of the pair along a certain perceptual dimension is greater, *and everything else is kept equal*; this can lead to a fixed ranking of many pairs of faithfulness constraints.

Along the place dimension, the vowel /i/ has a certain chance, say 10%, of being initially perceived as its perceptual neighbour /y/, In the recognition phase, the listener can correct this misperception, because she has learnt about confusion probabilities (§9.5). Suppose that initial misperceptions are symmetric, i.e., an intended /y/ also has a chance of 10% of being perceived initially as /i/. Thus,

$$P\left(perc = \mathrm{i} \mid prod = \mathrm{y}\right) = P\left(perc = \mathrm{y} \mid prod = \mathrm{i}\right) = 0.1 \qquad (16.36)$$

If all three high vowels are equally likely to occur in an utterance, the marginal probability of each possible intended production is

$$P(prod = \mathrm{i}) = P(prod = \mathrm{y}) = \tfrac{1}{3} \qquad (16.37)$$

Likewise, the probability that a random utterance is initially categorized as /i/ is

$$P(perc = \mathrm{i}) = \sum_{n=1}^{3} P(perc = \mathrm{i} \mid prod = x_n) \cdot P(prod = x_n) = 90\% \cdot \tfrac{1}{3} + 10\% \cdot \tfrac{1}{3} = \tfrac{1}{3} \qquad (16.38)$$

A table of all these probabilities is

prod↓ *perc*→	/i/	/y/	/u/	*P*(*prod* = *x*)
/i/	0.9	0.1	0	1/3
/y/	0.1	0.8	0.1	1/3
/u/	0	0.1	0.9	1/3
P(*perc* = *x*)	1/3	1/3	1/3	

(16.39)

In the recognition phase, the listener can try to reconstruct the speaker’s intended utterance by a search for the most likely produced utterance, given the initial perception. For this, she will have to compute the posterior probability of every possible produced utterance. For instance, if the listener initially categorizes the utterance, as /i/, the probability that the speaker actually intended /y/ is, with Bayes,

$$P(prod = \mathrm{y} \mid perc = \mathrm{i}) = \frac{P(perc = \mathrm{i} \mid prod = \mathrm{y}) \cdot P(prod = \mathrm{y})}{P(perc = \mathrm{i})} \qquad (16.40)$$

All the posterior probabilities are given by

perc↓ *prod*→	/i/	/y/	/u/
/i/	0.9	0.1	0
/y/	0.1	0.8	0.1
/u/	0	0.1	0.9

(16.41)

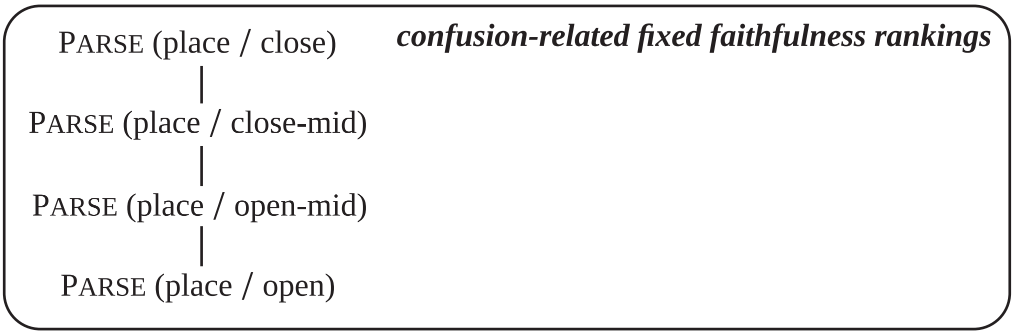
Thus, the probability that a perceived /i/ should be recognized as /y/ is equal to the probability that a perceived /y/ should be recognized as /i/. If we assume that a trained listener is capable of using these numbers in finding the most likely intended utterance (perhaps as a result of the learning algorithm described in chapter 15), we can conclude that it is equally bad for a speaker to pronounce an intended /i/ as [y], as it is for her to pronounce an intended /y/ as [i]: in both cases, the recognition problems for the listener are equally large. Now, because the speaker is also a listener, she can be supposed to “know” this. In a functionally-oriented constraint grammar, this means that the constraints *REPLACE (place: front, central) and *REPLACE (place: central, front) are ranked equally high, or, somewhat loosely, that PARSE (place: *x*) is ranked equally high for all three place values *x*.

The situation changes if we include the mid vowels in our story. Like the high vowel /i/, the mid vowel /e/ has a certain chance of being perceived as its central counterpart /ø/. But the range of F_2 values decreases as vowels become lower, as illustrated in (16.1) by the trapezoidal shape of the vowel space. Thus, the confusion probability of /e/ and /ø/ is higher than that of /i/ and /y/, say 20%. The listener has to base her recognition strategy on the following posterior probabilities:

perc↓ *prod*→	/e/	/ø/	/o/
/e/	0.79	0.20	0.01
/ø/	0.20	0.60	0.20
/o/	0.01	0.20	0.79

(16.42)

So, under a recognition strategy that maximizes the likelihood of the intended utterance, the chance that the listener successfully corrects a perceived /ø/ into the intended /e/, is larger than the chance that she corrects a perceived /y/ into the intended /i/. This means that a speaker, who knwos this because she is also a listener, can more easily get away with mispronouncing an /e/ as /ø/ than with mispronouncing an /i/ as /y/. Thus, the constraint *REPLACE (place: front, central / close) must outrank *REPLACE (place: front, central / close-mid). Simplifying this with PARSE constraints, we get the following local rankings in the (non-numerical) production grammar:



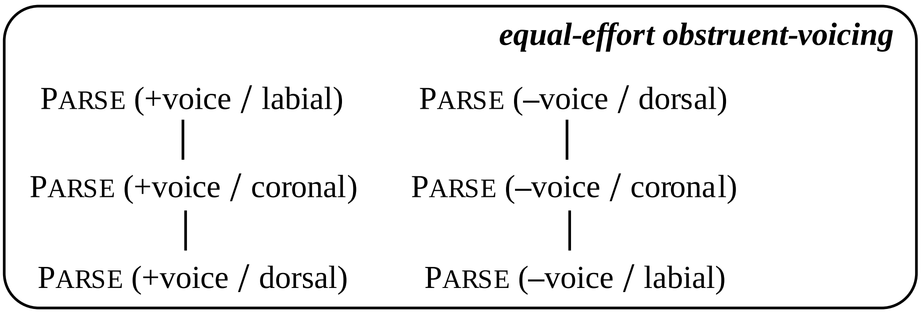
confusion-related fixed faithfulness rankings

PARSE (place / close)
PARSE (place / close-mid)
PARSE (place / open-mid)
PARSE (place / open)

(16.43)

This explains the fact that Frisian shows fewer place contrasts for lower than for higher vowels, but it does not yet explain where the gaps should be.

In our obstruent-voicing example, it will be clear where the gaps are. If the effort that the speaker wants to spend (instead of the perceptual contrast as in §16.3.1) is taken equal for all three places, the voicing contrast between [g] and [k] will be smaller than that between [d] and [t]. This leads to the following natural constraint ranking:



equal-effort obstruent-voicing

PARSE (+voice / labial)
PARSE (+voice / coronal)
PARSE (+voice / dorsal)

PARSE (–voice / dorsal)
PARSE (–voice / coronal)
PARSE (–voice / labial)

(16.44)

So, keeping the articulatory effort constant, we would have a homogeneous *GESTURE constraint and could get the following constraint interaction:

/d/	PARSE (+voice / coronal)	*GESTURE	PARSE (+voice / dorsal)
☞ [d]		*	
[t]	*!		

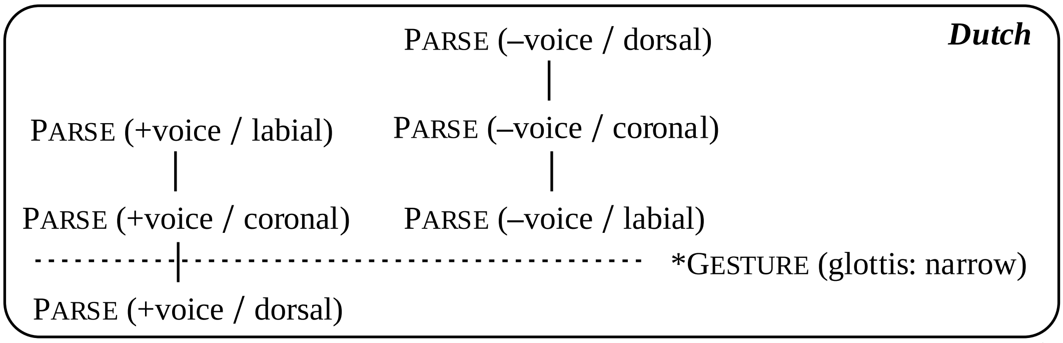
(16.45)

Thus, /d/ is parsed faithfully. The dorsal plosive, however, is devoiced:

/g/	PARSE (+voice / coronal)	*GESTURE	PARSE (+voice / dorsal)
[g]		*!	
☞ [k]			*

(16.46)

The Dutch system could be described as (cf. (16.31)):



(16.47)

16.5 Central gaps

The local rankings of §16.4 explained why languages tend to have gaps at articulatorily and/or perceptually peripheral locations: the articulatory effort often increases monotonically as we approach a more extreme articulation, and the perceptual contrast often decreases monotonically as a function of another dimension. We will now consider three proposals for the central location of the gap in the Frisian lower-mid vowel system, which has [ɛ] and [ɔ] but lacks [œ].

16.5.1 An articulatory explanation

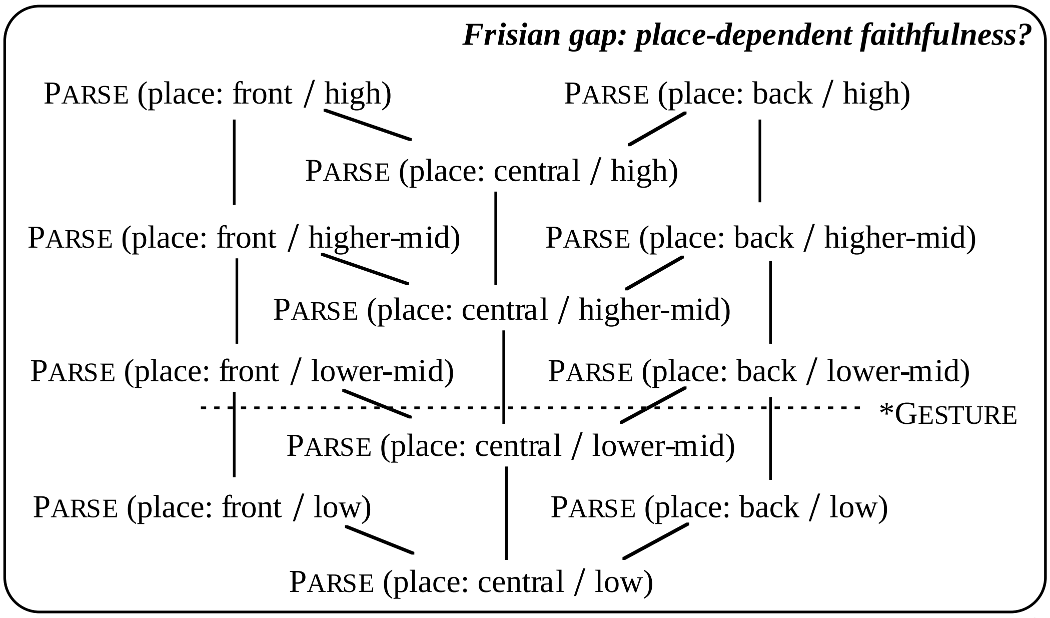
The distaste for [œ] could be explained if the effort needed for the rounding gesture is greater than that for the spreading gesture of [ɛ], so that we have the ranking *GESTURE (lips: rounded) >> *GESTURE (lips: spread). There are two problems with this approach.

First, this is not a local ranking, because different articulators are involved in lip spreading and rounding. Secondly, the ranking of the PARSE constraints of (16.43) does not depend on place, so any ranking of the two gestural constraints would treat [ɔ] in the same way as [œ]: either these two sounds are both licensed, or they are both forbidden. The same goes for the fronting gesture of [œ]: if the relevant gestural constraint is ranked higher than PARSE (place / lower-mid), [ɛ] and [œ] are both forbidden; otherwise, they are both allowed. There is no way to derive the correct system with a place-independent PARSE (place).

16.5.2 A contrast-based explanation

If we make PARSE (place) dependent on place, we may be able to account for the Frisian gap. Grammar (16.48) accurately represents the Frisian vowel system. I will now show how the listener's quest for an optimal recognition strategy can give rise to asymmetries in PARSE rankings along a single dimension.

The three place values are not equally well suited for use in a language. Table (16.39) showed that central values along a perceptual dimension give rise to twice as many confusions as peripheral values. In the history of a language, this could give rise to a pressure towards choosing peripheral values in the process of lexical selection. In our Frisian example, this would take the average confusion between high vowels down from 13.33% in the direction of 10%. For instance, a vocabulary with 40% /i/, 20% /y/, and 40% /u/, would reduce the average confusion probability to 12%, almost half-way the minimum. This lexical shift would reduce the information content per vowel, but not by much: from 1.58 to 1.52 bit. Getting rid of /y/ altogether would reduce the confusion probability to 10%, or 5% after recognition, or 0% after suspension of the central category, but it would also reduce the information content to 1 bit per vowel; this would require a much longer utterance for the same information, violating heavily one of Passy's (1891) functional principles.


Frisian gap: place-dependent faithfulness?
PARSE (place: front / high)
PARSE (place: back / high)
PARSE (place: central / high)
PARSE (place: front / higher-mid)
PARSE (place: back / higher-mid)
PARSE (place: central / higher-mid)
PARSE (place: front / lower-mid)
PARSE (place: back / lower-mid)
*GESTURE
PARSE (place: central / lower-mid)
PARSE (place: front / low)
PARSE (place: back / low)
PARSE (place: central / low)

(16.48)

In the recognition strategy, the skewed distribution of the place values leads to a shift of the /i/-/y/ discrimination criterion along the continuous F_2 axis in the direction of the centre of the distribution of the /y/ productions (fig. 8.2). If the production distributions are Gaussian, this narrowing of the /y/ category will cause an asymmetry to arise in the confusion probabilities. For instance, the chance that an intended /i/ is categorized as /y/ is 7%, and the chance of an /y/ being categorized as /i/ is 14% (this commonness-related asymmetry is the explanation for the fact that an English intended /θ/ has a larger chance of being perceived as /f/ than the reverse). The perception probabilities of our Frisian example become

prod↓ *perc*→	/i/	/y/	/u/	$P(prod = x)$
/i/	0.93	0.07	0	0.4
/y/	0.14	0.72	0.14	0.2
/u/	0	0.07	0.93	0.4
$P(perc = x)$	0.4	0.2	0.4	

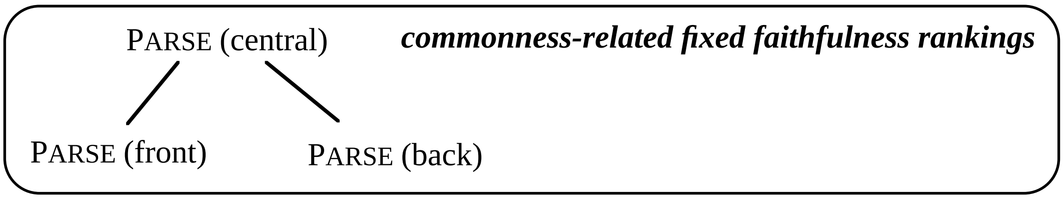
(16.49)

The posterior probabilities of a certain production given a perceived value are

perc↓ *prod*→	/i/	/y/	/u/
/i/	0.93	0.07	0
/y/	0.14	0.72	0.14
/u/	0	0.07	0.93

(16.50)

Thus, an initially perceived /y/ suggests an /i/ recognition candidate more strongly than the reverse. Therefore, it is less bad for recognition to perceive a spurious /y/ than to perceive a spurious /i/. Therefore, it is less bad for the speaker to pronounce an /i/ as [y] than to pronounce an /y/ as [i]. This gives the local ranking *REPLACE (place: front, central) >> *REPLACE (place: central, front), or, more loosely,



(16.51)

The general empirical prediction from this kind of rankings is that less common perceptual feature values have stronger specifications. For instance, if rounded vowels are less common than unrounded vowels, /i+o/ will have more chance of being assimilated to [yo] than /ye/ to [ie]; if coronals are more common than labials, it is more likely that /n+p/ becomes [mp] than that /m+t/ becomes [nt]; and if nasals are less common than non-nasals, /p+n/ will become [mn] more easily than /m+t/ will become [pt]; no theories of underspecification or privative features are needed to explain these three cross-linguistically well-attested asymmetries.

While the rankings in (54) exhibit a desirable property of phonological processes, they are the reverse of what would be needed to explain the Frisian gap. This becomes dramatically clear when we compare (16.51) with (16.48)...

16.5.3 A confusion-based explanation

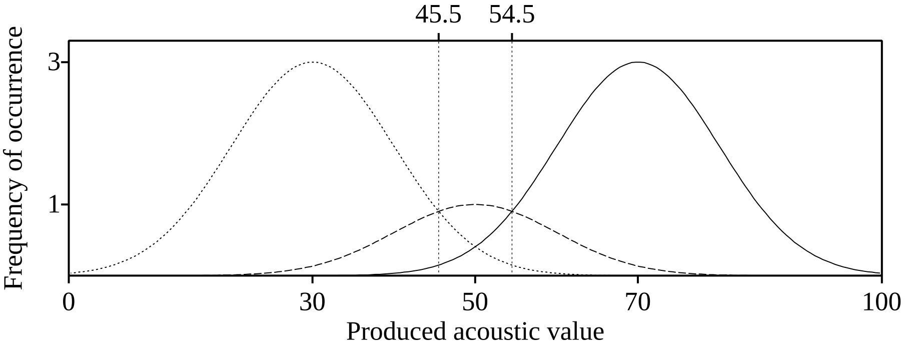
After discarding two other explanations, we are still left with a plausible explanation for the Frisian gap: the F_2 space for lower-mid vowels is considered too small to easily maintain a three-way contrast. Fewer confusions will arise if the language has an /ɛ/-/ɔ/ contrast than if it has an /œ/-/ɔ/ contrast.

In a production grammar, we could try to describe such a thing by a positive REPLACE (œ, ɛ) constraint (without the asterisk). However, this would effectively introduce an extra level in the phonology! The family *REPLACE (*x*, *y*), though formulated as a two-level constraint (a relation between input and output), can actually be seen as an output-only (i.e., one-level) constraint that says "the output should contain no *y* here". No such move would be possible with a structure-changing positive REPLACE.

Instead of accepting such anti-faithfulness constraints, we should note that the problem of the three-way contrast is in the perception grammar: because of the variation in production and perception, correct categorization is difficult, and not relying on noisy categories will make a better recognition strategy than relying on them. Now suppose that a language has a problematic three-way contrast. The following steps may happen.

Step 1. The middle category gets weaker, i.e. loses some of its lexical occurrences, as described above in §16.4.2. Variations within and between speakers will lead to random distributions of the acoustic input to the listener's ear. If the speakers implement three categories with midpoints at [30], [50], and [70] along a perceptual dimension with values from [0] to [100], the inputs to the listener's perception grammar are distributed as follows:

Production distributions of the three categories /30/, /50/, and /70/.



(16.52)

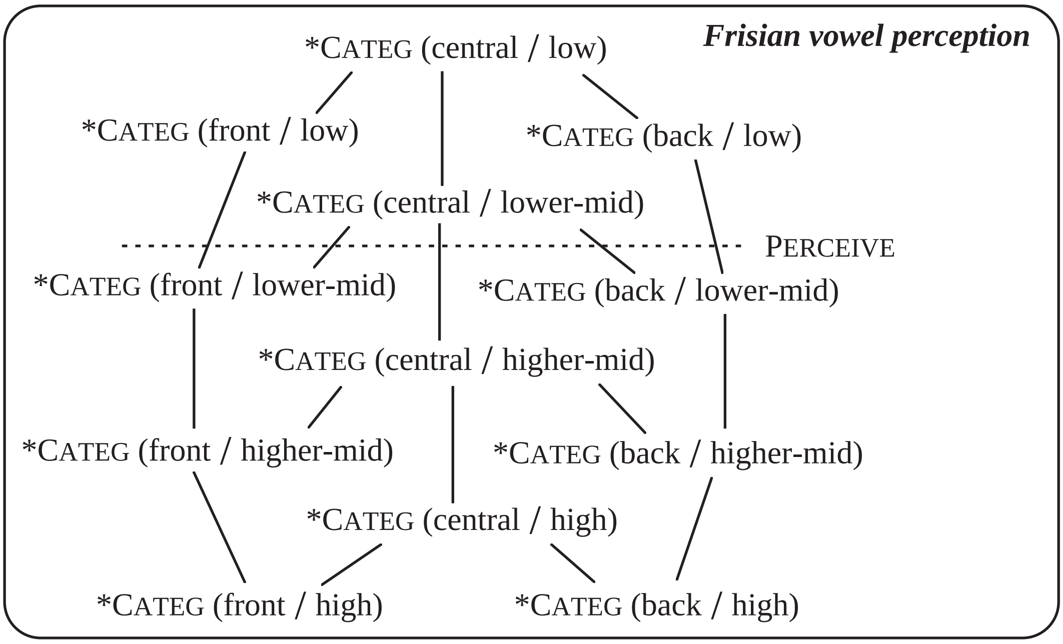
Step 2. The listener will make the fewest mistakes in initial categorization if she uses the criterion of maximum likelihood, i.e., if she chooses the category that maximizes the a posteriori probability (16.40). For instance, if the acoustic input is [44], an optimal listener will choose the /30/ category because the curve of the distribution of the production of /30/ in figure (16.52) is above the curve associated with the production of the category /50/, although the value [44] is nearer to the midpoint of the /50/ category than to the midpoint of /30/. Therefore, she will initially categorize all inputs below the criterion [45.5] into the class /30/, all the values between [45.5] and the second criterion [54.5] into the class /50/, and all values above [54.5] into the class /70/. In chapter 15, I showed how an OT listener manages to establish these criteria as a result of an automatic gradual learning process (though she will not actually become a maximum-likelihood listener).

Step 3. If the adjacent categories are close to each other, the criterion shifts can be described as a raising of the *CATEG (central / lower-mid) constraint (fig. 8.4).

Step 4. As the category gets narrower, more utterances of the middle class will be perceived into the neighbouring, broader, classes. Figure (16.52), for instance, shows that an intended /50/ is perceived as /70/ approximately four times as often as an intended /70/ is perceived as /50/.

Step 5. This will lead to the middle category getting still weaker, i.e., because of the large amount of misperception, the learner will lexicalize many adult /50/ as /70/.

These five steps form a system with positive feedback. Unless checked by requirements of information flow, the process will not stop until all the occurrences of the middle category have vanished, and a newly categorized feature is born. This situation can be described as


Frisian vowel perception
*CATEG (central / low)
*CATEG (front / low)
*CATEG (back / low)
*CATEG (central / lower-mid)
PERCEIVE
*CATEG (front / lower-mid)
*CATEG (back / lower-mid)
*CATEG (central / higher-mid)
*CATEG (front / higher-mid)
*CATEG (back / higher-mid)
*CATEG (central / high)
*CATEG (front / high)
*CATEG (back / high)

(16.53)

The ranking of the four constraints above PERCEIVE means that lower mid vowels cannot be categorized as central, and that the low vowel is not categorized at all along the F_2 dimension. Thus, the gaps in inventory (16.1) are the result of limitations of categorization, and no constraints against [œ] have to be present in the production grammar, since a hypothetical underlying |œ| could never surface faithfully: even if an underlying |œ| is pronounced as [œ], it will be perceived by the speaker herself as /ɛ/ or

/ɔ/; see figure (6.1) for the role of the perception grammar in the evaluation of faithfulness.

If we vary the ranking of PERCEIVE with respect to the *CATEG (back) family, we see that the following four systems are possible with three heights for non-low vowels:

i	u	i	y	u	i	y	u	i	y	u
e	o	e		o	e	ø	o	e	ø	o
ɛ	ɔ	ɛ		ɔ	ɛ		ɔ	ɛ	œ	ɔ

(16.54)

Precisely these four systems are fairly common: apparently, grammars are allowed a considerable degree of freedom in ranking the PERCEIVE constraints relative to the *CATEG constraints, but no freedom at all to reverse the universal ranking within the *CATEG (back) family; this gives strong evidence for the local-ranking hypothesis. The first system is more common than the fourth: global (cross-dimensional) contrast measures may predict which of these systems are the most common ones, but cannot preclude any of them beforehand; the local-ranking principle ensures that.

Finally, note that our theory not only tells us *which* sounds there are in an inventory, but also how many, given the number of low *GESTURE and *CATEG constraints; in previous 'phonetic' accounts (§16.1.5-6), this number used to be *posited*.

16.6 Conclusion

This chapter showed that a combination of functional principles, interacting in the production and perception grammars under the regime of Optimality Theory, allows accurate explanation of the symmetries and gaps in vowel and consonant systems.

Symmetry results because the listener interprets a finite number of categories along each of the language's perceptual dimensions, and because the speaker implements a finite number of articulatory tricks and their combinations. Note that this does not mean that the listener *cannot hear* other feature values or *cannot perform* other gestures: it is the constraint balance that determines what categorization constraints and gestural constraints are weak enough to let their categories and gestures come to the surface in the actual situation of language use.

Gaps are explained by local rankings of functional constraints:

(a) Local rankings of *GESTURE explain articulatorily peripheral gaps.
(b) Local rankings of PARSE explain perceptually peripheral gaps.
(c) Local rankings of *CATEG explain perceptually central gaps.

The global optimization criterion of maximal dispersion is a derivative of these local phenomena.

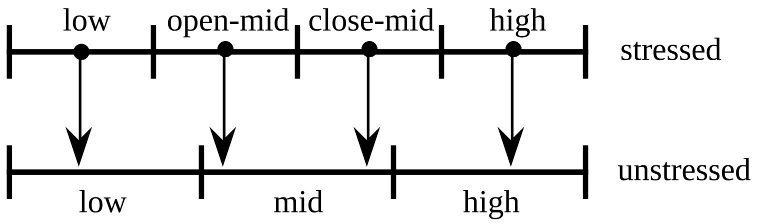
To explain inventories, we need assume no innate features, feature values, or constraints.

Postscript: the role of categorization in the production grammar

Superficially, only gestural and faithfulness constraints seem to play a role in the production grammar; categorization constraints are only made explicit in the perception grammar. This would suggest that I invoked the categorization constraints only in order to account for inventory structure, so I would be open to one of the same criticisms that I voiced on Flemming's MINDIST and MAINTAIN-N-CONTRASTS proposals in §16.1.8.

However, I will show that the categorization constraints actually play an essential role in the production grammar. To see this, we go back to our Italian mid-vowel-merger rule. In stressed position, we have a seven-vowel system, in unstressed position a five-vowel system. This reduced unstressed-vowel system can be caused by a dependence of *CATEG on stress. After all, Italian unstressed vowels are much shorter and less loud than stressed vowels, two properties that make them less resistant against background noise and cause confusion probabilities to be much greater for unstressed than for stressed vowels (with the same distance along the F_1 axis). With the usual dependence of the ranking of *CATEG on confusion probabilities, this may lead to different categorizations of the height dimension for stressed and unstressed vowels. While the four categories for stressed vowels were "low", "lower-mid", "higher-mid", and "high", we are left with three categories for unstressed vowels, which we will call "low", "mid", and high.

Now consider the categorization of the four "stressed" vowel qualities in unstressed position. If the F_1 space is equally large for stressed and unstressed vowels and the categories are equally wide, the F_1 space will be categorized as:



(16.55)

As we see in the figure (suggested by the re-categorization of the midpoints of the "stressed" categories), all of the "low" values map to "low" in the "unstressed" perception grammar, all of "high" maps to "high", and *most* of the open-mid and close-mid realizations will be categorized as "mid". So we see that the names of the three "unstressed" categories are appropriate and that Italian follows the default strategy of the merger of the two central categories: if the speaker pronounces unstressed [a ɛ e i], the listener will initially categorize this as /a "e" "e" i/, where /"e"/ is a vowel halfway between /a/ and /i/. In finding the underlying form, the listener will have to reconstruct |ɛ| or |e| as appropriate, with the help of the biases of lexical access, syntax, and meaning.

We can now see how the production grammar causes an underlying |ɛ| to be pronounced as [e] in unstressed position:

Input: |ɛ| (unstressed as in |spɛnd+iˈamo|)	*REPLACE (open-mid, high)	*REPLACE (open-mid, low)	*REPLACE (open-mid, mid)	*GESTURE (jaw: open)	*GESTURE (jaw: half open)
[a] /a/		*!		*	
[ɛ] /“e”/			*	*!	
☞ [e] /“e”/			*		*
[i] /i/	*!				

(16.56)

The ranking of the *REPLACE (*underlying category*, *surface category*) constraints depends on the distance between the midpoints of the underlying and surface categories; as we see in (16.55), the distance from the “open-mid” underlying category is smallest for the “mid” surface category (1/8 of the scale, as opposed to 5/24 for the “low” surface category). This local ranking of *REPLACE invalidates the candidates [a] and [i][5]. The two remaining candidates [ɛ] and [e] are both perceived as /“e”/, so they both violate the same faithfulness constraint. The buck is passed to the gestural constraints, specifically, to any small differences in jaw-opening effort. As most of the surrounding consonants are usually pronounced with a rather closed jaw, this effort will be larger for [ɛ] than for [e], giving a local *GESTURE ranking that causes the easier [e] candidate to win.

[5] Actually, /a/ and /“e”/ are on opposing sides of the specified value, so local rankability can be questioned: replacement of /ɛ/ with /a/ would also be thinkable (in another language), though Italian follows the globally determined default.

17 Sound change[1]

Abstract. Sound systems may never stop changing, not even if only internal factors are present, because there may always be a better system. Non-teleological random variation of constraint ranking defines a pressure that explains the existence of perpetually rotating cycles of sound changes.

In chapter 16, I showed that the ***symmetries*** in inventories of sounds can be described and explained by general properties of motor learning and perceptual categorization, and that the ***gaps*** in these inventories can be described and explained by asymmetries in articulatory effort or perceptual confusion probabilities. Intimately related with the problem of inventories is the problem of sound change. After all, the inventories have been created in a long series of sound changes, and if inventories seem constructed according to functional principles, these same functional principles must be the driving forces for many sound changes as well.

However, speakers cannot be expected to be able to see ahead to the state that their language will be in after the change: their goal is not to improve the language, but to make themselves understood as quickly, clearly, and easily as possible (Passy 1891). Thus, we have three levels of looking at sound change:

(1) **The grammar**. In Functional Phonology, the production grammar (i.e., the system that determines the shape of the utterance, given an underlying form) contains constraints that express the drives of maximizing articulatory ease and minimizing perceptual confusion either directly (by disfavouring gestures and favouring faithfulness) or indirectly (by their relative rankings). Strong evidence for the presence of these principles in the production grammar, which handles discrete phonology as well as phonetic implementation, is found in pragmatically based reranking of faithfulness: people are capable of speaking more clearly if they are required to do so. At this level, therefore, goal-oriented processes play a role; in Passy's (1891: 229) words, "one speaks in order to be understood". For instance, if a language has a voicing contrast in /b/ and /p/, and underlying |b| and |p| are usually pronounced as [b] and [p], a speaker may enhance the contrast by implementing |p| as [p^h], [p'], or [pː], thus reducing the probability that the listener will perceive the intended |p| as /b/, or by implementing |b| as [β], [ɓ], or [^{m}b].

(2) **The change**. If many speakers let a certain constraint prevail, e.g., if many speakers implement |p| as [p^h], new learners will include the aspiration in their specifications, thus creating a new underlying segment (= bundle of perceptual features) |p^h|. This change is automatic; it was not a goal of anyone at any time. Though change, therefore, is not ***teleological*** (there is no ***final causation***), it is ***functional*** in the sense that it is the result of local optimization in the production grammar.

[1] This chapter appeared on Rutgers Optimality Archive as Boersma (1997d).

(3) **The inventory**. As a result of the change, the inventory has improved: a voicing contrast has changed into a voicing-and-aspiration contrast, reducing the average number of confusions. At this level, we can talk in teleological terms again, if only we know what we are doing. This is completely analogous to the common use of teleological jargon in discussions on biological evolution ("why giraffes have long necks"), where everyone realizes that what appears to be a historical gene change is the automatic result of the survival of the fittest (those with the longest necks) in the struggle for life (Darwin 1859), not the result of any goal. However, we saw above that in contrast with genetic change, whose ultimate sources are random mutations, the ultimate sources of (several types of) sound change are directly related to communicative principles in phonetic implementation. In order not to confuse the concrete (1) and abstract (3) uses of teleological terminology, we should refrain from describing change as goal-oriented at the inventory level.

Thus we expect the following types of changes to be frequent:

Shifts with conservation of symmetry. The clearest examples can be read off from small regional variations. Dutch /eː/ is diphthongized to /ei/ in exactly the same regions where /oː/ is diphthongized to /ou/. This may reflect pure diachronic autosegmental behaviour of height contours (like we are used to in the case of tone contours), or result from a quick restoration of symmetry after an initial small imbalance. This restoration may take place as follows. If /eː/ slightly diphthongizes but /oː/ remains the same, listeners have to distinguish two very similar F_1 contours. Quite probably, learners will put these two contours in the same perceptual category and subsequently see no reason to distinguish them in their own productions; the extra F_1 contour has been temporary.

Filling of gaps. If (as the result of a blind sound change) an unnatural gap emerges in a system, subsequent sound changes or lexical selections will hurry to fill that gap. For instance, Latin inherited from Proto-Indo-European the stop system /p t k b d g/ with an unnaturally skewed distribution of voiced stops: though the implementation of voicing would be easiest for the labial plosive, only about 1.2% of all Latin words started with /b/, whereas /d/ accounted for 6% (without the de- and di(s)- words: 2%), and /g/ for 1.5% (as a page count in several dictionaries teaches us). In French, these numbers have become 5%, 6% (2%), and 3%.

A weak interpretation of these facts is that the Proto-Indo-European gap at /b/ was in an unnatural position and that the ***local-ranking principle*** (§11) caused this gap to be a lexical ***accident*** in the learner's grammar, thereby allowing Indo-Europeans to freely borrow words with /b/ faithfully (Latin, Greek, Sanskrit) or to fill in the gap with a sound change (Greek: gw → b). A stronger interpretation of the same facts is that French borrowed /b/ to a larger extent than the other voiced plosives; this active de-skewing would presumably involve phonologically-determined choices between synonyms in the lexicon. Though I believe that these choices can be made in the production grammar ("choose the best candidate"), a proof of the controversial factuality of this procedure would require a large empirical investigation.

Whether ***natural*** gaps, like the lacking /g/ in { p t k b d }, can also be filled, depends on the relative importance of the various factors involved, i.e. it depends on the rankings of the faithfulness, gestural, and categorization constraints (chapter 16).

Emergence of new gaps. If a system obtains a phoneme at a location where it would be natural to have a gap, subsequent sound changes may create such a gap. Many of the defective stop systems /p t k b d/ used to have a /g/. A “passive” explanation would be that a learner does not hear the difference between /g/ and /k/ as well as the differences in the other pairs, and merges the two. An “active” explanation would be that speakers selectively modify their /g/ so that it becomes perceptually more distinct from /k/. In §17.1.2, I will show that these active modifications are actually used.

The main idea to be learned from this small typology of functionally explainable sound changes, is that symmetrizing principles (“I have learnt a finite number of types of articulatory gestures and perceptual categories”) are just as “functional” as those depending on the biases of the human speech and hearing apparatuses (“minimize articulatory effort and perceptual confusion”). The functional tradition (Passy 1891, Martinet 1955) has always recognized that these principles conflict with each other and that every sound system shows a balance that is the result of a struggle between these principles.

The important question, however, has always been: can and should these principles be expressed directly in a grammar of the language? In part II, I have shown that they can be represented in a production grammar, thanks to the formal constraint-based phonology of Optimality Theory (Prince & Smolensky 1993). In part III, I argue that a phonological theory based on these principles adequately describes the data of the languages of the world without the need for positing any innate features, representations, or constraints.

After elaborating on the controversy (§17.1), I will use the remainder of this chapter to propose an answer to the irritating question:

> **Q**: “if functional principles optimize sound systems by causing sound change, why do not all languages get better, and why do languages never stop changing?”

The proposed answer will simply be: “because there will always be a better system”.

17.1 Criticisms of functionalism in sound change

Several criticisms have been directed to the unclear definitions and lack of formalizability that used to go with the idea of functionalism. These criticisms come together in Trask’s (1996) definitions of both ***maximum ease of articulation*** and ***maximum perceptual separation*** as “somewhat ill-defined principle[s] sometimes invoked to account for phonological change”. With the gestural and faithfulness constraints of Functional Phonology, however, the principles have received formal definitions that are capable not only of explaining, but also of describing sound patterns.

Beside the definitions, the concerted effects and interactions of the functional principles have also met with a poor press. Labov (1994) criticizes the simultaneous functional explanations of ***chain shift*** as an expression of the preservation of contrast and of ***parallel shift*** as an expression of rule generalization (i.e. preservation of symmetry):

> “the entire discussion will quickly become vacuous if we lump together explanations based on the facilitation of speech with those that are based on the preservation of meaning.” (Labov 1994: 551)

However, we can use the constraint-ranking approach of Optimality Theory to combine many explanations without ‘lumping’; instead, they are ***interleaved***, which makes all the difference. Several functional principles can play a role simultaneously. The existence of parallel chain shifts, by the way, proves that.

17.1.1 Ohala’s “phonetic” approach to sound change

With the definition and formalizability issues out of the way, we can turn to another criticism, directed at the idea that functional explanations invoke goal-orientedness. Ohala (1993) aggressively argues against language change involving “goals”:

> “reliance on teleological accounts of sound change is poor scientific strategy. For the same reason that the mature sciences such as physics and chemistry do not explain their phenomena (any more) by saying ‘the gods willed it’, linguists would be advised not to have the ‘speaker’s will’ as the first explanation for language change”. (Ohala 1993, p. 263).

Ohala’s own proposal involves synchronic unintended variation, ***hypo-correction***, and ***hyper-correction***. In his model, normal speech perception involves the process of ***correction***, which occurs when a listener restores a phoneme from its contextually influenced realization. For instance, in a language with no contrasting nasality for vowels, the utterance [kɑ̃n] can be reconstructed by the listener as the phoneme sequence |kɑn| that was intended by the speaker, because she knows that every vowel is nasalized before a nasal consonant. Hypo-correction occurs if she fails to restore a phoneme, perhaps because the [n] was not pronounced very clearly, and analyses the utterance as |kɑ̃|. Hyper-correction refers to the listener restoring a phoneme from the troubled environment although it was not intended by the speaker. Ohala’s example is the Latin change /kwiːŋkwe/ ‘five’ → */kiːŋkwe/. The first [w] in [kwiːŋkwe] may be interpreted by the listener as resulting from the spreading of the second, in which case it would be correct to reconstruct the word as |kiːŋkwe|.

Ohala’s theory accounts for several attested sound changes; for instance, it explains most of Kawasaki’s (1982) data of a general avoidance of /wu/ sequences. However, his anti-teleological position denies the possibility of sound changes (or lexical choices) that seem to preserve the contrast between segments, like (to stay with the */wu/ example) the avoidance of /um/ insertion into /w/-initial stems in Tagalog (Schachter & Otanes 1972). In the following section, I will discuss a case that probably does involve contrast enhancement.

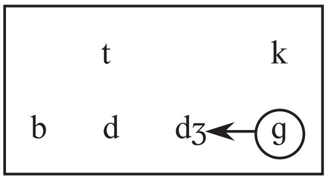
17.1.2 The story of the fugitive /g/

Ohala (1993) maintains that languages change by misunderstanding of the input, and that goal-oriented drives are never at work. However, we will show in this section that speakers do try to solve the problems that arise when two sounds are hard to distinguish. We will do this by looking at an example that Ohala himself (Ohala & Riordan 1979) has noticed as a tendency in the languages of the world.

Analogously to Ohala’s (1993) account described above, the relative voicelessness of [g], which is due to its short distance to the glottis, would result in misinterpretations of an intended /g/ as /k/. According to Ohala’s reasoning, the only thing /g/ could do, due to the small contrast with /k/, is to be misheard as /k/. If a language used to have [b]-[p] contrasts as well as [g]-[k] contrasts, and now still shows a [b]-[p] contrast but no [g]-[k] contrast, this would have to be due to a coalescence of the velar stops, in particular the conversion of /g/ into /k/.

This would give a merger of /g/ and /k/ in all cases. Surely the /g/ could not travel away from this danger zone, which would be a bad case of teleology? Nevertheless, exactly this is what real languages seem to do: most languages that lost the /g/-/k/ contrast while retaining the /b/-/p/ contrast, did so by converting their /g/ into something perceptually more distinguishable from /k/. Here are a few examples.

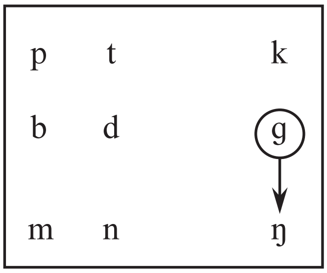
Arabic. In Arabic, an older /g/ was fronted and affricated, and became the palato-alveolar affricate /dʒ/ (Moscati, Spitaler, Ullendorff & Von Soden 1964):



(17.1)

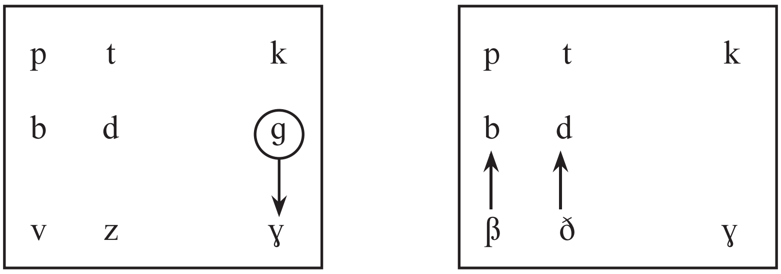
In Arabic, the /g/ problem was solved by creating a new place of articulation.

Japanese. In older Tokyo Japanese, /g/, which acts as a plosive stop throughout the phonology, is pronounced word-internally as the velar nasal stop [ŋ], except in geminates. This makes voicing of the /g/ easier by opening the velopharyngeal port:



(17.2)

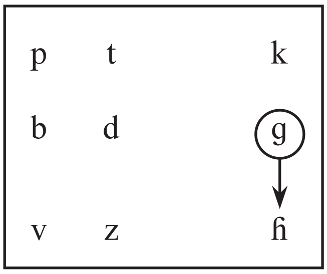
Low German. The Low German /ɣ/ may derive from /g/, which is still heard in some other Germanic languages (left figure):



(17.3)

The /ɣ/ stayed a voiced velar fricative in most Low Franconian dialects (including Limburgian), though it is /ɦ/ in West Flemish; in most Low Saxon dialects (including Westphalian), it turned voiceless (/x/); in Holland Dutch (including Afrikaans), it became a voiceless pre-uvular fricative /χ/. Many people (Streitberg 1896, Lass 1994) state that the Low German /ɣ/ is the direct descendant of a Proto-Germanic /ɣ/ (right figure above). If this is true, our same functional principle could still explain why original /β/ and /ð/ did become plosives but original /ɣ/ did not, though we would no longer have a counterexample to Ohala’s statement here. Note that nearly all Germanic dialects that retained (or developed) a plosive /g/, also have an aspirated /kʰ/.

Czech, Slovak, Ukrainian. An original /g/, which still appears in most other Slavic languages, changed into the (usually) voiced “glottal fricative” /ɦ/ (by way of /ɣ/, as still occurs in southern Russian and Belorussian):



(17.4)

Andersen (1969) also attributes this change to an enhancement of the k/g contrast, noting that the only Slavic dialects that do not show any signs of change of /g/ (Polish), can be argued to have once possessed an aspirated /kʰ/. Again, we are not absolutely sure that the /g/ is original: a late change of /ɣ/ to /g/ is attested in several southern Russian dialects, giving an alternation /sʲnʲega/ ‘snow (gen.sg.)’ – /sʲnʲex/ ‘snow (nom.sg.)’, (Avanesov 1949: 142), as well as in northern standard German, giving /taːgə/ ‘days’ – /taːx/ ‘day’. However, other than in the Germanic case, it is generally assumed that in Slavic the plosive is original.

Thus, far from merging /g/ with /k/, most languages solved the /g/ problem by replacing their /g/ with a more distinctive sound. Now, is this a bad case of teleology? No, we will see that it is an automatic result of the confusion-reducing principles of phonetic implementation, namely, maximal expression of contrastive perceptual feature values and low faithfulness requirements for non-contrastive feature values. Like the laws of sound change, the biological laws of Darwin are automatic and blind, too, but there, too, we find goal-orientedness on a higher level of abstraction.

17.1.3 Graduality

The second criticism touches the factual untenability of the requirement for Ohala’s model that sound changes cannot be phonetically gradual. This is demonstrably wrong in the case of his own nasality example: languages show all kinds of reproducible sound sequences between [ɑn] and [ɑ̃], as exemplified by the Japanese moraic nasal (Akamatsu 1997) and by all those languages that replace sequences of vowel + nasal by a nasalized diphthong with an approximant second element; here we find Portuguese ([nɐ̃ũ] ‘not’), Frisian ([mɛɪ̃skə] ‘human being’), and Alemannic ([deɪ̃χu] ‘think’; [ɑũɻ̃χə] ‘butter’).

The consistency of graduality and catastrophe is true even within a theory that presupposes universal features. For instance, if the sound /ɔ/ gradually converts into the more close /ɔ̝/, the underlying form may still specify [+open,+mid] or so, and the actual height would be determined at the level of phonetic implementation. However, a learner may reinterpret this sound as a somewhat open /ọ/, resulting in a [–open,+mid] specification. Though the actual pronunciation does not change, the grammar suffered a discrete change, and the new speakers are ready to develop some phonological processes that are appropriate for an underlying /o/. Though a functionalist standpoint could hardly accept this kind of universal feature values, the reshuffling of natural classes after gradual changes is a fact, and so are the discrete changes resulting from reinterpretation of /r/ as /ʀ/ or of /k^w/ as /p/, which can only be performed by the new language learner.

Below, we will see that small changes will often quickly become large changes, because of a positive-feedback loop: free variation wil cause lowering of faithfulness constraints, and this will increase variation again.

17.1.4 Merger

The fact that two segments often merge in the history of many languages, is sometimes forwarded as a fact contradicting the functionalist hypothesis. E.g., Labov (1994: 551) states that arguments that explain chain shifts with the need to avoid mergers or loss of distinctivity “fail to deal in an accountable way with the fact that mergers are even more common than chain shifts, and that massive mergers do take place, with a concomitant increase in homonymy”. I will show why this criticism is not justified. Apart from the fact that articulatory constraints may outrank faithfulness requirements and thus cause merger, there may also be a positive functional drive involved: merger may actually decrease the probability of perceptual confusion.

When populations of speakers who speak different but related dialects come together, the variation in the input to the listener will be high, and the chances of confusions will become high, too. This means that the listeners cannot rely any longer on all the original imported distinctions, and it may become preferable not to take into account any distinction heard between certain sounds.

We can distinguish ***discrete*** and ***gradual*** mergers.

New Sittard Limburgian. This is an example of a change in progress: a *discrete* merger, enjoying awareness by the speakers. Original Sittard Limburgian has the reflex of /ôː/ (or, with umlaut, /ø̂ː/) for West-Germanic */au/ before coronals, /h/, and /w/, and word-finally (/ʀôːt/ ‘red’, /nôːt/ ‘need’, /ʃôːn/ ‘beautiful’), while most neighbouring dialects have the older /ôa/ (/ø̂a/). Besides, it has /ɔ́u/ (with umlaut: /œ́i/) for West-Germanic /oː/ (/ɣɔ́ut/ ‘good’, /ɣʀœ́in/ ‘green’), while the other dialects continue /óː/ (/ǿː/). Both classes have invariable ***acute*** (falling) accent (Dols 1944) and remain distinct from their ***circumflex*** (high-mid level) counterparts (/nɔ̄ːt/ ‘nut’, /zɔ̄ːn/ ‘son’, /ɣɔ̄ut/ ‘gold’ in all dialects). Comparable relations exist for unrounded front vowels. Now, with the 20th-century mobility around the town of Sittard, the two populations are mixing, and listeners who hear an /ôː/ can hardly use that information to decide between the two word classes. In newer Sittard Limburgian, the contrast between the classes is given up in the production as well: both merge into /ôː/; the imported /ôa/ is considered markedly foreign, and /ɔ́u/ is judged as a distinctive ‘real’ Sittardism. Given these facts, social factors, like a cooperative effort for indistinguishability, may be involved as well.

If a listener does not take into account a possibly present unreliable distinction, she may actually lower the chances of confusion. Therefore, the typical stages in the process of a *gradual* merger of two segments are:

Stage 1: full contrast. The distinction is psychologically real, is produced in a reproducible manner, and is used for disambiguating utterances.

Stage 2: unreliable contrast. The distinction is psychologically real for people who produce the contrast, but because not everybody does so, they do not use it for disambiguating utterances. In this stage, there are many people who produce the contrast and believe that they can also hear it, though they cannot. The reality of the distinction can only be shown by methods external to the speakers in this stage, for instance by acoustical measurements or a perception test with forced choices (preferably with outsiders for subjects).

Stage 3: near merger. The distinction is not psychologically real and is not used for disambiguating utterances, though some people still produce it in a more or less reproducible manner.

Stage 4: complete merger. The distinction is not produced or perceived any longer.

Thus, **the loss of the perception of the distinction precedes the loss of its production**. Labov (1994) devotes several chapters to this phenomenon. Here are some examples of the intermediate stages of gradual merger.

San Giorgio Resian. Steenwijk (1992) reports a problem in distinguishing a rounded and an unrounded central mid vowel in a Slovene dialect in the valley of Resia. Informant ML says that there is a distinction (stage 1 or 2), her son LB denies it (stage 3 or 4). A forced-choice classification experiment with ML’s speech, with six listeners including four from outside Resia, proved that her three “e”s differed from her three “o”s along the continuum defined by the exaggerated categories “e” and “o”, which is compatible with ML being in stage 1 or 2. A forced-choice identification experiment with the same data showed that ML had trouble distinguishing her own utterances, which confines her to stage 2 (LB performed somewhat better; unfortunately, there are no data on his production).

Dutch /ɔ/ and /o/. In the larger part of the Netherlands, the distinction between original short /ɔ/ and /o/ is not lexical, but instead the choice of the allophones depends on the fine phonetic structure of the adjacent consonants. More than half of the speakers use both the closed variant before a homorganic cluster of nasal and plosive, and the open variant in words with no nasal or labial consonants. Because nobody knows or hears the difference, it must be stage 3 or 4.

Schouten (1981) performed a number of experiments by which he tried to prove that standard Dutch had a phonemic distinction between the short vowels /ɔ/ and /o/ for speakers from the east of the country (where the distinction occurs uncontroversially in the local Low Saxon dialects). First, Schouten stated that for every word in a long list, he could tell which of the two phonemes it contained. This psychological reality (stage 1 or 2) was confirmed in a comparison with his brother, who agreed on more than 90% of the words. Nevertheless, Schouten himself could not use the distinction for identification (apparently, stage 2), having been trained for years in *not* hearing the distinction between these sounds. Unfortunately, the experiment stopped here. Rather than doubting the psychological reality of the distinction, Schouten could have proved stage-2 behaviour with a forced-choice classification experiment (say, with exaggerated response categories [a] and [u]) involving listeners from other language backgrounds.

Thus, while mergers do reduce contrasts, they may also reduce confusion. So, while we could defend that the general functional principle of the minimization of confusion is usually implemented as the maximization of perceptual contrast, we see an example of exactly the reverse implementation here.

17.1.5 Comprehensive approach

Only a comprehensive approach will prove appropriate for the explanation of sound change. Thus, sound change is both preservation of contrast *and* merger, because both can reduce confusion; it is both reduction of confusion *and* facilitation of speech, because both are functional drives; and it is both gradual *and* catastrophic, because acoustic cues determine distinctivity while perceptual categorization determines recognition.

17.2 A procedure for function-driven sound change

If we can model sound change, we can also model sound structures by starting from a random set of sounds and letting many sound changes convert this impossible sound system into a natural one. Every sound system is the result of centuries of sound changes. A procedure for arriving at a good sound system could be:

(1) Start out with any vocabulary (e.g., existing or random), and determine its grammar.
(2) **Variation**: propose many different rankings of the constraints in this grammar. Many of these rerankings will propose new sound systems.
(3) **Selection**: choose from the pool of variation the sound system that occurs most often in this pool: a majority decision among the speakers of the language.
(4) Return to step 2.

The criterion for step 3 boils down to a majority decision between competing constraints. We could test this hypothesis by looking at all historical sound changes of any language. Some properties of the criterion are:

- **Locality**: because of the constraint-ranking approach, we do not have to measure effort and confusion in absolute terms; our only concern is whether the satisfaction of articulatory and perceptual constraints improves or not.
- **Unidirectionality**: if the criterion for step 3 is taken strongly, then if a certain sound change would “improve” the sound system and would therefore be allowed to take place, the reverse change would not be allowed to take place. This is the part of our hypothesis that makes it in principle testable and falsifiable. A weaker form of our criterion would have a probabilistic interpretation: if a reranking proposes a sound

change, this sound change would be possible; sound changes that are more often proposed than others (because they are more common in random variation) have a larger probability of occurrence.

- **Circularity**: because of the constraint-ranking approach, which cannot assign absolute quality measures to sound systems, it is possible that there does not exist any optimum system and that sound systems keep changing forever. It might be true that the larger the sound system, the smaller the chance that it will ever settle down as an optimal system. In this way, a sequence of several "system-improving" sound changes may eventually result in the re-emergence of the original sound system. A historical example of this circularity is shown in the next sections.

Note that our procedure is a mechanical process without goal orientation, and teleological only at a higher level of abstraction, just like Darwin's survival of the fittest.

17.3 Changes in an obstruent system

Our example will be a small language with only nine utterances, taken from the "universal" set of voiced, voiceless, and aspirated plosives and voiced and voiceless fricatives articulated with the labial, coronal, and dorsal articulators, followed by [a]. Thus, the utterances of our sample language are drawn from the following set:

ba	pa	p^ha	fa	va	
da	ta	t^ha	θa	ða	
ɡa	ka	k^ha	xa	ɣa	(17.5)

I shall further assume that the system is symmetric (no more than three manners are used at a time) and stays symmetric across sound changes, and that there are only changes in "manner", not in place. So, if /p/ changes to /p^h/, /t/ and /k/ will change to /t^h/ and /k^h/.

17.3.1 Hierarchy of articulatory effort

The first hierarchy that we consider, is that of articulatory effort.

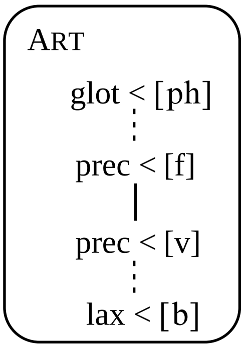
If a complete closing and opening gesture of the tongue tip is superposed on [a] and the larynx muscles are not adjusted, the vocal folds will stop vibrating soon. This is due to the increase in the oral pressure, which causes a rising intraglottal pressure which pushes the vocal folds apart. The resulting sound is a lenis voiceless stop [d̥]. To pronounce a fully voiced [d], the speaker will have to adjust the width of her glottis and may take recourse to several aiding tricks, such as a lowering larynx or slack supraglottal walls. Thus, if we compare /ta/ with /da/, we see that if /ta/ is allowed to be implemented with a passive larynx, it involves one gesture less than /da/ does (though many languages will enhance the contrast with /da/ by means of such tricks as an active widening of the glottis, a raising larynx, or stiff supraglottal walls). That constitutes the first part of the articulatory effort hierarchy: [ta] is easier to pronounce than [da]. This is expressed by the implementation constraint "any laxness gesture should be less effortful than the laxness gesture associated with a typical voiced plosive”, or in short: “lax < [b]” (now taking the somewhat easier [b] as a representative of the three voiced plosives).

The remaining parts of the articulatory-effort hierarchy are the following. Plosives are easier to pronounce than fricatives. A ballistic movement is easier than a controlled movement (§7.4). So [pa] is easier to pronounce than [fa]. This is expressed with the constraint “any articulatory precision should be less than the precision associated with a typical [f]”, or in short: “prec < [f]”.

If /v/ is realized as an approximant, it is spontaneously voiced and therefore does not involve an active gesture to make it voiced, and it requires less precision than /f/ does (it could halt anywhere between one and five inches from the wall). If we allow this freedom for /v/ (in accordance with its use in, say, English or French), the comparison of [f] and [v] only involves a pair of fixed (locally) ranked precision constraints: “prec < [f]” >> “prec < [v]”.

Finally there is [ph]. It involves an abduction of the vocal folds almost to the position during respiration and is therefore more difficult before a vowel than [p], which has no glottal contours at all. We could use the constraint *GESTURE (spread glottis), or its implementation formulation *[aspiration], but sticking with the style of the other articulatory constraints, we will use “glot < [ph]”, which is short for “any amount of glottal widening should be less than the widening associated with a typical [ph]”.

If we assume that making an active glottal opening gesture is more difficult than the precision needed for a continuant, and that this is more difficult again than the implementation of obstruent voicing, we get the following ranking within the articulation (“ART”) family (solid lines denote fixed local rankings, dotted lines denote language-specific rankings):



(17.6)

With a simplifying move, I shall promote the global effort criterion to the status of a fixed ranking. Thus, I will regard the hierarchy (17.6) as fixed, though this goes against my cherished local-ranking principle (§11), which would only regard the relative ranking of the two precision constraints as fixed, and all the other rankings as language-specific. In this special case, I justify the fixed global ranking with the idea that global effort measures can predict ***tendencies*** of asymmetry, and it is these tendencies that will cause some constraints to be on top more often than others in the pool of free variation that we consider the breeding place for sound change.

The hierarchy (17.6) would prefer [p] as the optimal implementation of an underlying underspecified |labial obstruent + a|:

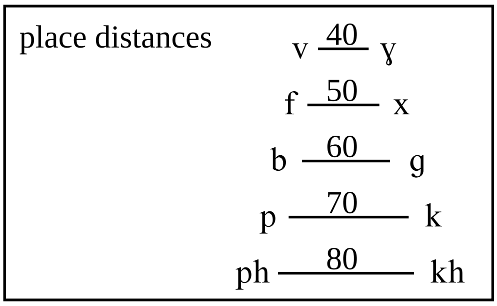
|labial obstruent + a|	glot < [ph]	prec < [fa]	prec < [v]	lax < [b]
[pha]	∗!			
☞ [pa]				
[ba]				∗!
[fa]		∗!	∗	
[va]			∗!	

(17.7)

Later, we will see that such an underspecification is actually a *weak* specification of a non-contrastive feature value. An evaluation as in (17.7) could occur in a language with a single labial obstruent.

17.3.2 Hierarchy of perceptual place distinctions

We now assume that the language, next to a series of labials, also has velar consonants with the same manner features. It is likely that voicing obscures place distinctions, so that the /p/-/k/ contrast is larger than the /b/-/g/ contrast, and the /f/-/x/ contrast is larger than the /v/-/ɣ/ contrast. If we take into consideration the commonness of changes of labial fricatives into velar fricatives (Dutch /ɑxtər/ 'behind' < /ɑftər/) and the other way around, we can tentatively propose the following place-distinction hierarchy:



(17.8)

This hierarchy is the same as the one used by Boersma (1989, 1990) to express the functional principle of maximization of *salience*, which was defined by Kawasaki (1982) as the change in perceptual features as a function of time. Thus, /pha/ has the largest salience, as the two segments /ph/ and /a/ differ in sonorance, voicing, continuancy, and noise. Likewise, /va/ has the lowest salience of the five syllables, because /v/ shares with /a/ its voicing and continuancy features. The hierarchy (17.8) is also reminiscent of the ***sonority hierarchy***.

17.3.3 Teleological inventory-oriented accounts of distinctivity

On the level of inventory evaluation, we could translate the distinctivity requirement directly into a family of inventory constraints:

Def. $*\text{N}\textsc{ear}\,(f\!: d) \equiv \exists x \in \Im \wedge \exists y \in \Im \wedge x \neq y \wedge \big(\forall g \neq f : g(x) = g(y)\big) \Rightarrow \big|f(x) - f(y)\big| > d$

“If two non-identical segments *x* and *y* in the inventory ℑ contrast along the perceptual dimension *f* only, their perceptual distance (along this dimension) is greater than *d*.” (17.9)

Its ranking within a family associated with a specific perceptual tier *f* would be fixed: the lower the distinctivity, the higher the constraint against having the corresponding contrast in the language:

$$*\text{N}\textsc{ear}\,(\textit{feature}\!: d_1) >> *\text{N}\textsc{ear}\,(\textit{feature}\!: d_2) \Leftrightarrow d_1 < d_2 \qquad (17.10)$$

In our example (17.8), the pair /f/ - /x/ violates *NEAR (place: 55) and, a fortiori, the lower ranked *NEAR (place: 65). Let’s loosely call these constraints “Δplace > 55” etc.

With evaluation of all pairs, the inventory { ph b v kh g ɣ } turns out to be more contrastive than { p b v k g ɣ }:

	Δplace > 35	Δplace > 45	Δplace > 55	Δplace > 65	Δplace > 75	Δplace > 85
p b v k g ɣ		*	*	**	***!	***
☞ ph b v kh g ɣ		*	*	**	**	***

(17.11)

In the inventory candidate { p b v k g ɣ }, the constraint “Δplace > 75” is violated three times: by the pair /v/ - /ɣ/, by the pair /b/ - /g/, and by /p/ - /k/. In the other candidate, the last violation is removed, so this inventory shows up as the winner.

This approach is equivalent to the non-OT pairwise inventory evaluation procedure in Boersma (1989, 1990). It also bears a similarity to the OT inventory evaluation approach of Flemming (1995), who, however, only evaluates the *minimal* distance in the system, not all the pairs. For instance, in Flemming’s approach, the two systems in (17.11) would be equally contrastive:

	MINDIST (place) > 35	MINDIST (place) > 45	MINDIST (place) > 55	MINDIST (place) > 75	MINDIST (place) > 85
☞ p b v k g ɣ		*	*	*	*
☞ ph b v kh g ɣ		*	*	*	*

(17.12)

Even in the inventory candidate { p b v k g ɣ }, Flemming’s constraint “MINDIST (place) > 85” is only violated once, because it evaluates only the least distinctive pair /v/ - /ɣ/. In contrast, our all-pairs approach is more in line with the intuitive idea that the need for the enhancement of the /p/ - /k/ contrast (by shifting it to /ph/ - /kh/) should be considered independent from the presence or absence of a /v/ - /ɣ/ pair.

The drawback from this technique is that it suggests that speakers are actively striving for better inventories: both (17.11) and (17.12) express teleology at the level of the inventory, a situation that we marked as undesirable in the introduction. Also, this technique suggests that optimization of inventories is a process separated from the OT production grammar. In the next section, I shall replace it with a more accurate and non-teleological formulation based on normal gestural and faithfulness constraints that evaluate the free variation in the production grammar.

17.3.4 Distinctivity in the production grammar

As phonologists, we all accept the existence of production grammars, i.e. a system that processes lexical information so that it can be pronounced. If the facts of inventories could be described by production grammars, we should not posit the existence of dedicated inventory grammars. As I have shown in chapter 16, inventories can be successfully described with the properties of articulatory, faithfulness, and categorization constraints, all of which are independently required in the production grammar. So, the inventory constraints introduced in §17.3.3 do not exist, until further evidence forces us to believe in them.

In the inventory formulation, the perceptual-distance constraints evaluated pairs of segments. In a production-grammar formulation, the functional principle of minimization of perceptual confusion has to be expressed indirectly (but locally) in terms of faithfulness constraints and their rankings. For instance, a segment specified as [labial] should be perceived as labial; this may be expressed with the faithfulness constraint PARSE (labial). If the segment contrasts with a dorsal segment, it is desirable that it is pronounced as *very* labial; the more labiality is given to the segment in the process of phonetic implementation, the lower the probability that the listener will mistake it for its dorsal counterpart.

This maximization of unary or peripheral feature values in phonetic implementation is expressed with the following constraint (§9.4):

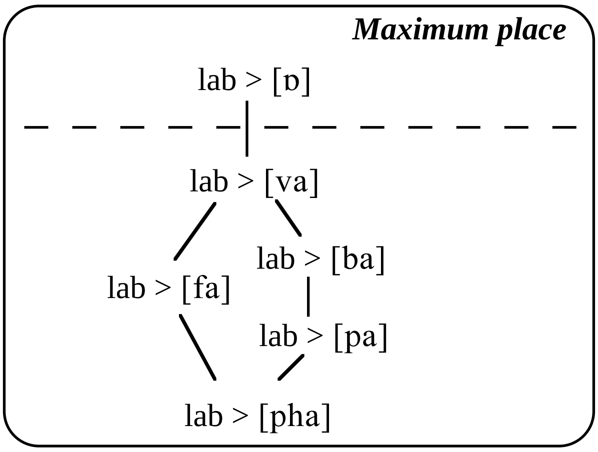
Def. MAXIMUM $(f\colon v) \equiv \exists x_i \in f_{spec} \wedge \exists y_i \in f_{ac} \Rightarrow \left(x_i = \text{“max”} \Rightarrow y_i > v\right)$

“If a feature *x* on a tier *f* is specified as “max”, the value of its acoustic correspondent *y*, if present, should be greater than any finite value *v*.”

(17.13)

Because we cannot assign numeric values to the degrees of labiality needed for the family MAXIMUM (labial), I will abbreviate these constraints in a by now familiar way, giving things like “lab > [ba]”, short for “any segment specified as labial should be pronounced with better labiality cues than those present in a typical [ba]”. The “any” in this formulation is justified because all labial obstruents in our example are considered to have coronal and dorsal counterparts.

From the distance hierarchy (17.8), we can infer that labiality cues are best in [pha] and worst in [va]. Therefore, producing labiality cues that are better than those of [pha] is much less important than producing labiality cues better than those found in [va]; this leads to a partial grammar that is parallel to the place hierarchy in (17.8). Restoring our ignorance of the relative place distinction of the voiceless fricatives with respect to the place distinctions of the non-aspirated plosives, we get



(17.14)

The location of the somewhat arbitrary constraint in the upper stratum (above the dashed line) expresses the notion of ***sufficient*** contrast: a segment specified as [labial] should get better labiality cues than the poor ones present in a typical [ɒ]. The lower stratum contains the constraints that express the notion of ***maximum*** contrast: depending on their ranking with respect to other families, they may be violated or not, as long as the fixed rankings (the drawn lines) are honoured.

We can evaluate labiality directly on an underlying/surface pair, without reference to dorsal obstruents (the ranking values of the labiality constraints, however, do depend on the presence or absence of dorsal obstruents):

|voiceless labial plosive + abava|	lab > [ɒ]	lab > [va]	lab > [b]	lab > [p]	lab > [ph]
pabava		*	**	***!	***
☞ phabava		*	**	**	***

(17.15)

Like in (17.7), the underspecification can only occur if |p| is not contrastively specified for [–noise], i.e., if |v| is specified as [–plosive] (or [+continuant]) instead of [+noise]. Our five consonants have the following perceptual features:

	|ph|	|p|	|f|	|b|	|v|
voiced	–	–	–	+	+
plosive	+	+	–	+	–
noise	+	–	+	–	+

(17.16)

Depending on their contrastive load, some of these specifications are strong and others are weak.

17.3.5 Poverty of the base

In a production-grammar formulation, we should derive sound changes with the mechanisms of richness of the base and filters, as performed for synchronic descriptions of inventories chapter 16. Let’s start with the obstruent system { p b v }.

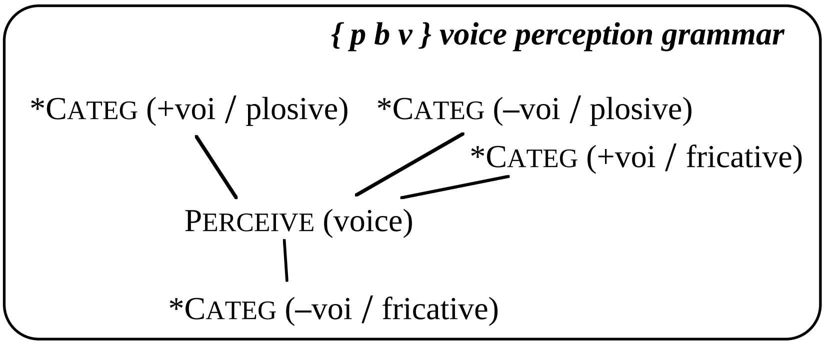
We should first identify a plausible synchronic description of the obstruent system. Apparently, this system uses the two binary-valued perceptual features [voice] and [noise]. If the perceptual categorizations along these two dimensions were independent of one another, the listener would be able to discriminate the four obstruents { p b f v }. In that case, the production grammar should prevent the faithful surfacing of an underlying |f|, presumably realizing it as [v], perceived as /v/. The grammar can achieve this by ranking PARSE (±voice) lower for fricatives than for plosives:

|pabafava|	PARSE (±voice / plosive)	prec < [f]	prec < [v]	PARSE (±voice / fricative)
[pabafava] /pabafava/		*!		
☞ [pabavava] /pabavava/				*

(17.17)

We see that any underlying |f|, supplied by richness of the base, is realized as [v], so that the surface inventory is { p b v }.

Some listeners may find it disadvantageous to maintain a perceptual voicing contrast in fricatives, only to merge the perceived /f/ and /v/ into a single underlying segment |v| in the recognition phase. These listeners will probably divide the perceptual voicing scale into two values for plosives, and into a single value for fricatives. This can be described by the following perception grammar:

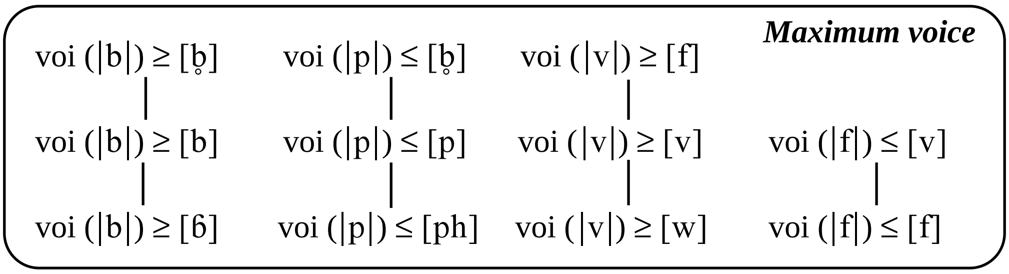

{ p b v } voice perception grammar
*CATEG (+voi / plosive)
*CATEG (–voi / plosive)
*CATEG (+voi / fricative)
PERCEIVE (voice)
*CATEG (–voi / fricative)

(17.18)

This grammar will ensure that a pronounced [f] is directly categorized as /v/. Now, the underlying form |pabafava| will always be perceived as /pabavava/, whether |f| is realized as [f] or as [v]. Unless we feel the need for a ghost segment, it is not very useful to maintain an underlying |f| - |v| contrast. In other words, the adaptation of the categorization system to the inventory { p b v } leads to an “impoverished base”, without any occurrences of underlying |f|.

17.3.6 Hierarchies of manner distinctions

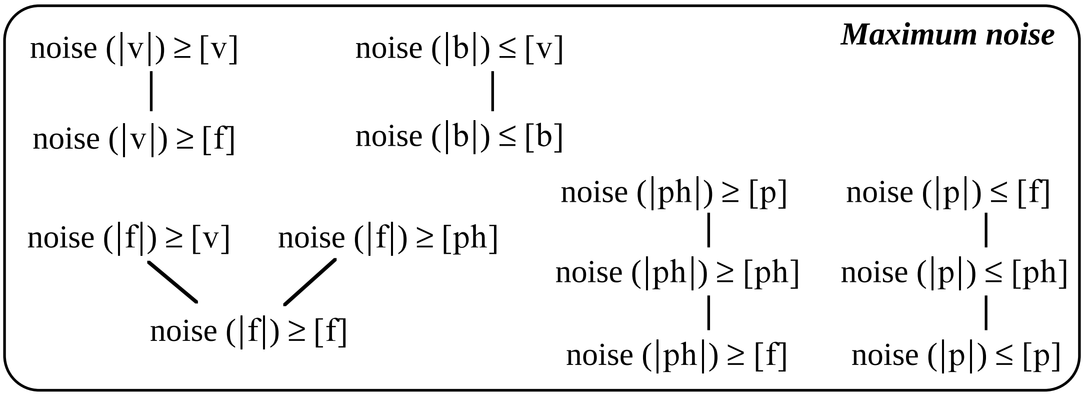
Bringing the manner contrasts into the picture is more complicated, because these cannot be expressed in a single fixed hierarchy like (17.14): if the perceptual manner space is divided up into three binary features (voice, noise, continuant), we have three separate MAXIMUM (voice), MAXIMUM (noise), and MAXIMUM (continuant) hierarchies, together with three corresponding MINIMUM hierarchies. Moreover, the rankings of these families may depend on the segment involved, i.e. on whether its specifications for these features are contrastive or not. Nevertheless, the rankings *within* these families are fixed. Here are the MAXIMUM (voice) and MINIMUM (voice) families:



(17.19)

In the "voi (|b|)" column, we see that for a plosive specified as [+voiced], it is most desirable to be at least as voiced as the voiceless lenis plosive [b̥], but that it is not so important that it is as loudly voiced as the implosive [ɓ]. Likewise, the "voi (|p|)" column expresses the increase in voicelessness as we go along the continuum [b̥] - [p] - [ph]; the "voi (|ph|)" family is similar. Finally, we see that [w] is more voiced than [v], in the sense that [w] is less likely than [v] to be incorrectly categorized as [–voiced].

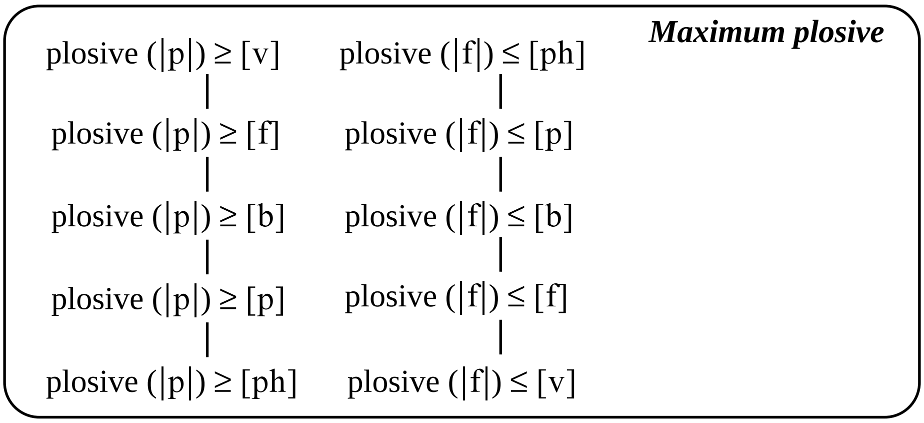
For the noise feature, the fixed hierarchies are



(17.20)

We know that [f] is more noisy than both [v] and [ph]; these relations account for five of the fixed rankings in (17.20). The "noise (|f|)" family is not totally ranked, because we cannot (and don't have to) tell whether [v] or [ph] is the more noisy of the two. Note that the constraints not connected via solid lines in (17.20) are not a priori ranked with respect to one another.

The two binary features noise and voice do not suffice for distinguishing the five obstruents. Specifically, /ph/ and /f/ have the same representation. So we need a perceptual feature like plosiveness, i.e. the degree to which the surrounding vowels are interrupted by something that resembles silence. If speakers distinguish two values for this feature, they will classify /p/, /b/, and /ph/ as [+plosive], and /f/ and /v/ as [–plosive]. This leads to the the following fixed hierarchies:



(17.21)

Apart from the subdivision between the plosives and non-plosives, these hierarchies express the idea that [ph] is the strongest and [b] the weakest plosive, and that [v] is even more non-plosive than [f].

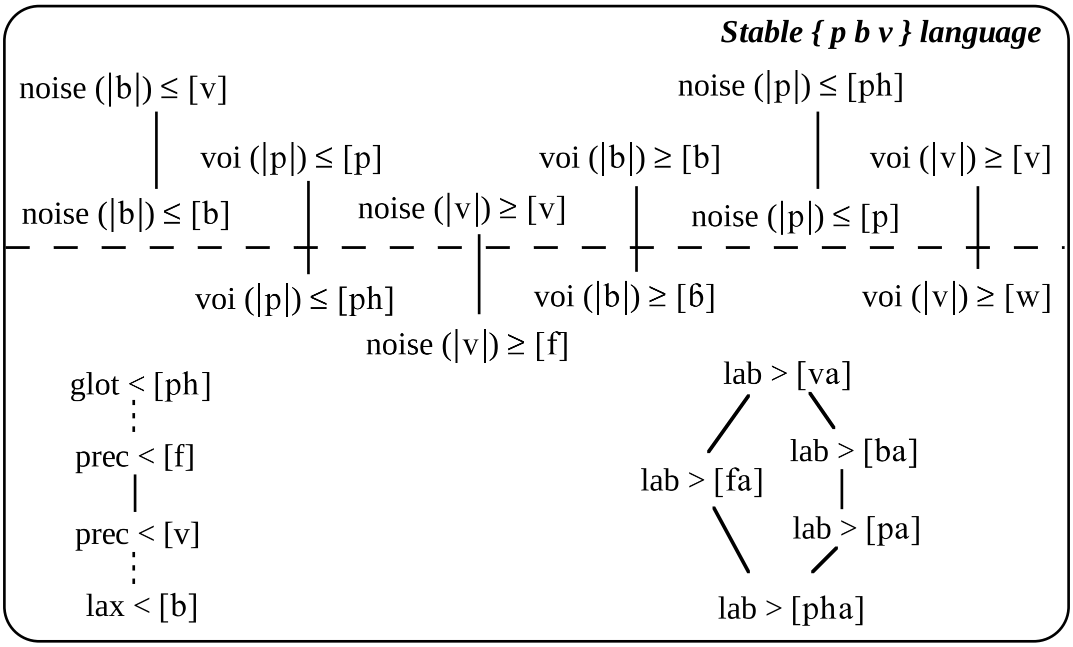
17.3.7 Stable systems

A stable { p b v } system will have the following specifications:

	p	b	v
[voice]	–	+	+
[noise]	–	–	+

(17.22)

The six specifications are translated as six families of faithfulness constraints, and the relevant member of each family is undominated:



(17.23)

Note the seemingly tautological formulation of the six constraints that are just above the dashed line: they are the expressions in phonetic-implementation language of the faithfulness constraints PARSE (+noise), PARSE (–noise), PARSE (+voice), and PARSE (–voice), or if you insist on unary perceptual features: PARSE (noise), FILL (noise), PARSE (voice), and FILL (voice). In (17.23), we see that the production grammar bundles seemingly discrete phonological constraints with continuous phonetic constraints.

In the grammars, an underlying |v| will always surface as [v], and |p| will surface as [p], independently from the relative rankings of any constraints in the lower stratum (the region below the dashed line). For instance, “noise (|v|) ≥ [f]” may favour the surfacing of |v| as [f], but the voicing-parsing constraint “voi (|v|) ≥ [v]” will prevent that, being in the upper stratum. Likewise, “lab > [pa]” would favour the rendering of |p| as [ph], were it not for the high ranking of the noise-filling constraint “noise (|p|) ≤ [p]”.

17.3.8 Direct contrastivity and free variation

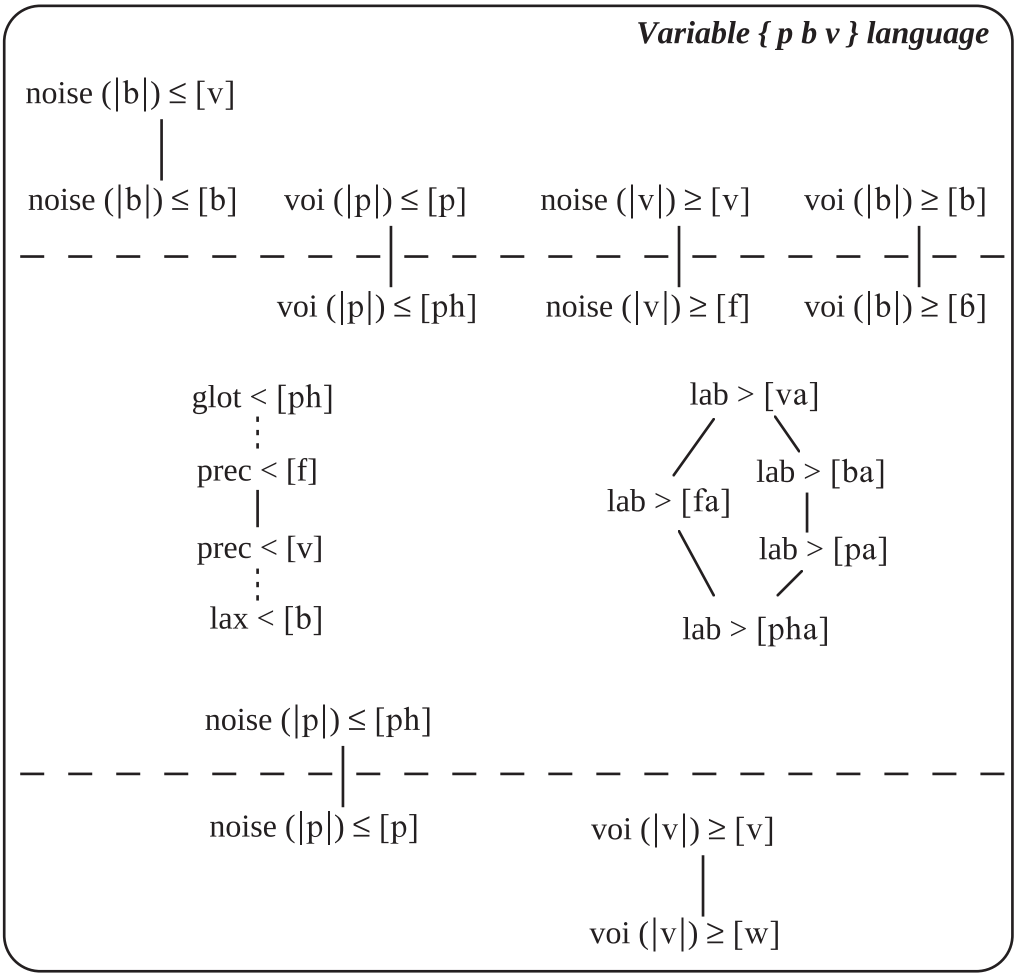
With the impoverished base that results from removing /f/ from the underlying inventory, the path is cleared for underspecified underlying representations. Contrastive specifications for { p b v } can be read from the following table:

	p	b	v
[voice]	–	+	(+)
[noise]	(–)	–	+

(17.24)

The specifications between parentheses are not directly contrastive, since /p/ has no direct [+noise] counterpart and /v/ has no direct [–voice] counterpart. Several theories connect contrastivity with phonological activity or passivity: in a derivational theory of contrastive underspecification (Steriade 1987), the [+voice] value for /v/ would be filled in at a late stage, probably after any [+voice] values would have been able to trigger assimilation; in a functional constraint-based theory, |v| would be *weakly* specified for [+voice], so that it can easily undergo devoicing, e.g. by assimilation from neighbouring segments more strongly specified for [–voice]. The functional idea behind this connection is that in a system without underlying voiceless fricatives, a |v| can be pronounced as [f] without too many problems with regard to perceptual confusion.

Let us translate the contrastivity argument into a constraint ranking. The voicing specifications for |p| and |b| are expressed as *REPLACE (voice: –, + / plosive) and *REPLACE (voice: +, – / plosive), or more loosely as PARSE (–voice / plosive) and PARSE (+voice / plosive), or more readably as the specification constraints |p| → [–voi] and |b| → [+voi] or their continuous counterparts like “voi (|p|) ≤ [p]”and “voi (|b|) ≥ [b]”. The strength of these constraints is expressed by putting them on top of the grammar. Likewise, the weakness of the specifications |v| → [+voi] and |p| → [–noi] is expressed by putting their continuous counterparts “voi (|v|) ≥ [v]” and “noise (|p|) ≤ [p]” at the bottom. The remaining constraints are ranked in an intermediate stratum:



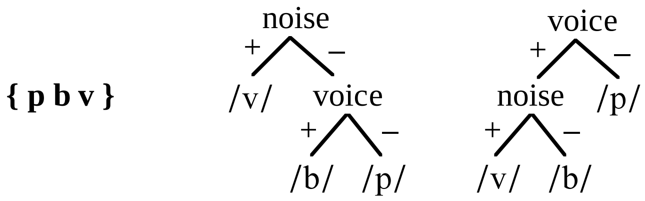
(17.25)

Note that the six “tautological” constraints of the faithful grammar (17.23) are now located either at the bottom of the highest or at the top of the lowest stratum. These six, therefore, express the principle of *sufficient contrast*; their lower-ranked fellows are in the lower strata, expressing the universally less important principle of *maximum contrast*.

If the rankings of the intermediate constraints are varied randomly, the division between the contrastive specifications at the top and the redundant specifications at the bottom will lead to a certain amount of free variation between the implementations [f] and [v] for |v|, and [p] and [ph] for |p|. Note that these alternations are exactly those that allow the speaker to be understood, i.e. those that allow the listener to reconstruct the underlying form as long as she either perceives [f] directly as /v/ or successfully reconstructs a perceived /f/ as |v| (and analogously for [ph] and |p|).

17.3.9 Indirect contrastivity and free variation

The concept of direct contrastivity in (17.24) yields an amount of underspecification too large for the purpose of determining the allowed degree of free variation: if /p/ is specified solely as [–voice] and /v/ is just [+noise], then there is no reliable contrast between these two segments. It would seem that we cannot freely vary the redundant feature values, as that would allow both /p/ and /v/ to be realized as [f]. So we should keep either the [+voice] specification of /v/ or the [–noise] specification of /p/, depending on what we consider to be the more basic feature. Both these possibilities can be represented by simple ***feature trees*** (in support of Jakobson, Cherry & Halle 1953, but contra Frisch, Broe & Pierrehumbert 1997), with segments as leaves:



(17.26)

The left-hand representation shows that /v/ can contrast with /b/ and /p/ at the same time. In such a situation, /v/ is allowed to be pronounced voiceless, and /p/ must stay noiseless. This is a reasonable description of the situation in Dutch (disregarding the somewhat marked lexically voiceless fricatives), where voiced fricatives are only weakly specified for [+voice] (they devoice after any obstruent).

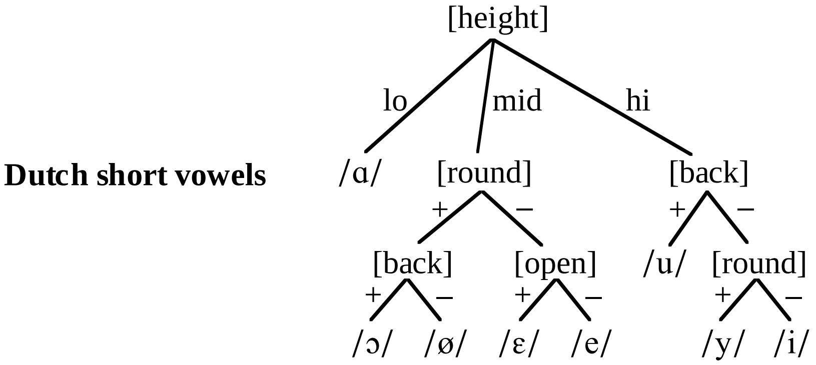
The feature-tree underspecification strategy is not only useful for predicting variation, but would also solve a technical problem (noted by Archangeli 1988) with an algorithm for contrastive underspecification that marks only those feature values whose reversal would result in a different segment in the inventory. For instance, the Dutch short-vowel system might be represented with contrastive underspecification as

	i	y	u	e	ø	ɛ	ɔ	ɑ
[open]	(–)	(–)	(–)	–	(–)	+	(+)	(+)
[mid]	–	–	(–)	+	+	(+)	(+)	(–)
[round]	–	+	(+)	–	+	(–)	(+)	(–)
[back]	(–)	–	+	(–)	(–)	(–)	(+)	(+)

(17.27)

The specifications between parentheses are not directly contrastive. The procedure leaves both /ɔ/ and /ɑ/ without any specifications at all. Surely these segments do contrast! This problem cannot be resolved by using different height features (unless /ɑ/ is changed to [+mid]; but the problem is in the algorithm, not in the data).

To solve the ambiguities in (17.27), we will have to mark some feature values as contrastive. We have to decide which is the most basic feature; for /ɔ/ and /ɑ/, this is perhaps the height distinction, but this will depend on the language. The solution can be graphically represented by feature trees, which guarantee exactly the right amount of specification. Several feature trees are possible for the Dutch data, but I will choose the one that I think most closely reflects the behaviour of the Dutch speaker/listener:



(17.28)

It may well be that this tree tells us more about the psychological realities of Dutch vowels than any feature matrix. There are seven underspecifications, most of which are reflected in regional or positional variations:

(1) The underspecifications of the rounding of /u/ and the backness of /ɛ/ and /e/ do not contribute to variation, because the gesture for non-low back unrounded vowels (styloglossus without orbicularis oris activity) does not belong to the Dutch repertory of articulations (any unrounded back /ɑ/ is implemented without styloglossus). For the high vowels, I could have exchanged the [back] and [round] nodes; this would have given an /i/ underspecified for backness.

(2) The underspecification of the backness and rounding of /ɑ/ corresponds to the high level of regional variation of this sound. Regarding the fact that the lip shape is quite different from that of the long low vowel /aː/, we may call its usual pronunciation rounded. Fronted realizations also are common. Specifying a single low vowel for no other than height features relieves us of the need of arbitrarily assigning it a [+back] or any other specification.

(3) The underspecification of the openness of /ɔ/ corresponds to the high degree of regional and positional variation of this sound. For instance, the pronunciations of Dutch *hondehok* ‘dog kennel’ are (from most to least common): [ɦondəɦɔk], [ɦondəɦok], and [ɦɔndəɦɔk]; any pronunciations with heights between [o] and [ɔ] are also perfectly normal. Underspecifying this vowel for openness allows us not to try to identify its height artificially with that of /ɛ/ or /e/.

(4) The underspecification of the openness of /ø/ corresponds to its special relation with the [œ] sound. Dutch listeners tend to hear foreign [œ] sounds, as in German [gœtɐ] ‘gods’ or French [œf] ‘egg’, as a Dutch /ø/. Even many speech scientists transcribe Dutch /ø/ as “œ”, though it is phonetically the rounded counterpart to /e/ (same F_1, same tongue-fronting gesture), so that we would expect the transcription “ʏ”, analogously to the traditional transcription “ɪ” for /e/. We may further note that /ø/ does not pattern with /e/ at all; instead, there are some words in which /ø/ alternates with /ɔ/ (/ʋɔʀɣə/ = /ʋøʀɣə/ ‘strangle’, etc.), so that the backness correlation between these sounds suggested in (17.28) is not strange at all.

We can now state the relation between contrastivity and free variation precisely:

The maximum free variation.

> “Segments are allowed to vary freely as long as the listener can easily reconstruct the underlying form. The largest allowed amount of free variation is achieved with random reranking of intermediate constraints, keeping directly or indirectly contrastive specifications fixed at the top and redundant specifications fixed at the bottom.” (17.29)

Our modification of the notion of contrast leaves us with a { p b v } system where only the specification of /v/ voicing ranks in the lowest stratum.

17.3.10 Where { p b v } will go

Suppose that a language goes from the stable { p b v } system (17.23) to the (modification of) system (17.25), which allows some variation of voicing in |v|. In a stable system, the relative rankings of the constraints in the lower stratum cannot be determined, so we may safely suppose that speakers have some invisible random variation here. When its non-contrastivity causes the voicing specification for |v| to drop below the formerly lower stratum, as in (17.25), the rankings in the emerging intermediate stratum become exposed: they determine whether an underlying |v| is pronounced as [v] or as [f].

The variation in (17.25) is subject to the local-ranking principle, which ensures that e.g. “lab > [ba]” will always be ranked above “lab > [pa]”. Constraint pairs that are not directly or transitively locally ranked, can be ranked in a speaker-specific manner. Whether an underlying |v| surfaces as [v] or [f], there are two constraints that are violated in either case: “prec < [v]” and “lab > [fa]”; and there is one constraint that is satisfied in either case: “noise (|v|) ≥ [v]”. For the voicing of an underlying |v|, there remain thus four relevant constraints: “noise (|v|) ≥ [f]”, “prec < [f]”, “lab > [va]”, and “voi (|v|) ≥ [v]”. Because of the redundancy of the voicing specification of |v|, the last of these is always ranked at the bottom, so the outcome of an underlying |pabava| will depend on the relative rankings of the remaining constraints. If “noise (|v|) ≥ [f]” is ranked on top, the noise contrast will be enhanced:

|pabava|	noise (|v|) ≥ [f]	prec < [f]	lab > [va]	voi (|v|) ≥ [v]
[pabava]	*!		*	
☞ [pabafa]		*		*

(17.30)

If “prec < [f]” is on top, the input surfaces faithfully:

|pabava|	prec < [f]	noise (|v|) ≥ [f]	lab > [va]	voi (|v|) ≥ [v]
☞ [pabava]		*	*	
[pabafa]	*!			*

(17.31)

And if “lab > [va]” is on top, the place contrast is enhanced:

|pabava|	lab > [va]	noise (|v|) ≥ [f]	prec < [f]	voi (|v|) ≥ [v]
[pabava]	*!	*		
☞ [pabafa]			*	*

(17.32)

If the stable system had a randomly ranked lower stratum, the variable system will have a randomly ranked intermediate stratum. We may guess, therefore, that the three rankings discussed here occur with equal frequency among the speakers. Thus, the majority of speakers with low-ranked |v| voicing will pronounce an underlying |v| as [f]. Underlying |pabava| now often surfaces as [pabafa], which is either directly perceived as /pabava/ or easily reconstructed as |pabava|.

17.3.11 Graduality and catastrophe

If a specification becomes a little bit redundant, it is going to fall down the constraint hierarchy by a little amount. This will cause the exposure of some formerly invisible randomly ranked constraints, and this will cause some free variation. This free variation will lead the listener into relying less on the presence or absence of the weakened specification. This will invite speakers (at least those who are listeners as well) to lower the faithfulness specifications again. This positive feedback loop may cause the original little sound shift to become large.

Returning to our { p b v } example: after the variation has caused low-ranked |v| voicing in all speakers, the majority of speakers will say [f]. If the implementation rule |v| → [f] becomes general, the next generation will almost certainly perceive the result as /f/ and see no reason to analyse it as anything but a fricative |f| that is underlyingly (weakly) specified as [–voice]. We then have a genuine { p b f } system.

Thus, to arrive at the preferred directions of sound change, we have to assess the hierarchies of articulatory effort and perceptual distinctiveness. We do this by randomly varying the ranking of the three constraint families that handle effort, place perception, and manner perception, though still keeping them internally contiguous.

Thus, the three constraint families together favour the inventory change { p b v } → { p b f }. In the same way, we can predict a lot of other sound changes.

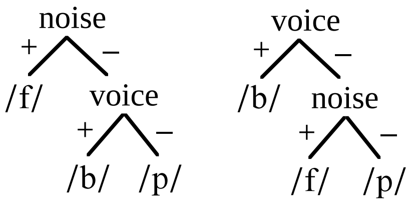
17.3.12 Unidirectionality

After the { p b v } system has become { p b f }, it will not return to { p b v }. This is because the representation of a { p b f } system that would show the free variation needed for going to { p b v }, would look exactly like the representation of a { p b v } system that would show the free variation needed for going to { p b f }, namely, the feature tree on the left-hand side of (17.26); and the relevant variation grammar is analogous to (17.25), with ranking of “voi (|f|) ≤ [f]” in the bottom stratum. The surface variation, therefore, will be the same as that discussed in §17.3.10, with a majority of speakers pronouncing an underlying |f| as [f]. Learners will specify this sound weakly, but as voiceless, not voiced. If the system ever becomes ***stable***, in the sense of (17.23), it will be with a strongly voiceless |f|.

17.3.13 Why { p b f } is such a good system

The { p b f } system is by far the most common three-obstruent system: in Maddieson (1984), it occurs twice as often as the second most common system, which is { ph p b } (Boersma 1989: 122). With the same features as in (17.25), its possible feature trees are

{ p b f }



(17.33)

As we saw in §17.3.12, the left-hand representation will not lead to sound change. I will now show that the right-hand representation is also optimal within the pool of variation that it allows. In the relevant grammar, /b/ is weakly specified for [–noise], so its realization is allowed to vacillate between [b] and [v] (a situation reminiscent of the Spanish obstruent system). The relevant constraints in the intermediate stratum (i.e., those that give different evaluations for [b] and [v]) are “voi (|b|) ≥ [v]”, “prec < [v]”, and “lab > [va]”. The following three tableaux evaluate the two candidates:

|pabafa|	voi (|b|) ≥ [v]	prec < [v]	lab > [va]	noi (|b|) ≤ [b]
[pabafa]	*!	*		
☞ [pavafa]		**	*	*

(17.34)

|pabafa|	prec < [v]	voi (|b|) ≥ [v]	lab > [va]	noi (|b|) ≤ [b]
☞ [pabafa]	*	*		
[pavafa]	**!		*	*

(17.35)

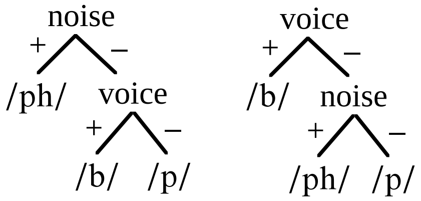
|pabafa|	lab > [va]	voi (|b|) ≥ [v]	prec < [v]	noi (|b|) ≤ [b]
☞ [pabafa]		*!	*	
[pavafa]	*!		**	*

(17.36)

Thus, the enhancement of the voicing feature is the sole supporter for [v]. The other two drives (enhancement of labiality and minimization of precision) prefer [f]. Hence, the system { p b f } will not change.

From the two representations in (17.33), we must conclude that the { p b f } system is a *sink* for sound change: both the { p b v } and { p v f } systems can be shifted to { p b f }, and the reverse shifts are impossible (with these two features). And there is a third system that may turn into our optimal system: the all-plosive system { ph p b }. Its feature systems can be described as

{ ph p b }



(17.37)

Plosiveness is not contrastive in this system, so *in both grammars* [ph] may alternate with [f]. There are three relevant intermediate constraints again. For an underlying |phabapa| with weak "plosive (|ph|) ≥ [ph]", the faithful candidate [phabapa] is preferred only by "lab > [pa]". The candidate [fabapa] wins, because it is supported by the other two, namely "noi (|ph|) ≥ [f]" and "glot < [ph]", which is ranked higher than "prec < [f]". This spirantization of the aspirates at all three articulators in an all-plosive system has probably occurred in Proto-Latin (Classical Greek /phero:/ 'I carry' versus Latin /fero:/). Look ahead to figure (17.44).

17.3.14 Tunnelling in Greek

When we compare Classical Greek with modern Greek, we see that an original { ph p b } system became { f p v }. There seem to have been two possibilities, given that the feature-tree representations have seemed to allow only single changes:

- The sequence { ph p b } → { f p b } → { f p v }. As we have seen, our model does not allow the second step.
- The sequence { ph p b } → { ph p v } → { f p v }. The first step is not allowed: it would involve enhancement of the [+voice] specification of |b| at the cost of its redundant specification for [–noise], but it would fail because two constraint families militate against it: maximum labiality ("lab > [ba]") and maximum ease (the allegedly fixed global ranking "prec < [v]" >> "lax < [b]").

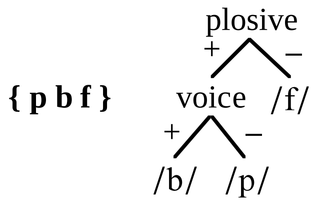
But note that if |b| is weakly specified for [–noise], |ph| can still be realized as [f]. This is because /ph/ and /f/ have the same perceptual noise and voice values; their values for [plosive] are different, but this feature is not contrastive in { ph p b }.

So [fapava] is actually a serious candidate for the implementation of |phapaba|. From the discussion above, we see that three relevant intermediate constraints favour the faithful [phapaba] candidate over [fapava]: maximum labiality in |ph| as well as in |b|, and "prec < [v]". Also three constraints favour [fapava]: maximum voicing in |b|, maximum noise in |ph|, and "glot < [ph]". So it looks a tie. But the constraint "prec < [v]" can never vote in favour of [phapaba], because it can never be ranked higher than "glot < [ph]" if our global fixation of the relation (17.6) between these constraints holds. Therefore, the articulatory constraints always favour [fapava], the labiality constraints favour [phapaba], and the enhancement constraints favour [fapava], which wins.

It seems, now, that the combined, “tunnelling”, change { ph p b } → { f p v } is possible. It requires the two Greek spirantization processes to have been simultaneous. According to Sihler (1995), they should both be dated around the first century A.D.

17.3.15 Why { p b f } is not the best system

The most common inventory { p b f } is good, but { ph b f } proves to be better if the inventory is represented with the help of the [plosive] feature:



(17.38)

Because of the choice of features (perceptual categorization), all [noise] specifications are absent, allowing |p| to surface as [ph]. This candidate wins on voicing contrast and on labiality, at the cost of implementing aspiration.

Tree (17.38) also shows that |f| has a weak specification for [–voice], suggesting that |pabafa| could be rendered as [pabava]. In (17.30)-(17.32), we have already compared these two candidates, but within a different feature set. The evaluations (17.31) and (17.32) can more or less be copied for the current situation, but (17.30) becomes

|pabafa|	plosive (|f|) ≤ [v]	prec < [f]	lab > [va]	voi (|f|) ≤ [f]
☞ [pabava]			*	
[pabafa]	*!	*		*

(17.39)

This is the small difference between noisiness and continuancy: [f] is more noisy than [v], but [v] is more continuant than [f]. Suddenly, { p b f } can change into { p b v }; it must have happened at some time in the history of Dutch.

The combined change of { p b f } directly into { ph b v } also seems possible, because (17.38) shows a combination of a weak voice specification for |f| and an absence of any specifications for noise. The aspiration of |p|, therefore, only enhances the voicing contrast with |b|, without diminishing any noise contrast, since listeners hear continuants instead of fricatives; the continuancy contrast is also enhanced. But the articulatory constraints disfavour the change: the precision gain in { ph b v } can never outweigh the aspiration loss, according to (17.6). Fatally, the labiality constraints also disfavour it: the gain in [p] → [ph] can never outweigh the loss in [f] → [v], according to (17.14).

So only the reverse combined change { ph b v } → { p b f } is allowed. But still, the inventory { p b f } can become { ph b v } in two steps. First, it can become { ph b f }, as we have seen. After that, { ph b f } can become { ph b v }, within the feature representation (17.38), since precision and continuancy improve, at the cost of labiality.

We now have a mini-cycle { p b f } → { ph b f } → { ph b v } → { p b f }. In contrast with the even shorter cycle { p b f } → { p b v } → { p b f } that we saw above, this 3-cycle works within a single feature representation (17.38), so it may well occur in a single language within a relatively short period (assuming that a feature tree tells us something about the grammar, so that we know that most speakers learn the same tree). We find another 3-cycle within the same inventory representation when we realize that { p b v } can become { ph b v }, for the same reasons as { p b f } → { ph b f }. Widespread regional variation between the four inventories compatible with the feature tree (17.36) (i.e., a voiceless or aspirated plosive, a voiced plosive, and a fricative), is found in the dialect continua of the Netherlands, Northern Germany, and Britain.

Since the aspiration of a voiceless plosive as a means for enhancing the perceptual contrast with its voiced counterpart is most urgent for dorsals, we can derive the following hypothesis of a correlation between aspiration and dorsality:

Suspicion of the dependence of aspiration on a velar voicing contrast:

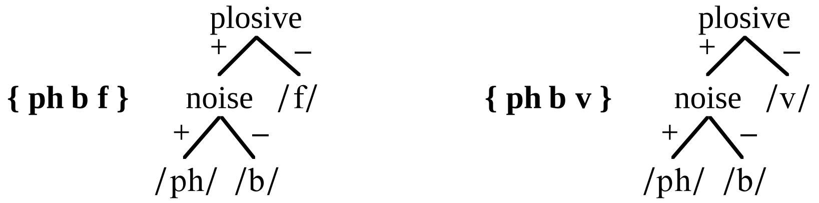
> "in languages that have a /g/ and a /k/, the /k/ may become aspirated (of course, if there is not already a /k^h/). The /p/ and /t/ may then, and only then, also become aspirated. Therefore, we do not expect many aspirates in languages with a gap at /g/." (17.40)

This hypothesis, which we saw to be able to account for the Germanic and Slavic distributions, could be checked by the data of more of the world's languages, as they have been compiled in Maddieson (1984), which contains information on the sounds of 317 typologically evenly distributed languages. Of these languages, 88 have a coronal and velar stop system of { t d k g } and 13 have { t^h d k^h g }. However, the languages with a gap at /g/ show a slightly different pattern: 10 languages have a coronal and velar stop system of { t d k } whereas no language has { t^h d k^h }. So, of the languages with /g/, 13% has aspirates, and of the languages with a gap at /g/, 0% has aspirates. Alas, there are not enough data statistically to confirm the dependence of aspiration on the presence of a velar voicing contrast (a Fisher exact test gives $p = 0.27$).

17.3.16 The second Germanic consonant shift

The proposed change { ph b v } → { p b f } feels somewhat strange: it combines two heterogenous sound changes, in contrast with the Greek double spirantization. Furthermore, its evaluation hinged on a certain way of counting how many of the six relevant constraint rankings spoke for and against it: we could have said that it only got the articulatory and labiality votes, and that two faithfulness constraints opposed it; this is in sharp contrast with the Greek { ph b p } → { f v p } change: no matter how we count votes, it is always favoured. So let us concede that inventories that adhere to (17.38) will eventually get an aspirated plosive.

I will now show that these { ph b f } and { ph b v } systems favour devoicing of their /b/ segments. First, we can immediately see that "lab > [ba]" and "lax < [b]" will always favour the [b] → [p] change. So we only have to show that this change does not violate any undominated faithfulness constraints, i.e., that we can find feature trees that show weak voicing specifications for /b/. Here they are:



(17.41)

In both trees, |b| is unpecified for voice, since the voicing feature does not belong to the listener’s repertoire of perceptual dimensions. Therefore, speakers are free to vary their voicing of this sound, and the articulatory and place-distinction drives will force its voiceless realization, despite the resulting proximity between |ph| and the new (reanalysed) |p| segment. This devoicing of originally voiced plosives is known in various degrees from Old High German, present-day Alemannic, English, Danish, and Icelandic; the result is typically “lenis” (weak and short) voiceless, and this sound contrasts with an aspirated or “fortis” (longer and stronger) voiceless sound.

The resulting inventories are { ph p f } and { ph p v }. Still within the representation of (17.39), the first of these can change into the second, because that would improve continuancy (the [–plosive] specification) of |f| and decrease the required precision, at the cost of a loss in labiality cues.

We are left with a single { ph p v } inventory. We met it before in the discussion of Greek: with a voice-noise tree like the one in the right-hand side of (37), its |v| is weakly specified for [+voice], and it will become pronounced as [b] because maximum labiality and ease numerically outweigh its [+voice] specification.

We have now found the five-cycle { p b f } → { ph b f } → { ph p f } → { ph p v } → { ph p b } → { f p b }. Note that the lexicon is not back where we started: the three obstruents have rotated.

The spirantization of |ph| could have occurred directly in { ph p v }, again with the voice-noise features (17.37): the noise contrast between |ph| and |p| is enhanced (note that |v| was only weakly specified for [+noise]), and the loss of aspiration is an articulatory gain (labiality cues deteriorate). The result is the { f p v } inventory that we met before, and we have identified another cycle.

The presence of a [pf] or [f] is what distinguishes High German from the other Germanic languages, which have not come further than [ph].

Here is a summary of all the allowed sound changes discussed above:

From:	**To:**	**Feature tree:**	**Constraints for:**	**Constr. against:**
p b v	p b f	noi voi	+noi, lab	prec-f
	ph b v	plos voi	–voi, lab	glot
p b f	ph b f	plos voi	–voi, lab	glot
	p b v	plos voi	–plos, prec-f	lab
p v f	p b f	voi noi	prec-v, lab	+voi
ph p b	f p b	voi noi, noi voi	+noi, glot	lab
	f p v	voi noi	+noi & +voi, glot	lab
ph p v	ph p b	voi noi	prec-v, lab	+voi
	f p v	voi noi	+noi, glot	lab
ph b v	p b f	plos voi	lab, glot	–plos & –voi
	ph p v	plos noi	lab, lax	
ph b f	ph b v	plos voi	–plos, prec-f	lab
	ph p f	plos noi	lab, lax	
ph p f	ph p v	plos noi	–plos, prec-f	lab

(17.42)

All these cases are also visible in (17.44).

17.4 The Germanic consonant shifts

Let us now set out to explain the circularity of the Germanic consonant shifts in word-initial position.

17.4.1 Data

In the first Germanic consonant shift (between Proto-Indo-European and Old Germanic), original voiceless stops became voiceless fricatives, original voiced stops became unvoiced, and original 'murmured' stops became plain voiced stops (or voiced fricatives). The second Germanic consonant shift (from Old Germanic to High German) more or less repeated the first. Here is a simplified review of these historical sound changes (cognate Latin words have been added for comparison):

Proto-IE	*p	(*b)	*bɦ	*t	*d	*dɦ	*k	*g	*gɦ
Latin	pel		fol	tri	dw-	(fa-)	kan	gel	(host)
Old Germanic	fel	poːl	bal	θri	tw-	doː-	xund	kald	gast
High German	fel	pfuol	pal	(dri)	tsw-	tuo-	(hund)	kxalt	kast
gloss	'skin'	'pool'	'leaf/ ball'	'three'	'two'	'do'	'dog'	'ice/ cold'	'guest'

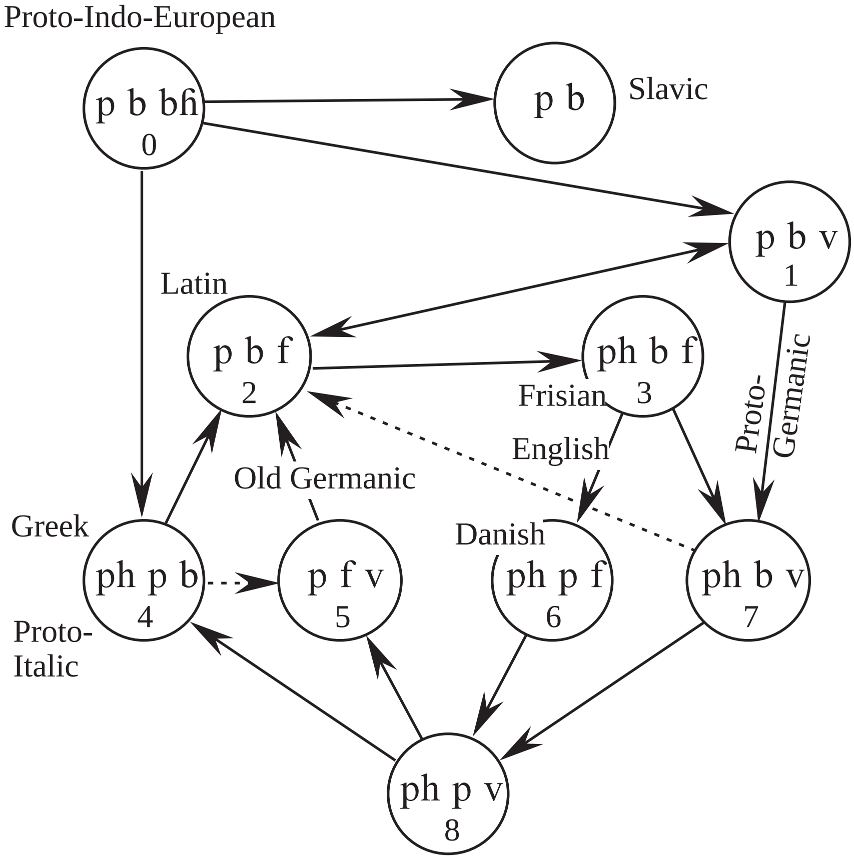
(17.43)

The endings of the words have been suppressed. Between parentheses, we find the results of some changes that disturb the general pattern:

- The Latin change from the PIE system { bɦ dɦ gɦ } through word-initial { ph th kh } and the non-strident fricatives { ɸ θ x } to the classical { f f h }, with merger of the two anterior fricatives.
- The Common Germanic change of /x/ into /h/.
- The German change of /ð/ into /d/.

17.4.2 The preferred direction of change

The eight three-consonant systems discussed in §17.3.16 are all shown in figure (17.44). The arrows represent the directions of sound change that are preferred according to §17.3. These preferred directions are equal to the directions of the first and second Germanic consonant shifts in word-initial position. The solid arrows denote the changes of a single segment, the dotted arrows those of two segments. The two possible systems that are not shown are { b, v, f } and { ph, f, v }. These would only have arrows that point away from them.



(17.44)

17.4.3 Proto-Indo-European obstruents

There has been some debate about the nature of the plosives in Proto-Indo-European. Gamkrelidze & Ivanov (1973) considered the system { p bɦ, t d dɦ, k g gɦ } typologically unnatural, because in most languages with voiced aspirated plosives, there are also voiceless aspirated plosives. Moreover, there was a gap in the labial system, which lacked the plain voiced plosive [b]. This is a typologically marked situation, too: if a language lacks one of the voiced stops { b d g }, it is usually [g] that is missing (for good reasons, as we saw). However, from a system of ejectives { p' t' k' }, the first element to drop out is usually the labial [p'] (for equally good reasons). Therefore, Gamkrelidze & Ivanov's proposal included ejectives instead of plain voiced stops, and the voiceless plosives would have been aspirated, because the voiced aspirates would otherwise have no unmarked (voiceless) counterparts. The system resulting from this theory is thus { ph bɦ, th t' dɦ, kh k' ɡɦ }. A related theory (Beekes 1990) does have plain voiceless plosives, but the aspirates are voiceless: { p ph, t t' th, k k' kh }.

These 'glottalic' theories may be typologically and phonetically more satisfactory than the traditional account, but they run into severe problems if we try to find the paths that led from the PIE system to the systems attested in the daughter languages. The theory by Beekes poses the largest problem: the aspirated stops would have become fricatives in Latin, voiced aspirates in Sanskrit, and voiced stops in Slavic and Germanic. By contrast, voiced aspirates are obviously problematic in many sound systems, so it would only be natural that the problem has been solved in so many ways in the various branches. As we have seen in our discussion on the fate of [ɡ], problematic consonants are struggling to find a different place in the system, and they do so in any suitable way. So, while [ɡ] had the choice to develop into [ɣ], [ŋ], [ɦ], or [dʒ], the breathy stop [dɦ] had the choice of changing into any neighbouring consonant, be it [th] (Greek and Latin), [d] (Slavic), or [ð] (Germanic and Latin). Thus, the diversity of the reflexes in the daughter languages points to a problematic ancestor. However, if we make the unconventional move of regarding the Slavic data as evidence of an earlier stage with voiced fricatives (as in Germanic, see §17.1.2), the sequence th → θ → ð → d may be defensible.

Secondly (in both glottalic theories), the ejectives would have become plain voiced stops in Italic, Greek, Celtic, and Slavic, independently. The classical theory has far more straightforward reflexes of its plain voiced stops: they stayed that way in most branches, but became voiceless in Germanic, a natural process which repeated itself later in High German. For detailed criticism on the glottalic theory, see Hayward (1989) and Garrett (1991).

Can we have both? Yes. The glottalic theory may well be the result of an internal reconstruction on Proto-Indo-European, and the change of { th t' dɦ } or { t t' tɦ } into { t d dɦ } may well have occurred before the split.

The top left of figure (17.44) thus shows the PIE system '0', *with* the notoriously problematic /bɦ/. In Old Greek, for instance, this sound changed into /ph/, giving system '4'.

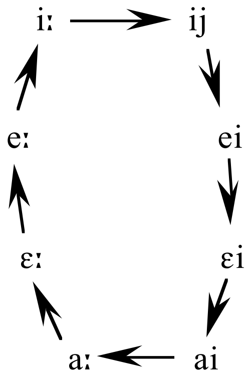
17.4.4 Circularity

In (17.44), we see several cycles of optimization, as discussed in §17.3.15 and §17.3.16.

If we assume a Pre-Germanic system like { p b β } (numbered '1' in figure (17.44); we ignore the difference between /β/ and /v/), we see that there are two different routes to arrive at the favoured sound system { p b f }. Most authors state that Latin took the same route as Greek (Sihler 1995), at least in initial position (Stuart-Smith 1995). The second route to the preferred { p b f } starts with aspirating the /p/, which is what Proto-Germanic may have done. This language then necessarily obtained system ‘8’, then ‘4’ or ‘5’, and then number ‘2’. This is why Latin and Old Germanic both feature the favoured { p, b, f } system, though these systems are shifted relatively to one another.

17.5 Vowel shift

The eternal circular optimization loop is also seen in the vowel shifts that occurred around 300 years ago in many West-Germanic dialects:



(17.45)

The arrows denote the direction of the chain shift. Basically, monophthongs rose (Dutch slaːpə ‘sleep’, Old English slæːpɑn, Old Frisian sleːpan, English sliːp), and diphthongs fell (Limburgian wiːn ‘wine’, Dutch ʋɛin, English wain, Flemish waːn). The chain may have started as a result of the lengthening of short vowels in open syllables, which crowded the height dimension if their lengths and tone contours came to equal those of the originally long vowels (Limburgian, which preserves the West-Germanic long vowels, developed a three-way length contrast and a tone contrast).

The rise of the monophthongs can be understood from an asymmetry between adults and children. An adult open mid vowel [ɛː] with an F_1 of 600 Hz will be imitated by young children as a vowel with an F_1 of 600 Hz instead of as an open mid vowel, if we assume that it takes the learner some time to develop an adult-like vowel-height normalization strategy. To implement a vowel with an F_1 of 600 Hz, the child produces a close mid vowel [eː]. If the associated articulation persists into adulthood, the vowel will have risen from one generation to the next. Thus, monphthongs tend to rise if their primary acoustic cue is F_1.

The fall of the first part of the diphthongs can be understood from the contrastive representation of diphthongs. If a language has a single diphthong, its primary perceptual feature may well be its **diphthongal character** (e.g., the presence of an unspecified F_1 fall), by which it is contrasted with all the other vowels. Lowering of the first part of the diphthong amounts to enhancing the contrast with the other vowels: the more the two parts of the diphthong differ from one another, the more they will contribute to the diphthongal character.

17.6 Conclusion

The strict-ranking approach allows us to model an eternally improving sound system. Even when there are no external factors, sound change may go on forever. The possibility of circular optimization is a property of random constraint variation in general, not of the details of the more involved production grammars that I proposed in §17.3.

When viewed from a distance, the procedure seems teleological, because the inventory has improved. The changes were, however, automatic, not goal-oriented; the apparent teleology arises because the constraints themselves are functional principles.

18 The Obligatory Contour Principle

Abstract. Though seemingly a good candidate for a universal output-oriented constraint, the OCP does not occur as a constraint in the production grammar. Instead, it handles, in interaction with the No-Crossing Constraint, the correspondence between acoustic cues and perceptual feature values in the perception grammar. Because faithfulness constraints use the perception grammar to evaluate the similarity between the perceptual specification and the perceptual output in the production grammar, the OCP does influence the evaluation of candidates in the production grammar. As a result, adjacent identical elements are avoided because they constitute PARSE violations. Dissimilation at a distance, by contrast, is due to a constraint against the repetition of articulatory gestures.

In this chapter and the next, I will point out the advantages of distinguishing between articulatory and perceptual features in autosegmental phonology. According to McCarthy (1988), the only phonological processes that can be accepted as primitives in autosegmental phonology, are ***spreading***, ***deletion***, and the ***obligatory contour principle*** (OCP). While the next chapter will centre on spreading, the current chapter will tackle the OCP.

McCarthy (1986) expresses the Obligatory Contour Principle (OCP) in its naked form as follows:

> “adjacent identical elements are forbidden”

As we will see, many phenomena have been ascribed to this principle.

18.1 Functional interpretation of the OCP

From a functional standpoint, the OCP is not a single primitive principle, but branches into two fundamentally different ones. Furthermore, one of these two principles is naturally embedded in a set of constraints on simultaneous and sequential combinations of gestures and perceptual features.

18.1.1 Articulatory and perceptual motivations

If we distinguish between articulatory and perceptual phenomena, the OCP branches into two principles.

The first is a general principle of human perception, not confined to phonology. In designing a country map of Europe, the cartographer can choose to fill in the countries with the minimal number of four colours that are needed to give every pair of adjacent countries different colours. If she decided to paint both the Netherlands and Belgium red, the reader of the map would not be able to identify them as separate countries; thus, in cartography, adjacent identically coloured countries are avoided.

Likewise, if a morph ending in /-m/ is concatenated with a morph starting with /m-/, the usual timing of syllable-crossing clusters will result in the long consonant [-mː-]. The perceptual identity of one of its constituents is therefore lost, violating featural faithfulness. Some of the information about the existence of two morphemes is kept in the timing, but if the language is adverse to geminates, it may just end up with [-m-], violating some more faithfulness.

The problem of the long perceptually homogeneous sound can be levied by inserting a pause between the two consonants (i.e., drawing a black border between the Netherlands and Belgium): giving [[-m_m-]]. This violates a FILL (pause) constraint: a pause can be perceived as a phrase boundary. Another strategy would be to insert a segment (declaring the independence of the southern provinces of the Netherlands, and painting them blue), which will give [-məm-] or so: another FILL violation. Language-specific rankings of all the faithfulness constraints involved will determine the result.

The perceptual nature of this first functional correlate of the OCP is shown by the rules of vowel insertion in English, which are hard to capture with generalizations over single tiers in feature geometry. Thus, the insertion of /ɪ/ before the morpheme /-z/ occurs in *bridges* but not in *tents*, exactly because [dʒz] would contain a perceptually unclear boundary (The Netherlands in red, Belgium in purple), and [nts] would not; likewise, the insertion of /ɪ/ before the morpheme /-d/ occurs in *melted* but not in *canned*, because the boundary would be lost in [ltː] but not (or less so) in [nd].

The second functional correlate of the OCP is simply the tendency not to repeat the same articulatory gesture: an articulatory *REPEAT constraint. The features involved in this constraint are arguably of an articulatory nature: the Japanese constraint against two separate voiced obstruents within a morpheme obviously targets the articulatory gesture needed for the voicing of obstruents, not the perceptual voicing feature, which would also apply to sonorants. A clear difference with the first principle is exhibited by a morpheme-structure constraint in Arabic, which does not allow two labial consonants within a root; apart from disallowing two appearances of /b/, it does not even allow /m/ and /b/ to appear together. This generalization over plosives and nasals is typical of the articulatory labial gesture, which does not care whether the nasopharyngeal port is open or not, whereas the divergent behaviour of plosives and nasals in *melted* versus *canned* is exactly what is expected from a perceptually conditioned phenomenon.

The predicted correlations between near OCP effects and faithfulness constraints, and between distant OCP effects and articulatory constraints, are verified in this chapter.

18.1.2 Simultaneous and sequential combinations of features

I will identify the “perceptual” OCP as one of the four constraint clans that handle ***combinations*** of articulatory gestures or perceptual features. In a functional phonology (ch. 6), we express articulatory implementations in articulatory features (gestures) *or their combinations*, and we express perceivable sounds in perceptual features *or their combinations*.

Articulatory constraints on combinations of gestures.
From general properties of the acquisition of human motor behaviour (namely, the ability to group simultaneous or sequential gestures into a more abstract coordination or motor

program), we can posit the unity of often-used coordinated gestures (like the lip and tongue body gestures in [u], in a language where this sound is common), and the unity of common sequences of gestures (like the lip closing and opening gestures in [apa], in a language where [p] often occurs intervocalically), which leads to assuming separate constraints for these more abstract articulations:

*COORD (a_1: g_1; a_2: g_2 / ...): "do not combine the gesture g_1 on the articulator a_1 with the gesture g_2 on the articulator a_2." Cf. (7.29).

*SEQ (a_1: g_1; a_2: g_2 / ...): "do not follow the gesture g_1 on the articulator a_1 with a gesture g_2 on the articulator a_2."

Faithfulness constraints on combinations of perceptual features.

From general properties of the acquisition of human perception (namely, the ability to group simultaneous or sequential percepts into a more abstract representation), we can posit the unity of often-heard simultaneous features (like labiality and nasality in [m], in a language where this sound is common), and the unity of common sequences of features (like the nasal murmur, the silence, and the explosive burst in [ampa], in a language where [mp] often occur in sequence), which leads to assuming separate faithfulness and correspondence constraints for these more abstract percepts:

*REPLACEPATH ($f \times g$: $x \times z$, $y \times w$): "do not replace the values x and z on the combined perceptual tiers f and g with the different combination y and w." Cf. (9.49). For instance, depending on the language, a surfacing of /n/ as /m/ may involve a violation of *REPLACE (place: cor, lab / nas) or a violation of *REPLACEPATH (place × nas: cor × +nas, lab × +nas) (§12.7).

OCP (f: x; q_1 | m | q_2): "A sequence of acoustic cues q_1, q_2 with intervening material m is heard as a single value x on the perceptual tier f." (12.10)

The OCP is just one of the four combination constraints. It belongs in the ***perception grammar*** (ch. 6, 8, 15) since it handles the categorization of acoustic input into perceptual features. As such, it also plays a crucial role in featural correspondence in the ***production grammar*** (§12.3).

18.2 History of the OCP

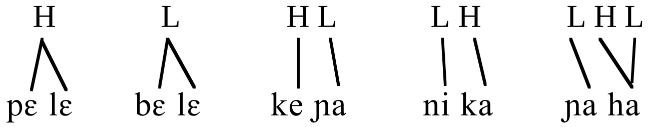
In order to be able to defend the descriptive adequacy of a functional account of the OCP, we have to investigate first the various interpretations it has suffered throughout the years, and the types of phenomena it has been invoked to explain.

18.2.1 The original OCP

The first expression of the OCP is commonly attributed to Leben (1973). In his defence of ***suprasegmental phonology***, he demonstrated that tone features and nasality show suprasegmental behaviour in several languages. For example, in Mende, mono- and bisyllabic monomorphemic nouns have the following possible tone sequences (H = high, L = low, HL = falling, etc.):

1. Nouns of one syllable have H (/kɔ́/ ‘war’), L (/kpà/ ‘debt’), HL (/mbû/ ‘owl’), LH (/mbǎ/ ‘rice’), or LHL (/mbǎ̂/ ‘companion’), but not HLH.
2. Nouns with two syllables have H-H (/pɛ́lɛ́/ ‘house’), L-L (/bɛ̀lɛ̀/ ‘trousers’), H-L (/kɛ́ɲà/ ‘uncle’), L-H (/nìká/ ‘cow’, or L-HL (/ɲàhâ/ ‘woman’).

The five tone sequences for the bisyllabic nouns can be seen to be equal to those of the monosyllabic nouns, if we represent Mende tone in a suprasegmental way:



(18.1)

The generalization is that Mende has a morpheme structure constraint that allows only tautomorphemic tone sequences H, L, HL, LH, and LHL.

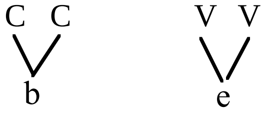
Leben’s analysis assumes that the segmental and suprasegmental information is specified in the lexicon as two independent sequences, and a phonological rule maps the tones to the syllables. This rule assigns tone to syllables in a one-to-one fashion from left to right. If there are more syllables than tones, the last tone spreads through the remaining syllables.

Thus, Mende provides evidence for the fact that an apparent sequence of high tones on adjacent syllables should sometimes be viewed as only one H. Leben (1973) does not suggest yet, however, that tautomorphemic HH sequences are universally impossible; the strongest statement in that direction is his assertion (p. 94) that for Mende, “the distinction between HLL and HHL is representable in McCawley’s notation, while in suprasegmental notation, both sequences reduce to HL”.

18.2.2 Morpheme-structure interpretation of the OCP

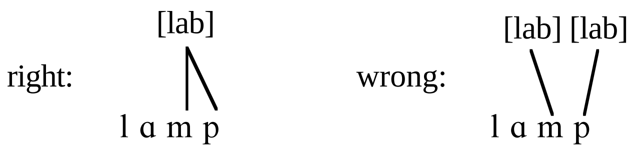
Goldsmith’s (1976) ***autosegmental phonology*** changed the language and the scope of the suprasegmental approach. From then on, Mende was said to have independent tonal and segmental *tiers*, which were ***linked*** with ***association lines*** (visible in the above representation of bisyllabic nouns). Goldsmith made Leben’s tacit assumption explicit, coining it the ***Obligatory Contour Principle***: “At the melodic level of the grammar, any two adjacent tonemes must be distinct. Thus HHL is not a possible melodic pattern; it automatically simplifies to HL” (Goldsmith then goes on to reject the OCP).

In autosegmental theory, this principle acts as a condition on representations, valid within any tier. On the root-node tier, the interpretation is that geminate consonants and long vowels should be represented as e.g.



(18.2)

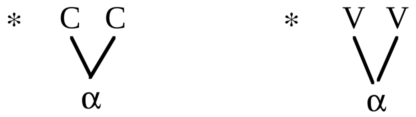
In these figures, association lines link the root nodes /b/ and /e/ to the CV skeleton (or *timing tier*). On the labial tier, the OCP ensures that the last two segments of /lɑmp/ share one [labial] specification, and not two:



(18.3)

The functional interpretation of these ideas is that /bː/ and /mp/ are both implemented with a single gesture of the lips, and that in /eː/, the tongue and jaw are held in one position.

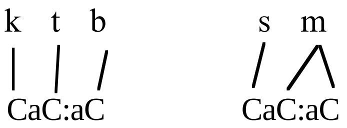
As a universal condition on phonological representations, the OCP helps us to express constraints on morpheme structure. For instance, if a language has no geminate consonants or long vowels, these constraints can be expressed as the following filters:



(18.4)

With this interpretation, the OCP seems to be satisfied in most cases in most languages. Exceptions, e.g., cases where it is favourable to represent a /mp/ sequence as having two separate [labial] specifications, are known as ***OCP violations*** (we can now note that the OCP has long been the only violable constraint in generative phonology). Within morphemes, the OCP is thought to be universal, and this can be tested by investigating the scope of phonological rules.

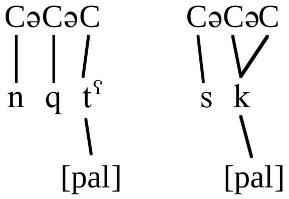
A good example is found in the root structure of Semitic languages. McCarthy (1986) formulates the OCP as follows: “At the melodic level, adjacent identical elements are forbidden.” McCarthy restricts himself to absolute identity of segments, and argues that in Semitic languages, consonants and vowels are on different tiers, where the consonant sequence can be identified with the lexical root. For instance, in Bedouin Hijazi Arabic (McCarthy 1982), /katːab/ ‘he wrote’ clearly has the root /ktb/ ‘write’, which is mapped onto the pattern /CaCːaC/ ‘(past 3 sg.)’. Likewise, /samːam/ ‘he poisoned’ seems derived from the root */smm/, but this is impossible according to the OCP, because two identical /m/ would be adjacent: the root must be /sm/, and the /m/ is spread to the last consonant (we already inserted the vowels into the template):



(18.5)

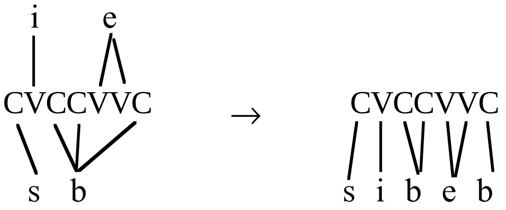
This analysis is corroborated by many facts, including from a language game (McCarthy 1982), which freely commutes the root consonants and leaves the pattern intact, so that /katːab/ may become /batːak/, /kabːat/, /takːab/, /bakːat/, or /tabːak/. By contrast, /samːam/ ‘is mapped to /masːas/ only, which is exactly what we expect if the root is /sm/ and the game works directly on the root. Apparently, the root is indeed /sm/. Because *all* /C$_1$aC$_2$ːaC$_2$/ words in Arabic behave in the same way as /samːam/, we can see that their roots must be /C$_1$C$_2$/; thus, all Arabic roots obey the OCP and Arabic cannot distinguish between /sm/-like and /smm/-like roots.

If Semitic roots must always be analysed as satisfying the OCP on the consonantal level, we can expect morphological and phonological rules to work on the two reflexes of the second consonant of biconsonantal roots. McCarthy provides the following example. In Chaha, the feminine form of the imperative is made by applying palatalization to the last coronal or velar consonant in the root. Thus, /nəqət^{ʕ}/ 'kick' becomes /nəqət^{ʕj}/. The biconsonantal /səkək/ 'plant in the ground', however, is taken to /sək^{j}ək^{j}/. Apparently, for this morphological operation, the two /k/ act as a single consonant:



(18.6)

Most *phonological* rules, however, do not show this behaviour. For instance, Tiberian Hebrew has a rule of postvocalic spirantization, which does not apply to geminates. Still, this rule takes /sibːeːb/ 'he surrounded', from the root /sb/, to /sibːeːβ/, changing only a part of the alleged single consonant. The solution is that this rule works on a form that is created by the ***conflation*** of the consonantal and vocalic tiers. Thus, /sb/ + /CiCːeːC/ is first converted to /sibːeːb/ in a process called ***Tier Conflation*** (a generalization of the process of ***Bracket Erasure*** known from concatenative morphology), which is applied after each stratum in the lexical phonology:



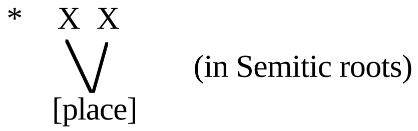
(18.7)

In /sibːeːb/, the two /b/ are no longer adjacent, and postvocalic spirantization will affect the last /b/; the branching /b/ in the geminate is still one consonant, because we do not get */siβbeːβ/. In the Chaha example, we saw that morphological palatalization occurred *before* the conflation of the root and pattern tiers.

As Odden (1988) points out, it remains a question why many phonological rules are never seen to precede Tier Conflation. For example, McCarthy (1986) notes that ***geminate integrity*** (the universal failure of phonological rules to target only one of the supposed members of a geminate) is only valid for ***surface*** tautomorphemic geminates.

McCarthy (1988) generalizes the OCP to all the tiers that are known from the theories of *feature geometry*. In part, this move adapted the OCP theory to a criticism by Odden (1988) and Yip (1988), namely that sequences of homorganic consonants, like /bm/, are ruled out in Semitic roots. If a separate morpheme structure constraint would be needed to capture this fact, this constraint would encompass the OCP, so the OCP would be superfluous as far as morpheme structure is concerned. To save the OCP, McCarthy extended it to include the labial tier and other articulator tiers. Thus, the root /btm/ is ruled out because the labial specifications of /b/ and /m/ are adjacent on the [labial] tier.

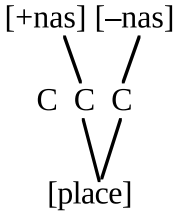
Suddenly, however, the near-universal OCP, which might be parametrized but whose "default value is *on*" (McCarthy 1986), was replaced by a highly language-dependent tendency-like constraint, whose default value must be *off*; most languages, after all, do not object to tautomorphemic homorganic consonants like /mb/, because no OCP is violated. Therefore, the OCP is not strong enough to rule out the root /tmb/ all by itself; we must enforce the additional constraint that the place node must not branch:



(18.8)

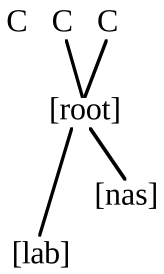
This well-formedness condition would do the job, with the help of the OCP, which universally does not allow separate identical place nodes; the root /btm/ must then be ruled out by the additional constraint that the [labial] node does not branch (since on the place tier, /b/ and /m/ are not adjacent). McCarthy does not state these constraints explicitly, probably because they are not entirely true: for many pairs of consonants, they are only tendencies, i.e. they express markedness criteria. Besides, introducing these non-branching constraints would have to change the formulation for the case of identical segments as well.

To see this, we must consider the interpretation of root-consonant formulas like /ktb/, /sm/, /smm/, and /tmb/. In early generative phonology, symbols like /m/ and /b/ were used to transcribe segments and were to be interpreted as shorthands for complete feature bundles. Since the advent of autosegmental phonology, the interpretation has changed, because some features are shared between adjacent segments. Consider first the interpretation of /tmb/. If it stands for a sequence of three separate complete feature bundles, the OCP is violated. But if it is interpreted in the usual way in which a segmental sequence like /tumb/ is interpreted, namely as



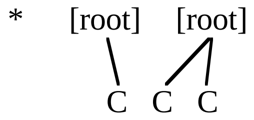
(18.9)

then the OCP is not violated. An analogous story can be told for the root node: the sequence /smm/ could mean three separate feature bundles, in which case the OCP is violated. However, interpreted in the same way as is usual for /summ/ (though that would be transcribed as /sum:/), the OCP is not violated:



(18.10)

The fact that we have a separate linear-notational device for branching root nodes (as in /sumː/), and lack one for branching place nodes (or it should be /tum͡b/), cannot be a reason for treating geminates in a different way from homorganic NC clusters. Thus, a grammar for Semitic root consonants must contain a filter that prohibits branching root nodes:

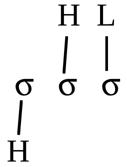


(18.11)

18.2.3 The OCP in phonological rules

Interesting things may happen when morphological or syntactical operations produce structures that threaten to violate the OCP. For instance, suppose that a morpheme with a H tone is concatenated to a morpheme with a HL contour. Before the morphological operation, the two tone sequences can be thought of as sitting on different tiers, so that the OCP is not violated (McCarthy 1986). Any of the following representations of this situation will do (σ = syllable):

H + HL



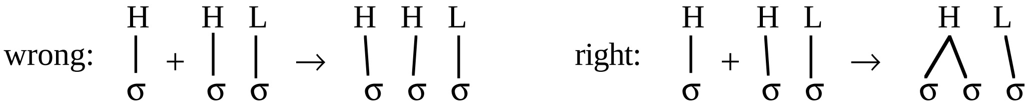
[[H][HL]]

(18.12)

After the morphological operation, which removes the '+', conflates the two tiers, or erases the inner brackets, a resulting HHL contour would violate the OCP. A number of theories about what happens, have been proposed.

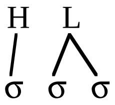
18.2.4 The fusion interpretation of the OCP

In order to satisfy the OCP, the result of the concatenation H+HL may be a HL contour, with the H doubly connected:



(18.13)

Evidence for this falling together of the two H tones must come from later phonological rules, which are predicted to treat the resulting single H as one toneme. For instance, if the default left-to-right one-to-one tone-mapping rule applies cyclically to the output of the above morphological operation, the result will be



(18.14)

which would be a clear proof of the fusion of the two H tones.

Thus, the OCP seems to play an active role in concatenation: it collapses two adjacent identical elements into one. This is the ***fusion*** interpretation of the OCP; it was assumed by Goldsmith (1976) for phonetics and by Leben (1978) for tone.

18.2.5 Antigemination: the blocking interpretation of the OCP

McCarthy (1986) does not agree with the fusion interpretation of the OCP. He notes that in Afar (Bliese 1981), the stressed suffix /ˈi/ ‘(nom.-gen.)’ takes the stress away from the root, thus creating a situation in which a syncope rule can delete the originally stressed vowel:

ħamˈil ‘swampgrass’ + ˈi → ħamilˈi → ħamlˈi (18.15)

The fusion interpretation of the OCP would handle a root with two equal consonants in the following way:

miɖˈaɖ ‘fruit’ + ˈi → miɖaɖˈi → *miɖɖˈi → *miɖːˈi (18.16)

We see that this is a scheme that would *repair* the OCP-violating *miɖɖˈi by fusion, like the tone example above can be interpreted as repairing an offending HHL sequence. But the actual result is /miɖaɖˈi/, not */miɖːˈi/. Apparently, the OCP blocks the syncope rule: the OCP violation is *prevented*, not repaired. This is the ***blocking*** or ***antigemination*** interpretation of the OCP.

McCarthy (1986) is quite explicit about his preference: “I reject the fusion interpretation of the OCP and hold instead to its blocking effect”. The first reason that he mentions is that “we never find application of syncope followed by restructuring of the output”. The second reason deserves to stand out:

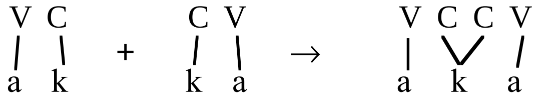
> “the idea that universal or language-specific constraints on phonological well-formedness function as negative rather than positive filters is far more typical of the vast majority of uses of constraints in the literature” (McCarthy 1986, p.222)

In this comparison, ***positive filters*** are repair rules: even the OCP can be seen as a *rule* that collapses identical adjacent elements. ***Negative filters***, on the other hand, evaluate the possible output of a rule and are capable of blocking that output. It is crucial that these cases should be seen as blocking of syncope, not of triggering of epenthesis.

What, then, with the simple fusion that is found so often when a morpheme ending in /-ak/ is connected to a morpheme starting with /ka-/? McCarthy (1986) maintains that either of two possibilities arise. The first is:

$$\begin{array}{cccccccccccc} V & C & & C & V & & V & C & C & V \\ | & \backslash & + & | & \backslash & \rightarrow & / & | & \backslash & \backslash \\ a & k & & k & a & & a & k & k & a \end{array} \qquad (18.17)$$

This violates the OCP. However, according to McCarthy, this is not a serious problem, because it “arises from a conflict between the OCP and another universal principle, Tier Conflation”, in which “we are free to dictate priority between the two as we choose”. This formulation suddenly introduces out of nothingness a new device into the repertoire of phonology, namely, the idea of strict ranking of violable constraints. For the moment, McCarthy rejects this idea, and prefers the hypothesis that Tier Conflation causes fusion:



(18.18)

18.2.6 The OCP as a rule trigger

Fusion and antigemination are only two ways of satisfying the OCP. Yip (1988) extends the power of the OCP in such a way that the OCP works as a trigger for rules that repair OCP violations:

> "all rules involving identity of target and trigger with an output in which they are no longer identical and adjacent are OCP-triggered rules." (Yip 1988, p. 73)

In Yip's analysis, Tier Conflation exists of two stages. In the first stage, the two morphemes are concatenated. This may raise an OCP violation, which can be repaired by a repair rule if such a rule exists. In the second stage, if no repair rule has cancelled the OCP violation, the remaining adjacent identical elements are automatically fused.

18.2.7 Long-distance OCP effects

Yip (1988) and McCarthy (1988) work within a theory of underspecification and feature geometry. Under such a regime, elements can be far apart and still be adjacent on a tier with underspecified or privative features. For instance, the two /p/ in /pap/ would either share their [labial] specification, or otherwise violate the OCP. Because in most languages phonological rules normally treat the two /p/ as distinct segments, even if only labiality is targeted, we must conclude that the OCP, understood in this way, is violated by default and is nothing more than a markedness principle. Crisis results.

Steriade (1995) seems to interpret the OCP as a constraint against the repetition of a feature, and Myers (1994) formulates his language-dependent "OCP!" as follows:

> "A feature value [F] should not appear twice inside a specified domain".

18.2.8 Multiple strategies in satisfying the OCP

All the outputs proposed by the various theories have the common property that the OCP is not violated. The OCP seems capable of ***blocking*** any rule whose output would violate it, and of ***triggering*** any rule that would prevent or repair its violation. Yip (1988) concludes that the OCP:

> "acts as an MSC, rule blocker, rule trigger, constraint on the mode of operation of an ambiguous rule, and constraint on the form of possible rules." (Yip 1988, p. 97)

Now we have the situation that a constraint may induce a plethora of effects. As we have seen, this situation clearly poses problems for derivational approaches to phonology. It could only be handled well within a framework that allows the parallel evaluation of all thinkable candidate outputs; therefore, a constraint-based framework. McCarthy's preference for negative filtering would then be honoured, even for situations that would look like active repair in derivational approaches.

18.2.9 Violability of the OCP

Though he invented its name, Goldsmith (1976) did not accept the OCP. Odden (1986, 1988, 1995) has presented much evidence that the OCP is not a universal principle.

> "The strongest possible version of the OCP at this point is that there may be a dispreference for adjacent identical tones; languages are free to express this dispreference by constraining lexical representations, by adding rules of tone fusion or tone deletion, or by putting conditions on tone spreading rules. Ultimately, languages retain the option of doing nothing about OCP violations." (Odden 1995, p. 464)

Our constraint-based framework, therefore, should be one in which constraints may be violable, so that the winner is determined by a language-specific ranking of the constraints; in that case, there would be nothing special at all about the OCP: it would be a violable constraint like all others.

If Prosodic Morphology had not existed, the OCP situation might have induced the paradigm shift that came with Optimality Theory.

18.3 Functional interpretation of concatenation problems

As is natural in an OT framework, our typology of phonological phenomena will centre around structures and their ***problems***, not about rule types and their applications. The first structure that we will consider, is |ak+ka|.

18.3.1 Acoustic results

If a morph ending in |ak| is concatenated with a morph that starts with |ka|, the most straightforward result would be an implementation of |akka| with two dorsal closing gestures, timed in the same way as is usual in the language for |apka| and |akpa|. If the language would normally overlap two heterorganic closing gestures (giving [[ap˺_ˑ^{k}a]] and [[ak˺_ˑ^{p}a]][1]), the acoustic result of |ak+ka| would be [[ak˺_ˑ^{k}a]], or [akˑa], with a prolonged closure: a short geminate, acoustically indistinguishable from a single closure with a hold phase. If the language would normally not overlap two heterorganic closing gestures (giving [[ap˺_p^{h}_ka]] etc.), /ak+ka/ would still give [[ak˺_ː^{k}a]], or [akːa]: a long geminate. This is because two adjacent partially overlapping styloglossus commands would result in a single long dorsal closure (§18.3.5).

[1] The double brackets indicate a *microscopic transcription*: a transcription with one symbol for every perceptual change (§1.3.3). The sequence [aka], for instance, is transcribed as [[ak˺_^{k}a]]. The three components of the plosive are: transition, silence, burst. The symbols [ˑ] and [ː] mark half-long and long elements, respectively.

18.3.2 Perceptual results

To compute the perceptual output of the production grammar, we must run the acoustic result [[ak˺_ː^{k}a]] through the perception grammar (fig. 6.1). Nearly all languages would perceive a *non*-geminate [[ak˺_^{k}a]] as /aka/, which is a shorthand for something with a single perceived dorsal value on the perceptual place tier (§12.3). This is because the sequence transition-silence-burst often occurs together in most languages, so that it is advantageous for the listener to perceive them as one. The relevant near-universal ranking of the relevant OCP (12.10) and no-crossing (12.11) constraints is

OCP (place: dor; tr | sil | bu)
|
NCC (place: dor; tr | sil | bu)

Short-plosive perception

(18.19)

The perception tableau is

[[k˺_^{k}]]	OCP (place: dor; transition | silence | burst)	NCC (place: dor; transition | silence | burst)
dor dor \\ | k˺_ k	*!	
☞ dor /\\ k˺_ k		*

(18.20)

Note that the winner violates a weak NCC constraint, since the silence intervenes between the two dorsal cues. If more material intervenes between the two cues, the OCP constraint will be ranked lower, and NCC higher. Thus, for geminate plosives, with their longer silences, the constraints are closer than they are in (18.19), or may even be reversed. In a language that maintains frequent geminate consonants, we expect that [[ak˺_ː^{k}a]] is perceived with a single dorsal long consonant, just as in (18.20):

[[k˺_ː^{k}]]	OCP (place: dor; trans | long silence | burst)	NCC (place: dor; trans | long silence | burst)
/kk/	*!	
☞ /kː/		*

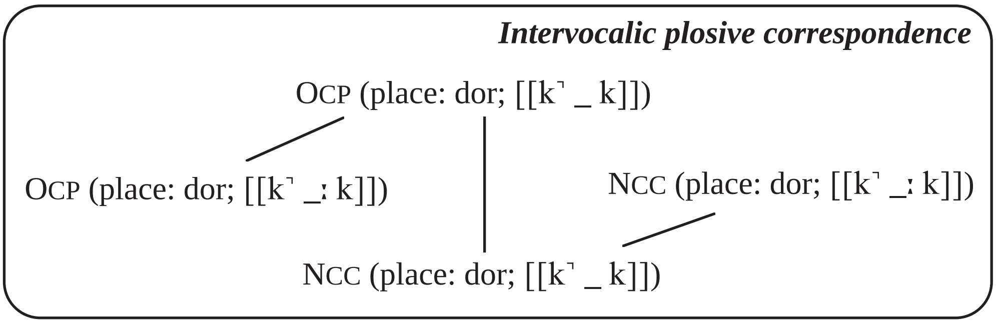
(18.21)

In a language without geminates, the acoustic input [[ak˺_ː^{k}a]] probably tells something about the heteromorphemic descent of the two dorsal place cues. In such a language, the ranking may well be reversed:

[[k˺_ː^{k}]]	NCC (place: dor; trans | long silence | burst)	OCP (place: dor; trans | long silence | burst)
☞ /kk/		*
/kː/	*!	

(18.22)

The following graph shows the (near-) universal rankings of the correspondence constraints in the perception grammar:


Intervocalic plosive correspondence
OCP (place: dor; [[k̚ _ k]])
OCP (place: dor; [[k̚ _ː k]])
NCC (place: dor; [[k̚ _ː k]])
NCC (place: dor; [[k̚ _ k]])

(18.23)

This leaves a binary typology, as exemplified in (18.21) and (18.22).

18.3.3 The influence of the OCP on the production grammar

Whether the listener hears *one* dorsal feature value as in (18.21) or *two* as in (18.22), determines whether or not the perceptual result of the acoustic form [[ak̚_ː$^{\text{k}}$a]] from the underlying form |ak+ka| violates PARSE (place: dorsal). This is the modest influence of the OCP constraint on the production grammar. An asymmetry between heterorganic and homorganic gestures is one of the causes of some “OCP effects”: whereas [ap͡ka] faithfully parses both the labial and the dorsal specifications present in /ap+ka/, the geminate [akˑa] may parse only one of the two dorsal specifications present in /ak+ka/, thus violating PARSE (dorsal): the identity of one of the consonants is lost. The only difference with a rote deletion is the faithful rendering of the timing slot, which may be a sufficient sign of the double specification in languages without tautomorphemic geminates. The following phonetic truth can be stated:

> Adjacent identical gestures may be heard as a single gesture.

It depends on the ranking of PARSE (dorsal) whether the language does anything about the problem. For instance, it could try to produce [[ak̚_$^{\text{kh}}$k̚_$^{\text{k}}$a]], with two temporally separated complete dorsal gestures, but this would need a quite different syllable timing, effectively inserting a boundary normally used for separating intonational phrases. Let’s say that this would violate FILL (]$_{I\,I}$[), probably a strong constraint in all languages. A somewhat less radical solution would be the insertion of a segment between the two /k/, so that /ak+ka/ would end up as /akəka/ or /akska/ or so[2]. That would violate FILL (segment). The following tableau shows six of the most obvious possible outputs of /ak+ka/:

[2] In a theory of Feature Geometry with privative features, neither /akəka/ nor /akska/ seem to satisfy PARSE (dorsal), because the two /k/ are still adjacent on the dorsal tier; we should say, in that case, that /akəka/ satisfies PARSE (root), and /akska/ satisfies PARSE (place).

|ak+ka|	Fill $(]_I\,{}_I[)$	Fill (noise)	Parse (k)	Parse (timing)	*Hold (tongue)	Fill (ə)
[ak_ka] /ak]$_{\varphi}$ $_{\varphi}$[ka/	*					
[akːa] /akːa/ (18.21)			*		*	
[akːa] /akka/ (18.22)					*	
[aka] /aka/			*	*		
[akəka] /akəka/						*
[axka] /axka/		*				

(18.24)

We see that all six candidates are possible winners, depending on the ranking of the seven constraints and the ranking of Ocp and Ncc for long silences. This proposal makes empirical predictions: for instance, we expect that languages with low *Hold constraints, i.e. languages with geminates, will not produce [aka], because [akːa] will always be more harmonic. Also, languages with geminates (18.21), with their Parse violations, will be more liable to choosing one of the epenthesis or dissimilation candidates. In the following sections, we will examine the conditions that give rise to the various outcomes.

18.3.4 Separation: ak_ka

Suppose we have the following levels of prosodic organization: discourse, utterance (*Σ*), intonational phrase (*I*), phonological phrase (*φ*), phonological word (*W*), foot (*F*), syllable (*σ*). The higher constituents may be separated by intonation breaks and several sorts of pauses. Each of these phonological boundary markers should appear on the surface, and we can expect that this is more important for higher constituents: Parse $(]_{\Sigma}\,{}_{\Sigma}[)$ >> Parse $(]_I\,{}_I[)$ >> Parse $(]_{\varphi}\,{}_{\varphi}[)$. For this reason, languages have stronger markers (longer pauses, etc.) for higher constituents. These boundaries should not appear where they do not belong, and we expect that the stronger boundaries are more offending in this respect: Fill $(]_{\Sigma}\,{}_{\Sigma}[)$ >> Fill $(]_I\,{}_I[)$ >> Fill $(]_{\varphi}\,{}_{\varphi}[)$. Also, a certain boundary will be more offending if it occurs in a lower domain:

$$\text{Fill } (]_I\,{}_I[\ /\ [\ \_\ ]_W) >> \text{Fill } (]_I\,{}_I[\ /\ [\ \_\ ]_{\varphi}) >> \text{Fill } (]_I\,{}_I[\ /\ [\ \_\ ]_I) \qquad (18.25)$$

If we assume that the lowest domain that uses pauses as boundary markers, is the intonational phrase, then the insertion of a pause in order to faithfully parse both /k/ root nodes in the word-level concatenation of |ak+ka|, involves a violation of the relatively high-ranked Fill $(]_I\,{}_I[\ /\ [\ \_\ ]_W)$. This explains why languages avoid this situation. We also see why the OCP is often said not to apply across higher prosodic boundaries; across intonational phrases, for instance, [ak_ka] is the usual result.

Nevertheless, there may be situations where pragmatic requirements rerank the constraint system. For instance, if we make up a new compound, which we do not expect the listener to have heard before, we can insert pauses as efficient metalinguistic signals in the utterance itself. On the day that I am writing this[3], I was engaged in the following dialogue with my son (aged 5:0), who had trouble finding the word *wapens* ‘weapons’:

Jelle: “We played at soldiers”.
Paul: “With the pirate flag?”
Jelle: [nei ʔɑlʎeimɛt fˈ˦ɛ˥xt deŋə] “No, only-with *fighting things*.”

The new compound /fˈɛxt deŋə/, pronounced with a fully released /t/, was in contrastive focus and had a clearly unaccented /deŋə/ part which signalled that the preceding pause was not a syntactic prosodic boundary[4]. The relevant constraint ranking is

|ak+ka|	FILL (noise)	PARSE (dorsal)	FILL (syllable)	FILL (]$_{I\,I}$[)	*HOLD (dorsum)	*GESTURE (dorsum)
☞ ak-ka				*		**
[ak͡ka] /aka/		*!				**
[akːa] /aka/		*!			*	*
aka		*!				*
akəka			*!			**
axka	*!					*

(18.26)

All the segmental faithfulness constraints have risen and are undominated. We see that in Jelle’s grammar, [akːa] must violate PARSE (dorsal), although Dutch has no geminates; otherwise, [ak͡ka] (with a double dorsal gesture, see §18.3.5) would always have been a better candidate. This can have two causes. First, Jelle may not have considered the transition a sufficient cue to dorsality, so that he violates PARSE (k) unless he hears two release bursts. Secondly, he may ignore the length in [ak͡ka], so that he ranks OCP as high as for short plosives; this is what I have suggested in (18.26) by marking two perceptions as /aka/.

Another example was reported by McCarthy (1986): in a Moroccan Arabic language game that reverses the consonants in a word, /ħbib/ ‘friend’ becomes /b-biħ/, with a released first /b/ (at least for one speaker; two others have /bħiħ/, reversing the root sequence before Tier Conflation). McCarthy suggests that the phonetic reflex of an OCP violation is exactly as described here:

> “contrast[s] between singly associated and multiply associated (...) geminates (...) would conceivably have transparent phonetic consequences (like medial release for clusters versus medial closed transition for geminates).” (McCarthy 1986: 255)

[3] September 22, 1996.

[4] In Dutch, compounding, in contrast with inflection, has the postlexical trait of dependence on the pragmatics, though the choice of the expression of the juncture is lexically and morphologically governed; here, we have the zero allomorph because the first constituent is a verb stem.

If this is indeed the interpretation of the OCP, it can be identified in the production grammar with FILL (]$_{I\,I}$[/ [_]$_W$), and if the occurrence in the languages of the world is restricted to the exceptional types described in this section, this constraint is nearly universally undominated in normal adult language. Much more interesting is a phonological interpretation of the OCP, namely, that a geminate element always acts as a single element in phonological processes; with this interpretation, the OCP is strong, but not unviolable.

18.3.5 Gemination: akka

With a normal timing of the neural commands for the two dorsal gestures, we get a result that sounds like one prolonged gesture. This may occur in languages without tautomorphemic geminates, as was measured for English by Stetson (1951: 61) and Lehiste, Morton & Tatham (1973). These languages thus have a high *HOLD constraint, and produce the geminate by violating *GESTURE (dorsum) twice:

|ak+ka|	FILL (]$_{I\,I}$[)	*HOLD (dorsum)	PARSE (dorsal)	FILL (noise)	FILL (σ)	*GESTURE (dorsum)
ak-ka	*!					**
☞ [ak͡ka] /akka/						**
[akːa] /akka/		*!				*
aka			*!			*
akəka					*!	**
axka				*!		*

(18.27)

This will occur in English compounding: *stock car*, and in the Dutch adult version of the pragmatically conditioned ‘analysing’ pronunciation of an otherwise unclear compound, as in /eːt/ ‘eat’ + /taːfəl/ ‘table’ → [ˈeːt·aːfəɫ] ‘eating table’ and [[vɛxt˺_·ᵈeŋ]] ‘fighting things’, which we can compare with the child’s [[fɛχt˺_ᵗ˷ᵈeŋ]] of §18.3.4.

18.3.6 Fusion: ak:a

A language with tautomorphemic geminates will presumably implement a geminate not by two overlapping gestures, but by a single prolonged gesture (low *HOLD constraints). This has been measured for Cellefrouin Gallo-Roman (Rousselot 1891: 86), Hungarian (Hegedüs 1959), and Estonian (Lehiste, Morton & Tatham 1973). The ranking of an organizational articulatory trick like a locational hold probably depends on what the speaker can gain from learning it. This is larger as more geminates occur in the language: for every geminate, a dorsal gesture is saved:

|ak+ka|	Fill (]$_{I\,I}$[)	Fill (noise)	Fill (σ)	Parse (timing)	Parse (dorsal)	*Gesture (dorsum)	*Hold (dorsum)
ak-ka	∗!					∗∗	
[ak͡ka] /akːa/					∗	∗∗!	
☞ [akːa] /akːa/					∗	∗	∗
aka				∗!	∗	∗	
akəka			∗!			∗∗	
axka		∗!				∗	

(18.28)

Since in these languages the geminate result violates a Parse constraint, this constraint has to be low-ranked; otherwise, the language will opt for epenthesis.

18.3.7 Degemination: aka

Parse (timing) is the constraint that requires an underlying C slot on the timing tier to appear in the output. If this is ranked lower than both articulatory constraints and all other faithfulness constraints, the result will be deletion of one of the original /k/:

|ak+ka|	Fill (]$_{I\,I}$[)	Fill (noise)	Fill (σ)	*Hold (dorsum)	*Gesture (dorsum)	Parse (dorsal)	Parse (timing)
ak-ka	∗!				∗∗		
[ak͡ka] /akka/					∗∗!		
[akːa] /akka/				∗!	∗		
☞ aka					∗	∗	∗
akəka			∗!		∗∗		
axka		∗!			∗		

(18.29)

Note that if the language hears a geminate as two consonants, the output [aka] can only win if Parse (dorsal) is dominated by the gestural constraints; since intervocalic dorsal consonants will usually surface faithfully, the Parse constraint must be conditioned for the geminate environment. Another possibility is that Ocp (place: dorsal; [[ak˺_ː^{k}a]]) is highly ranked after all, so that the two geminate candidates are perceived as /akːa/, which would violate Parse (dorsal), so that this constraint may be highly ranked.

The result of (18.29) represents the usual action in Dutch lexical phonology. For instance, the past tense of weak verbs is formed with the suffix |-də|. With the stem |mɛld| 'notify', we get /mɛldə/ 'notified'. In the phonology, Dutch appears not to be concerned with preserving the identity of the morpheme; however, the homophony between present- and past-tense forms, which only occurs in the plural forms (/ʋəmɛldə/ 'we notify' or 'we notified'), is circumvented by avoiding the past-tense forms, at least in the spoken language.

On the sentence level, the situation is somewhat different. The following Dutch examples illustrate the relation between the length of a geminate and syntactic constituency[5]:

(a) *We gaan even voor een nieuwe klok kijken* [ɔkːɛi]. "We'll go shopping for a new clock." The prepositional phrase can be extraposed: *We gaan even kijken voor een nieuwe klok.*

(b) *Even op de klok kijken* [ɔkˑɛi]. "I'll look at the clock." Extraposition is hardly allowed: the verb is strongly subcategorized for looking at gauges, with the preposition *op*. *Klok* is still clearly a noun.

(c) *Ze kan nog niet klokkijken* [ɔkɛi]. "She cannot read the time yet." *Klok* is incorporated into the verb, as witnessed by the verbal negation *niet* 'not', as opposed to the nominal negation *geen* 'no' (which is also allowed here).

Cases (a) and (b) have the similarly timed counterparts [pk] (released [p]) and [p͡k] (unreleased [p]); case (c) has no counterpart for heterorganic clusters.

We can now see that a language without geminates does not have the anti-root-node-branching rule (18.4), interpreted with the help of the OCP, but the constraint ranking *HOLD (C) >> PARSE (C). The question, of course, is whether these analyses are empirically different. If *HOLD (C) and PARSE (C) did not interact with other constraints, the situation would be difficult to decide. However, PARSE (C) interacts with *GESTURE: the constraint ranking set *HOLD (tongue body) >> PARSE (C) >> *GESTURE (tongue body) describes a language like English, which does not allow tautomorphemic geminates, but allows geminates across some morpheme boundaries (stratum-2 affixation; compounding). This only works under McCarthy's assumption that tautomorphemic geminates are always inseparable, i.e. that they are phonologically single long consonants.

18.3.8 Epenthesis: akska

Many strands of epenthesis have articulatory causes: satisfaction of synchronization constraints (Dutch [mɛlᵊk], English [tɛnᵗs]). Epenthesis is also possbile in order to satisfy the perceptual distinction between two underlying segments or features that would otherwise merge into a cluster that could be analysed in a different way:

Limburgian. Limburgian inserts an /s/ between nouns stem ending in a velar and the diminutive suffix /-kə/ (a full list is in §12.2). Thus, /dɛ́ŋ-s-kə/ 'thing (dim.)' inserts an /s/ so that it cannot be confused with /dɛŋ̄kə/ 'think'. Of course, it is not this actual word pair that is the problem to be solved; rather, Limburgian listeners like to be

[5] Of course, these data are highly variable.

confident that they can always interpret an [ŋkə] as belonging to a single morph[6]. It may be relevant that the epenthesis originated at a time that the language must still have had geminate consonants.

English. Plurals and past tense: /hʌg-z/ ‘hugs’ but /kɪs-ɪz/ ‘kisses’: vowel insertion between two sibilants; /lɪv-d/ ‘lived’ and /kæn-d/ ‘canned’ but /wɔnt-ɪd/ ‘wanted’ and /niːd-ɪd/ ‘needed’: vowel insertion between two coronal stops. McCarthy’s (1986) generalization of the idea that both /tt/ and /dt/ appear to constitute OCP violations, is that the sequence /dt/ would consist of two identical elements after voicing assimilation. Yip’s (1988) rule of “Coronal Epenthesis” reads thus:

Domain: Coda
Tier: (i) Strident
      (ii) Continuant
Trigger:
Change: Insert

Thus, this rule repairs sequences like /ʃz/ and /td/ in coda, by epenthesizing a vowel that puts the first consonant in an onset position; the sequences must have a specification on the stridency tier (presumably, all coronal obstruents), agree in stridency (to rule out /θs/), and agree in continuancy (to rule out /ðd/). The trigger of the rule is not specified, because it is a universal trigger: the OCP. There are several problems with Yip’s approach.

I would like to express the fact that */ʃz/ and */td/ are morpheme-structure constraints as well: in English morphology, they do not occur until the stratum of compounding[7]. However, Yip has to restrict the domain of the rule to the coda because if the domain were a phonological word, the consonants would still be adjacent on the stridency tier; epenthesizing a vowel, with no specification on the stridency tier, would not make the OCP satisfied on that tier.

Fricatives are often economically divided into *non-strident* (/ɸ/, /θ/, /x/) and *strident* (/f/, /s/, /ʃ/, /χ/). In contrast with what the label suggests, this division is based on distributional grounds: the strident fricatives are louder (make more noise) than their non-strident counterparts on the same articulator, and are, therefore, on the average more suitable for human communication in a world with background noise and distances; the non-strident fricatives, on the other hand, often alternate, or are historically related to, plosives at the same place of articulation; as so happens, plosives tend to occur at locations where perfect closures are easy to make (bilabial, corono-postdental, dorso-

[6] The [ə] may be crucial here. In the morphology of the third person singular of the present tense, basically expressed as the suffixation of a /t/ weakly specified for place, the words /zɛ́ŋ-k/ ‘sings’ and /zɛŋ̄k-∅/ ‘sinks’ can only be distinguished by their different tones, which was not enough for */dɛ́ŋ-kə/ versus /dɛŋ̄k-ə/. Alternatively, we could argue that diminutive formation is at a more superficial level of derivation than verb inflexion, causing different degrees of importance to be attached to the prevention of the two /ŋ+k/ clashes: PARSE (dorsal / _ +*dim*) >> FILL (s) >> PARSE (dorsal / _ +*3sg*). We see yet another possibility when we realize that [dɛ́ŋs.kə] can be syllabified perfectly, whereas *[zɛ́ŋsk] is problematic.

[7] Morphologically, compounding takes place after inflection. The morphological bracketing [[apple] [[pie][s]]] does not necessarily coincide with the semantic bracketing [[[apple][pie]] s].

velar), and fricatives prefer locations with small holes (labio-dental, corono-dental) or unstable structures (dorso-uvular). From the perceptual standpoint, however, we could divide the stridency scale into three levels:

1. **Mellow**: low-frequency noise (“grave”): [ɸ], [x].
2. **Strident**: high-frequency noise ([f], [θ]) or amplitude modulation ([χ]).
3. **Sibilant**: strong noise: [s], [ʃ].

The problem with */fɪʃs/ is the unpronounceability of a boundary between /ʃ/ and /s/: the same articulator is involved and has to move from one shape and position to another; if the gestures overlap, the result is a long sibilant without much temporal structure, so that it will be heard as one; thus, PARSE (sibilant) is violated, if that constraint is considered as an evaluation of the relation between perceptual features in the input and the output. By contrast, the two sibilants in /fɪʃɪz/ ‘fishes’ and /hɔːsɪz/ ‘horses’ are clearly separated in time, and PARSE (sibilant) is not violated.

Thus, though the same articulator is involved in both segments, the problem of their concatenation is of a perceptual nature. We can find some more evidence for this when we look at other weakly contrasting fricatives in English.

First, there is /θs/, which seems not to pose many problems in /klɔθ+s/ ‘cloths’; though both segments use the same articulator, the temporal separation between the [θ] and [s] parts is clear, because the mode of noise generation changes from local (between the teeth) to dipole (smashing a jet of air against the teeth), resulting in a large change of intensity (somewhat smaller for the voiced versions). We find an optional fusion in the morphologically hardly analysable word /klou(ð)z/ ‘clothes’, and an occasional dissimilation in /sikst/ ‘sixth’.

Secondly, there is /fθ/. Though produced on different articulators, these dental fricatives are perceptually very similar. It is no coincidence that we find /fɪft/ ‘fifth’ and /twɛlft/ ‘twelfth’: in this dissimilation, the conflict is between losing the root node and losing the [fricative] specification, which must be underlying because the morpheme ends up as [θ] on other numerals. This is a clear case of perceptually motivated dissimilation; articulator tiers have nothing to do with it.

Another problem in Yip’s analysis, is that, in order to account for */td/, she has to assume that /t/ and /d/ have non-redundant [–strident] specifications. A generalization to “coronal stop” is impossible, because that would include /n/, which is inappropriate in the view of /kæn+d/ ‘canned’, where we find no epenthesis. Reality is simpler again. If there were no epenthesis, the past tenses of /wɔnt/ ‘want’ and /niːd/ ‘need’ would be equal to the present-tense forms (assuming the same voicing assimilation as with other final consonants, and degemination); so we would have a lot of morphologically related homonyms in a language that otherwise expects tense marking: a functionally undesirable situation, which a language may choose to prevent. This situation is much less severe for /n/-final verbs: /fain+d/ ‘fined’ may be homophonous to /faind/ ‘find’, but in practical situations its /-d/ has a much larger chance of being recognized as a tense marker than the same suffix in */niːd/ ‘needed’ would have. This functional explanation is straightforward and not very controversial, and in a functional view of phonology it should be directly formalizable. In */niːd/ ‘needed’, we would have violations of all

PARSE constraints that can be associated with /d/, most notably segmental identity, which we could write as PARSE (root); in /fain+d/, which makes the final cluster undistinguishable from a monomorphemic one, we have a violations of PARSE (coronal) if the relevant OCP is ranked high, but PARSE (timing) and PARSE (nasal & coronal) are satisfied. The intuitive idea that more is lost in /d+d/ → /d/ than in /n+d/ → /nd/, can be formulated as the near-universal ranking PARSE (root) >> PARSE (coronal). This ranking becomes crucial as soon as an interacting constraint is allowed to intervene. In our case, this intervening constraint is a constraint against the insertion of a syllable or a vowel:

|wɔnt+d| |kæn+d|	PARSE (root)	FILL (ɪ)	PARSE (coronal)
wɔnt	*!		*
☞ wɔntɪd		*	
☞ kænd			*
kænɪd		*!	

(18.30)

This analysis seems more straightforward than Yip’s statement that /n/ does not cause an OCP violation because it is not specified for stridency (as opposed to the [–strident] of the plosives).

We can now predict a typology depending on the ranking of the FILL constraint:

(1) FILL is ranked high: no epenthesis; fusion may result. The situation in Dutch past-tense formation (see below).
(2) FILL is ranked in between: epenthesis between homorganic plosives. The situation in English and German (see below) past-tense formation.
(3) FILL is ranked low: epenthesis between all homorganic consonants. The situation in Limburgian diminutive formation (see above).

Hungarian. In the formation of the Hungarian preterite, a mid vowel is inserted after a stem that ends in /t/: /keːr+tɛm/ ‘I asked’ versus /ʃyt+øtːɛm/ ‘I baked’. With a stem in /-d/, the epenthesis occurs only in some verbs.

German. The German past tense is formed like /main+tə/ ‘meant’ and /max+tə/ ‘made’, but coronal plosives insert a schwa: /mɛld+ətə/ ‘reported’. The same ranking as in English.

Typologically, we see that epenthesis is more probable if the two segments are more alike. Thus, epenthesis into English /d+d/ is not joined by epenthesis into /n+d/. In Limburgian, on the other hand, the existence of epenthesis into /ŋ+k/ presupposes epenthesis into /k+k/. This dependence of the ranking of the OCP on perceptual similarity was predicted in §12.3.

18.4 Blocking of rules

Instead of as a rule *trigger*, the interaction of OCP with PARSE can also work as a rule *blocker*. With output-oriented constraints, the distinction between triggering and blocking often becomes meaningless.

18.4.1 Elision blocked

Strongly related to the phenomenon of epenthesis (in fact, often indistinguishable from it), is the idea that the force of clash prevention can cause exceptions to an otherwise general rule of elision:

Malayalam. Vowels can be elided between consonants, in casual speech, but not if that would produce a geminate (Mohanan 1986: 168):

waːkate ‘a type of tree’s’ → waːkte
kaʈaʈe ‘shop’s’ → kaʈaʈe

It is crucial here that geminates are very common in Malayalam, so that there would be a large parsing problem if spurious geminates would surface. For instance, if |kaʈaʈe| is pronounced as [kaʈʈe], it will be perceived as /kaʈːe/, violating PARSE (root / C). The grammar is something like PARSE (root / C) >> *SYLL >> PARSE (root / V).

Bärndütsch. The article /di/ ‘the’ surfaces as [d], except before /d/:

d-gægəd ‘the neighbourhood’
di-dœrfər ‘the villages’

Tiberian Hebrew. (McCarthy 1986): schwa deletion is blocked between identical consonants:

zaːχəruː → zaːχruː ‘they recalled’
saːβəβuː ‘they surrounded’ (from the root /sb/, after Tier Conflation)

English. /n/ engages in degemination: the fricative /s/ can be followed by syllabic /n/, which may come from the factitive morpheme: /lɪsn̩/ ‘listen’, /fɑːsn̩/ ‘fasten’. After coronal plosives, the morpheme vacillates between a syllabic nasal and /ən/: /flætən/ or /flætn̩/ ‘flatten’, /mædən/ or /mædn̩/ ‘madden’. After /n/, we find no syllabic nasals: /lɪnɪn/ ‘linen’. Again, syncope is more likely if the perceptual identity of the segments is preserved better.

18.4.2 Assimilation blocked

Another implementation of clash prevention at morpheme boundaries, is the refusal to do the usual assimilation:

Limburgian. Postlexical place assimilation of nasals, though uncommon, is thinkable, but in diminutive formation, it is out of the question: the diminutive of /mɑn̄/ ‘man’ is /mǽnkə/, and that of /bɛ̄n/ ‘leg’ is /bɛ́inkə/. Apparently, PARSE (place: dorsal / kə) is ranked quite high, as we already saw with /s/ insertion between it and a base ending in a dorsal consonant (§12.2): the diminutive suffix must stand out from the base.

Malayalam. According to Mohanan (1995), we find /a n-pə/ ‘kindness’, which constitutes an exception to a postlexical rule that assimilates the place of nasals to the following plosive.

According to Mohanan & Mohanan (1984), nasals assimilate to the following plosives, but not to nasals, which would give geminates:

b^{ə}aːlan+poːdʒ^{ə}i → b^{ə}aːlamboːdʒ^{ə}i ‘the boy went’
n̪an-ma ‘goodness’

This situation can be described with the grammar PARSE (nasal) >> *GESTURE >> PARSE (place). Note that */namːa/ would violate PARSE (nasal).

18.4.3 Sound change blocked

The case of English /fɪft/ and /sɪkst/, which we saw in §18.3.8, may well continue an exception to the Early Germanic spirantization of voiceless plosives: while an Indo-European /t/ became /θ/ in Germanic (compare Latin /treːs/ ‘three’ with English /θriː/), this change did not occur after a (old or new) fricative (Latin /staːre/ ‘stand’ vs. English /stænd/; Latin /nokte/ ‘night’ vs. Dutch /nɑxt/). If OCP (noise) is ranked high enough (for no intervening material), the sequences [sθ] and [xθ] would be perceived with a single value on the noise tier, which may lead to perceiving a single value on the place tier, violating PARSE (coronal).

18.5 Articulatory dissimilation

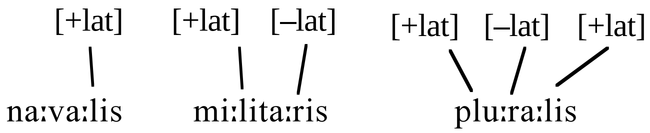
All the cases of §18.3 and §18.4 avoided a perceptual problem. We shall now see two cases of dissimilation at a distance, commonly ascribed to the OCP.

Dissimilation of [lateral] in Latin. The Latin suffix |-aːlis|, which produces adjectives, turns up as /-aːris/ if the base contains an /l/, except if that /l/ is followed by an /r/ (Passy 1891: 201):

naːv+aːlis	‘naval’
miːlit+aːris	‘military’
pluːr+aːlis	‘plural’

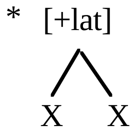
To account for this, we could say that /l/ and /r/ are ***lateral-bearing*** segments, and state the allomorphy rule as follows, without the OCP: “within a single word, there cannot be two adjacent lateral segments on the partial segment sequence that consists of all lateral-bearing segments”. Thus, in our three cases these partial sequences would be /l/, /lr/, and /lrl/, respectively, and */miːlitaːlis/ would be ruled out because its partial lateral-bearing sequence would be the offensive */ll/.

Though descriptively adequate, the above formulation of the rule does not explain the phenomenon in terms of any fundamental principle. The Obligatory Contour Principle may come to the rescue, since this principle has repeatedly been used for the “explanation” of all sorts of dissimilatory phenomena. First, we re-represent the three words in an autosegmental notation, as follows:



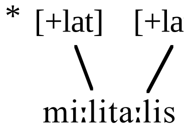
(18.31)

With the OCP, the rule would now be restated thus: “within a single word, [+lateral] cannot be multiply linked” (i.e., linked to more than one segment), or: “[+lateral] cannot branch”:



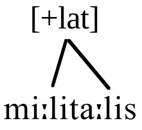
(18.32)

Now, the OCP helps to get the right result, because



(18.33)

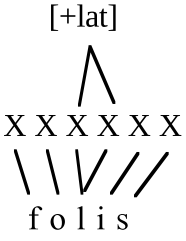
is an illicit representation ruled out by the OCP itself, and



(18.34)

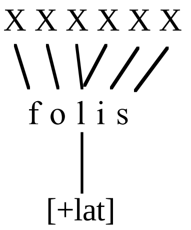
is ruled out by the Latin “don’t branch” rule (18.32). Crucial in this example is that the feature value [–lateral] is linked to any non-lateral lateral-bearing segment (here the /r/), otherwise the two *l*’s in *pluralis* would be adjacent on the [lateral]-tier and thus would have to be linked, because of the OCP, to the same [+lateral] specification. In other words, [–lateral] is phonologically active here.

Still, however, nothing has been explained: the OCP only enabled us to state the rule in two words. Fortunately, the “don’t branch” rule is even descriptively wrong: Latin has geminate laterals. The word /fol:is/ ‘leaf’, for instance, must be represented, according to the OCP, as



(18.35)

The solution, in this case, is the correct linking of [lat] to the root node:

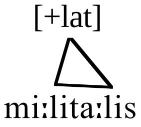


(18.36)

So there is no branching [+lat]. This procedure allows us to differentiate between repetition and lenghtening of a gesture: these processes are phonetically very different, and I cannot imagine that this difference would not be reflected in phonological processes. Nevertheless, in her analysis of Seri (Marlett & Stemberger 1983), Yip (1988) attributes both the rule /ʔaʔ/ → /ʔaː/ and the non-occurrence of the long glottal stop */ʔː/ to the same prohibition on branching glottal stops.

Going back to the *function* of the phenomenon may yield us genuine explanations. Certainly, we cannot take refuge to the dissimilarity-by-hypercorrection hypothesis by Ohala (1993a), because we are talking about transparent morphological alternations here.

The problem in */miːlitaːlis/ seems to be the repetition of the lateral gesture (or the lateral feature). But the fact that a /r/ would break the ban, suggests that the really offensive implementation is



(18.37)

Here, the lateral gesture would continue throughout the /litaːl/ sequence. This is a possible articulation, but the perceptual result, a violation of FILL (lateral & vowel), appears so offensive that /miːlitaːris/, violating /l/ → [+lateral], is a better candidate. In /pluːraːlis/, we find the strong specification /r/ → [trill]; the trill is *not* compatible with the labial gesture and thereby prevents the coalescence of the two lateral gestures.

There is a strange coincidence here: the allomorphy /-aːlis/ → /-aːris/ proves that /l/ and /r/ form a natural class (they must have something in common perceptually). On the other hand, /r/ breaks up laterality, so it must be articulatorily incompatible with /l/, and perhaps articulatorily very different.

An analysis with a privative [lateral] feature was proposed by Kenstowicz (1993). In /floːr+aːlis/, the delinking of [lat] is blocked because /floːraːris/ would violate the OCP on the [rhotic] tier. Steriade (1995) rephrases this in OT terms as

PARSE (rhotic) >> OCP (rhotic) >> OCP (lateral) >> PARSE (lateral) (18.38)

This elegant solution does not work. Besides the fact that (18.38) may not be general (Latin allows /kelː-ula/ 'small room' and /reːgaːl-iolus/ 'wren'), it would also predict that any /r/ in the base would cause selection of the /l/ allomorph. Thus, we would expect */reːgul-aːlis/ instead of /reːgul-aːris/ 'regular', because the latter would violate OCP (rhotic) on the privative [rhotic] tier. Unfortunately, in this example we may have transparently /reːg-ula/ as a base; monomorphemic /r-l/ sequences cannot be found, because of a restriction against two liquids in Indo-European roots[8]. We cannot ask the

[8] /flor-/ and /plur-/ derive historically from intervocalic /s/.

speakers of the language that seems to continue Latin phonology as its stratum I, as they seem to have reduced the maximum distance of this rule to the preceding syllable: witness *velar* and *alveolar*, but *palatal* and *laminal* (beside *laminar flow*); if Dutch *prullaria* 'knick-knacks' from *prul* 'piece of trash' + Lat. *a:lis* (neuter plural) were a Latin word, we would almost have disproved Kenstowicz's proposal.

Dissimilation of [voiced] in Japanese. In Japanese, there cannot be more than one voiced obstruent in the expression of one originally Japanese word stem. Thus, *gado gado* is not a Japanese meal. This constraint actively prevents the usual voicing of an obstruent in the morphological operation of *compounding*. Thus, while /ori/ 'fold' + /kami/ 'paper' gives /origami/, the same process is prohibited in /onna/ 'woman' + /kotoba/ 'word', which gives /onnakotoba/ 'women's speech' because */gotoba/ is ill-formed.

Though Itô & Mester (1986) invoke the OCP, the phenomenon resembles the Latin data given above. The difference is that the /t/ in /kotoba/ does not seem to license two separate [+voice] features: we cannot say that the two voiced obstruents are adjacent, if we define adjacency as follows (also Myers 1994 and Archangeli & Pulleyblank 1994): "two features are adjacent on a tier if there are no intervening feature-bearing units". Now, /t/ is obviously a voice-bearing unit, so /g/ and /b/ would not be adjacent in */gotoba/. Gesturally speaking: whatever articulatory trick causes the two obstruents to be voiced, that trick is not used on /t/, so that we have two separate gestures. Apparently, the problem here is a genuine problem of repetition, of articulatory or perceptual nature (for once, we do not know yet).

A simple answer to our problems is a *REPEAT(*f*) constraint. Whether *f* is an articulatory gesture or a perceptual feature, remains to be seen. The *REPEAT constraint works on a language-particular basis, and on selected domains: in Latin, we have *REPEAT (lateral) on the "base + *a:lis*" domain; in Japanese, we have *REPEAT (voice) on the domain of the Yamato morpheme. The question of the different behaviour of /r/ in Latin and [–voice] in Japanese, must be put aside for the moment.

18.6 Conclusion

Most alleged OCP effects can be reduced to interactions of more fundamental PARSE and *GESTURE constraints, under the influence of two families of correspondence constraints in the perception grammar, which I called OCP and NCC. The near universality of some OCP effects is due to the high ranking of a constraint against inserting pauses. Typological predictions can be made on the basis of the dependence of the perceptual OCP and NCC constraints on the perceptual similarity of the two segments involved, on the distance between them, and on the probability that the two segments will be heard as one when adjacent, which again depends on the frequency of the occurrence of the sequence in the language.

Long-distance "OCP effects" must be ascribed to the workings of an anti-repetition constraint. *REPEAT militates against long gestures across intervening segments. It can be reset by an interrupting conflicting gesture, required by a perceptual specification.

The OCP, to sum up, is not an autosegmental primitive, and does not have to be described as a separate innate phonological device.

19 Spreading[1]

Abstract. The occurrence of and the restrictions on the temporal spreading of phonological feature values (assimilation, harmony) are the results of interactions between the functional principles of minimizing articulatory effort and minimizing perceptual confusion. This proposal is tested on the typology of opacity to nasal spreading. While the sonority approach of Gnanadesikan (1995) meets with insuperable problems with regard to the position of /h/ in the hierarchy, and the feature-geometric representational approach of Piggott (1992) needs to take recourse to ad-hoc conditions in UG in order to get the hierarchy right, the functional approach accurately predicts the attested typology.

19.1 The functional approach to spreading

We can distinguish several fundamental functional principles, all of which can lead to the phenomenon of feature or gesture spreading.

19.1.1 Limiting the perceptual loss of an articulatory deletion

The Dutch words[2] |aːn| 'on' and |pɑsə| 'fit' concatenate as [aːmpɑsə] 'adapt'. Compared to the alternative [aːnpɑsə], the assimilated form saves us a complete closing-and-opening gesture of the tongue blade. Apparently, Dutch language users value this gain higher than the perceptual loss of replacing the perceptual [place: coronal] specification of |aːn| with a surfacing [place: labial] feature in /aːmpɑsə/, at least for a nasal consonant in the first position of a consonant cluster. In constraint language, the ranking of *GESTURE (tongue blade: close & open) above *REPLACE (place: coronal, labial / nasal / _ C) forces the deletion of the tongue-blade gesture.

The labiality of /m/ in /aːmpɑsə/ must have come about by the ***spreading*** (in this case, lengthening) of the closing-and-opening gesture of the lips: while the hold phase (closed lips) would be short in [aːnpɑsə], as in [pɑsə], it must be somewhat longer in [aːmpɑsə], approximately adding the durations of the lip closures of a [m] in coda and a [p] in onset. This spreading is forced by a perceptual requirement, namely the perceptual specification of simultaneous nasality and consonantality (or non-orality, see ch. 12). After all, if we just leave out the tongue-blade gesture without adjusting the lip gesture, the result would be /aãpɑsə/, with a vocalic (or oral) nasal. Apparently, a path constraint like *REPLACE (nasal × oral: +nasal & –oral, +nasal & +oral) is undominated[3]. In a short notation, the relevant evaluation reads:

[1] This chapter appeared on Rutgers Optimality Archive as Boersma (1998a).

[2] As before (ch. 6), I write underlying perceptual specifications between pipes, articulatory implementations between square brackets, and perceptual results between slashes.

[3] I would like to use terminology that is unbiassed with respect to the oral/nasal distinction, i.e., I would regard [p] and [a] as oral and non-nasal, [m] as nasal and non-oral, and [ã] as oral *and* nasal. The traditional term for this interpretation of 'oral' is 'continuant': an unfortunate leftover from the age of

|an+p|	*GESTURE (blade)	*DELETE (coronal)	*INSERT (nasal & oral)	*INSERT (nasal & labial)
[anp] /anp/	*!			
[aãp] /aãp/		*	*!	
[an͡mp] /amp/	*!	(*)		(*)
☞ [amp] /amp/		*		*

(19.1)

Note that the process /an+p/ → [amp] crucially involves *both* spreading and deletion: if we spread without deletion, we incur a perceptual loss without any articulatory gain; if we delete without spreading, the perceptual loss will not outweigh the articulatory gain. The Optimality-Theoretic approach serves us well in the evaluation of this kind of tunnelling processes.

The general function of this kind of spreading is that it limits the perceptual loss associated with the deletion of an articulatory gesture: in itself, the spreading gesture (lip closure) is unrelated to the lost gesture (tongue blade). This phenomenon of the correlation between labial spreading and coronal deletion is one of the reasons why the concept of *place node* has been advanced in theories of feature geometry (Clements 1985, Sagey 1986, McCarthy 1988, Clements & Hume 1995): the process described here would then be “explained” as “spreading of the place node”.

But there is no articulatory reason why the three articulators should act as a group: they can be moved independently from each other. The attested common behaviour must be caused by the perceptual specification of a nasal consonant: the only thing common to the lip, blade, and body closures, is that we can use any of them to implement faithfully the perceptual feature combination [nasal & not oral]: as long as there is a constriction anywhere in the mouth, the listener will hear the acoustic characteristics of an airstream that travels exclusively through the nose.

So **there is no place node**: the learner does not need such an innate feature grouping to learn that to realize a nasal consonant, she can choose any articulatory gestures [lips: closed], [blade: closed], and [body: closed].

19.1.2 Reducing articulatory synchronization

The perceptual specification |an| is a shorthand for:

binarism, when it had to perform the multiple roles of distinguishing fricatives from plosives, and nasal consonants from nasalized vowels.

Specify:	|a|	|n|
coronal		+
voice	voiced	voiced
noise	–	–
F1	max	
round	–	–
nasal	–	+
oral	+	–

(19.2)

An isolated |a| can fairly easily be realized as [a] (closed velum, wide tongue), and heard faithfully as /a/; an isolated |n| can equally easily be pronounced as [n] and heard as /n/. A faithful implementation of the concatenated |an|, however, requires two articulatory contours at the transition between the two sounds: an opening of the velopharyngeal port and an alveolar closing of the tongue blade. There are three possibilities for the relative timing of these contours. First, the nasal gesture may occur before the coronal gesture:

Articulate:

<table>
<tr><td>velum</td><td>closed</td><td colspan="2">open</td></tr>
<tr><td>tongue</td><td colspan="2">wide</td><td>closed</td></tr>
</table>

Perceive:

<table>
<tr><td>coronal</td><td colspan="2">trans</td><td>side</td></tr>
<tr><td>voice</td><td colspan="3">voiced</td></tr>
<tr><td>nasal</td><td>–</td><td colspan="2">+</td></tr>
<tr><td>oral</td><td colspan="2">+</td><td>–</td></tr>
<tr><td></td><td>a</td><td>ã</td><td>n</td></tr>
</table>

(19.3)

The value *side* for the feature [coronal] refers to the oral side branch between the velum and the coronal constriction; this branch causes a *zero* (depression) in the frequency spectrum, and the length of this branch puts a minor cue to the place of constriction into the location of this zero (which the visual cue of closed lips can easily override: a stationary nasal sound pronounced with closed tongue tip and closed lips will sound like /n/ only in the dark).

The output /aãn/ is quite faithful to the input: all specified features appear, and nothing is heard that was not in the input. Autosegmentally, the correspondence is perfect. Segmentally, of course, there is the misalignment of the left edges of [+nasal] and [–oral]. We can solve this problem by synchronizing the two gestures:

Articulate:

velum	closed	open
tongue	wide	closed

Perceive:

coronal	tr	side	
voice	voiced		
nasal	–	+	
oral	+	–	
	a	n	(19.4)

Perfectly faithful this time, but it violates a synchronization constraint. The third possibility is to put the coronal gesture before the nasal gesture:

Articulate:

velum	closed		open
tongue	wide	closed	

Perceive:

coronal	tr		side	
voice	voiced	unvoiced	voiced	
nasal	–		bu +	
oral	+	–		
	a t˺	_	q$^{\text{N}}$ n	(19.5)

This produces the terrible /at˺_q$^{\text{N}}$n/ (for want of a better notation, I represent the nasal release burst by /q$^{\text{N}}$/; /_ / means silence). Apart from the intrusion of a nasal burst, there may be a voiceless silence in the middle, though the result /ad˺ ˷ɢ$^{\text{N}}$n/ (broadly /a$^{\text{d}}$n/) is, depending on the glottal configuration, also a possible, though hardly less problematic, output (/˷/ stands for the sound of the vocal-fold vibrations radiated out through the vocal-tract walls).

The cross-linguistically favoured candidate will come as no surprise:

|an|	*DELETE (anything)	*INSERT (nasal burst)	*SYNC (velum: close, blade: close)	*INSERT (nasal & oral)	
☞ aãn				*	
an			*!		
at˺ _ q$^{\text{N}}$n		*!			(19.6)

Most languages seem quite willing to incur this minor violation of segmental integrity. The low ranking of the path constraint expresses the importance of the autosegmental approach.

To find out how far nasality spreads into the vowel (§19.1.8), we must first know with what precision the velar and coronal gestures are synchronized. The ranking of the synchronization constraint depends on this precision: synchronizing the two gestures within 20 milliseconds is more difficult than synchronizing them within 40 ms. If we can describe the realized timing difference with a Gaussian distribution, we can represent the imprecision as a standard deviation σ, expressed in seconds, and the universal ranking is

$$\text{*SYNC (velum, blade / } \sigma < x_1) \gg \text{*SYNC (velum, blade / } \sigma < x_2) \Leftrightarrow x_1 < x_2 \tag{19.7}$$

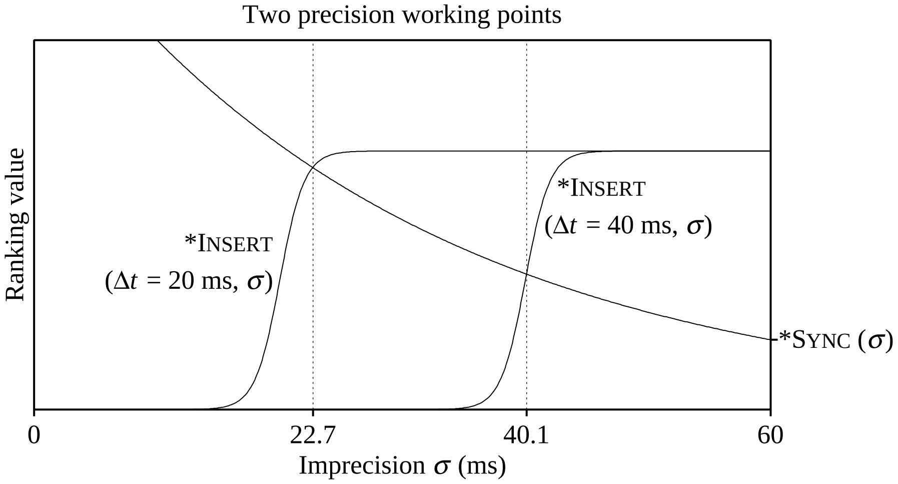
Likewise, the ranking of *INSERT (nasal burst) depends on the probability that a nasal burst is generated. This probability depends on the intended timing difference Δt between the velar and coronal gestures and on the imprecision σ with which this timing difference is implemented:

$$probability(\Delta t, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\Delta t}^{\infty} e^{-y^2/2\sigma^2} dy \tag{19.8}$$

This leads to the universal local rankings

$$\begin{gathered}\text{*INSERT (nas bu / } \Delta t \text{ / } \sigma = x_1) \gg \text{*INSERT (nas bu / } \Delta t \text{ / } \sigma = x_2) \Leftrightarrow x_1 > x_2 \\ \text{*INSERT (nas bu / } \Delta t = x_1 \text{ / } \sigma) \gg \text{*INSERT (nas bu / } \Delta t = x_2 \text{ / } \sigma) \Leftrightarrow x_1 < x_2\end{gathered} \tag{19.9}$$

The rankings of *SYNC and *INSERT are monotonically decreasing and increasing functions of the imprecision, respectively. For a given timing difference, this leads to the emergence of a working point (cf. figure 10.3):



(19.10)

In this example, a timing difference of 20 or 40 ms leads to a working point of 22.7 or 40.1 ms, respectively. We can see all three local rankings in the figure.

In reality, the ranking of *INSERT will not depend on any probabilities. Instead, its ranking will be determined by the number of times it is violated or not during the learning process (chapter 15).

19.1.3 Strong specifications spill over to weakly specified segments

The [+front] (i.e. maximum F_2) specification of |ɛ| in the English word |tɛns| ‘tense’ is implemented by keeping both the tongue body and the lips in non-neutral positions (fronted and spread, respectively) throughout the duration of [ɛ]. In constraint language, the faithfulness constraint *DELETE (+front / vowel) must dominate an articulatory constraint like *GESTURE (lips: spread). This *DELETE constraint is indeed expected to be ranked high, since the replacement of a high F_2 by a low F_2 would make a large acoustic difference for a vowel, and this would be expected to give a large *perceptual* difference as well. In fact, the perceptual difference between a front and a back vowel is large enough that English uses it to support meaning contrasts; in constraint language, the faithfulness constraints for the perceptual feature [front] are ranked so high (for stressed vowels) that any underlying [front] contrast reaches the surface.

The faithful implementation of [front] for a vowel comes with a cost. If lip spreading is fully realized during all of the vocalic opening phase, the gesture of returning the lips to their neutral position must occur after the vowel, i.e. during [n] or [s]. This will have an acoustic effect on the consonant. For instance, at least the first part of the /s/ in /maus/ ‘mouse’ will sound differently from the first part of the /mais/ ‘mice’. However, the acoustic difference between a rounded [s̹] and a spread [s̜] is much smaller than that between [ɛ] and [ɔ], so that the speaker will be understood much easier if she varies the lip shape of a sibilant fricative than if she varies the lip shape of a mid vowel. In constraint language, *INSERT (+front / sibilant) is ranked so low that the lip spreading needed to implement the perceptual place of a neighbouring vowel is allowed to extend well into the fricative; the general lowness of rounding faithfulness for consonants also leads English to not lexically contrasting rounded and spread fricatives.

19.1.4 Limiting the duration of an articulatory gesture

In English, the articulatory realization of a vowel seems to be governed by a scheme of “there and back again”: the [ɛ] in [t^{h}ɛɛ̃n_^{t}s] ‘tense’ tends to be realized as movements away from the neutral tongue-body and lip positions during the closure of [t], and as movements back to the neutral position during [s] or so. Apparently, this language likes to spend an articulatory gesture in order to return to the less fatiguing neutral position. In constraint language, we start from the four-parameter constraint family *GESTURE (lips: spread / *duration* / *precision* / *distance* / *velocity*), isolate the duration parameter, rename the resulting family for clarity to *HOLD, and realize that we must have a universal ranking within this continuous family exemplified by *HOLD (lips: spread / long) >> *HOLD (lips: spread / short).

If, as seems the case in English, duration is a strong determinant of articulatory effort, the *HOLD family will limit the amount of the spreading of the lip gestures that help implementing the place specifications of the neighbouring vowels. Now, vowel specifications are universally weaker in unstressed than in stressed syllables, since confusion probabilities are greater in unstressed syllables (ch. 10). If vowel faithfulness is *very* weak in unstressed syllables, and duration is a strong effort cue, unstressed vowels will tend to have a neutral position of the articulators. For instance, adding the unstressed comparative morpheme to |tɛns| yields [t^{h}ɛɛ̃n_^{t}sə] ‘tenser’.

For the comparative morpheme, of course, we cannot reconstruct any underlying non-neutral vowel quality. But English shows alternations between full vowels and /ə/, as in /prˈoutɛst/ ‘protest (noun)’ versus /prətˈɛst/ ‘protest (verb)’, from which we can posit a common underlying form |proutɛst|. The two surface forms prove the strong specification of vowel quality in stressed syllables, and its weak specification in pre-stress position:[4]

|proutɛst| + initial stress	PARSE (place / stress)	*GESTURE (lips: round)	PARSE (place / pre-stress)
☞ prˈoutɛst		*	
prˈətɛst	*!		

(19.11)

|proutɛst| + final stress	PARSE (place / stress)	*GESTURE (lips: round)	PARSE (place / pre-stress)
proutˈɛst		*!	
☞ prətˈɛst			*

(19.12)

Most crucially, however, the constraint *GESTURE (lips: round) depends on the duration of the lip closure, as we can see in the evaluation of /prəlˈɔŋ/ ‘prolong’:

|proulɔŋ| + final stress	PARSE (place / stress)	*GESTURE (lips: round / long)	*GESTURE (lips: round / short)	PARSE (place / pre-stress)
proulˈɔŋ		*!		
☞ prəlˈɔŋ			*	*
prəlˈəŋ	*!			

(19.13)

If the constraint *GESTURE (lips: round) had not depended on duration, the result would have been */proulˈɔŋ/.

[4] An alternative analysis would have that the effort needed to produce place information is greater in pre-stress than in stressed position, because pre-stress syllables are much shorter. The dependence of *GESTURE on the resulting velocity differences would be able to produce the attested asymmetry. However, a still more realistic account would describe the interplay between two continuous families: *GESTURE as a function of velocity, and MINIMUM (F_2) as a function of the realized F_2. The result would be the intersections of these two functions (see ch. 10); however, if the two functions do not intersect, i.e., if the minimum effort of lip spreading (namely, the organizational effort of the neural command) is greater than the maximum acoustic loss of place information (namely, the replacement of a full [ɔ] with a completely neutral [ə]) in unstressed position, the result would plainly be [ə].

19.1.5 Reducing the number of articulatory contours

We could imagine languages where the lip closing-and-opening gesture is divided into two separate gestures: a closing and an opening gesture. Constraints for such gestures have no *duration* parameter, so their general form is something like *MOVE (*articulator*: from *a* to *b* / *precision* / *velocity*). For lip rounding, we would have *MOVE (lips: from neutral to round) and *MOVE (lips: from round to neutral).

If the *MOVE constraints are separate, there must also be a separate *HOLD (*articulator*: *position* / *duration*) constraint, for instance *HOLD (lips: round / long). Note that this is different from our earlier *GESTURE (lips: round / long), which includes the actual closing and opening movements.

If *HOLD dominates *MOVE, we tend to have short combinations of closing and opening gestures, and these are likely to be incorporated organizationally into a single gesture, as described earlier. If *MOVE dominates *HOLD, however, the articulator tends to stay in its position until stronger constraints force it to move.

For instance, consider the Hungarian dative suffix |nɛk|. Its |ɛ| may be specified as [front], judging from the form /nɛkɛm/ ‘to me’. But since affixes are usually less strongly specified for their features than stems, beause of their lesser semantic content, the [front] specification of |ɛ| is weaker than that of the stem that it is added to. If *MOVE is highly ranked, the form |fɔl+nɛk| ‘wall+DAT’ will surface as /fɔlnɔk/:

|fɔl+nɛk|	*MOVE (tongue)	*REPLACE (place / stem)	*REPLACE (place / suffix)	*HOLD (tongue)
fɔlnɛk	*!			*
☞ fɔlnɔk			*	*
fɛlnɛk		*!		*

(19.14)

Thus, the principle of the minimization of effort lets us either limit or spread articulatory gestures. The limitation comes from high *HOLD constraints or from the universal dependence of *GESTURE on duration, which minimize energy expenditure; the spreading comes from high *MOVE constraints, which minimize the organizational effort, i.e. the number of muscle contours.

19.1.6 Limiting harmony

The spreading of an articulatory gesture, forced by *MOVE, can only extend so far until it reaches a perceptual specification that is stronger than the *MOVE constraint. For instance, leftward spreading of the articulatory gesture of velum lowering (a form of ***nasal harmony***) is blocked in some languages by the first obstruent encountered. This is not because obstruents are specified as [–nasal] in these languages, but because they are specified for the perceptual feature [plosive] or [fricative], which means that a release burst or friction noise should be audible during these segments. The high pressure drop across the constriction, needed for release bursts or friction noise to arise, is hard to attain if the velopharyngeal port is open. So, strong perceptual specifications can block spreading.

For instance, consider the rightward spreading of the velum-lowering gesture in Warao (Osborn 1966):

|mojo| ‘cormorant’	*MOVE (velum)	*INSERT (nasal / j)	*INSERT (nasal / o)
mojo	*!		
mõjo	*!		*
☞ mõj̃õ		*!	**

(19.15)

Apparently, Warao does not consider it very (perceptually) offensive to nasalize a glide or a vowel. This is relatively natural: under nasalization, a glide is still a glide, and a vowel is still a vowel, so that their main perceptual specifications are honoured in the output. On the other hand, Warao spreading is blocked by a plosive:

|mehokohi| ‘shadow’	*DELETE (plosive)	*MOVE (velum / σ _)	*MOVE (velum / σσ _)	*MOVE (velum / σσσ _)	*INSERT (nasal / h)	*INSERT (nasal / V)
mẽhokohi		*!				*
☞ mẽh̃õkohi			*		*	**
mẽh̃õŋõhi	*!			*	*	***
mẽh̃õŋõh̃ĩ	*!				**	****
mẽh̃õkõh̃ĩ			**!		**	****

(19.16)

Apparently, Warao does consider it quite offensive to nasalize a plosive. Again, this is relatively natural: under nasalization, a plosive becomes a nasal stop, so that its main perceptual specifications (silence and release burst) are violated. Note that the spreading must be implemented with a ***family*** of *MOVE constraints, crucially ranked by the moment of the gesture, thus expressing the strategy “move the velum up as late as possible”, which is one of the possible local strategies for globally minimizing the number of gestures (on the utterance level); if there had been a single *MOVE constraint, the candidate /mehokohi/ would have been the best candidate (of those shown here), and the plosive would throw its shadow leftward all the way to /m/.

Thus, perceptual features can block the spreading of an articulatory gesture. The spreading will not proceed beyond the block, because that would require a second articulatory gesture. In tableau (19.16), this is shown (schematically) by the double violation at the candidate /mẽh̃õkõh̃ĩ/. Thus, this kind of articulatory spreading often shows *opacity* effects.

19.1.7 Spreading of perceptual features

The spreading of ***perceptual*** features would reduce the perceptual salience within the utterance (if this were defined as the number of perceptual contours) and the perceptual contrast between utterances, without decreasing articulatory effort. So there are a lot of arguments against it, and languages use it much less than articulatory spreading. For instance, it is not probable that [ps] will become [fs] (the feature *fricative*), or that [ɔti] will become [oti] (the feature *vowel height*). We expect spreading of degree-of-constriction features only if the participants use the same articulator, i.e., we do expect [zn] to become [dn] and [ɛti] to become [eti].

However, there is also one argument in favour of perceptually motivated ‘spreading’: it could improve the probability of recognition of the feature, as hinted at in §19.1.3. This phenomenon would be associated with stem-affix vowel harmony, whole-word domains, etc. (the F-domain of Cole & Kisseberth 1994). The acoustic-faithfulness constraint MAXIMUM (*x*) which says that a feature specified for its maximum value should be realized with a value greater than *x*, has an analogue in LONG (*feature*: *value*, *t*): “a feature specified for the value *v* is heard at least as long as the period *t*”, with a universal ranking of LONG (*f*: *v*, *t*) >> LONG (*f*: *v*, *u*) ⇔ *t* < *u*. For Hungarian (19.14), the result would be the same as with articulatory spreading:

|fɔl+nɛk|	LONG (place: back, σ)	LONG (place: back, σσ)	*REPLACE (place / stem)	*REPLACE (place / suffix)
fɔlnɛk		*!		
☞ fɔlnɔk				*
fɛlnɛk			*!	

(19.17)

But it is not spreading (as Cole & Kisseberth note). ‘Transparent’ segments with incompatible articulations are expected, not ‘opaque’ ones, as we see from an example of Guarani (Rivas 1974):

|tupa [nas]|	*DELETE (plosive)	LONG (place: back, σ)	LONG (place: back, σσ)	*MOVE (velum)
tupã			*!	*
tũpa			*!	**
☞ tũpã				***
tũmã	*!			*

(19.18)

We see that |p| is transparent to nasal ‘spreading’; the winning candidate has the most velar movements of all, quite contrary to the winners in articulatory spreading sytems like Warao. Plosives are transparent to the spreading of [+nasal] but are still pronounced as plosives. Analogously to the situation in most other languages, where nasality can be seen as superposed on an oral string and implemented with a [lowered velum] gesture, these harmony systems may consider orality (in half of their morphemes) as being superposed on a nasal string and implemented with a [raised velum] gesture, i.e. /tũpã/ is the mirror image of /muna/.

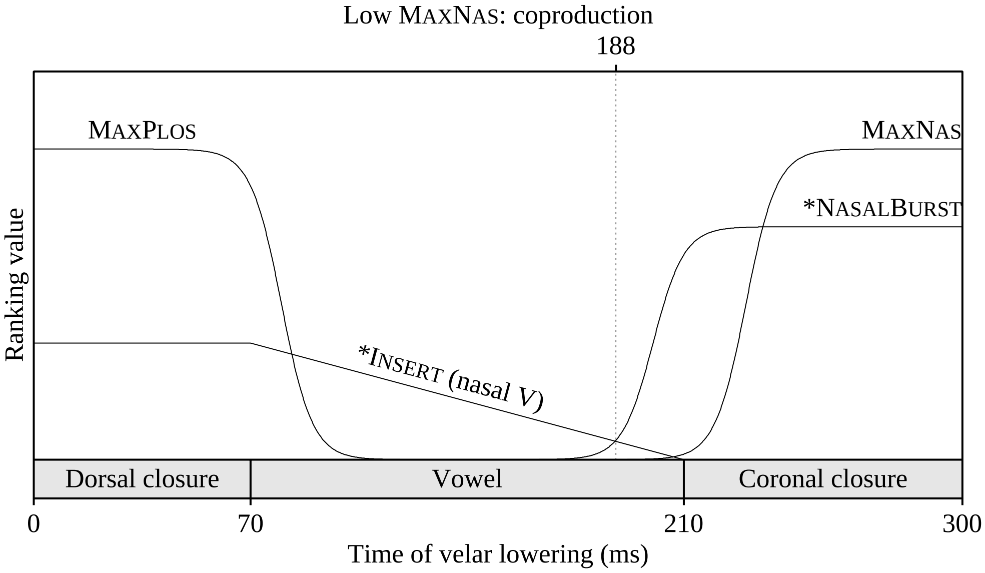
19.1.8 Coarticulation

There has been some controversy about the strategies that speakers use for the timing of articulatory gestures (Kent & Minifie 1977, Fowler 1980).

For instance, Benguerel & Cowan (1974) found that some speakers of French, when asked to pronounce a phrase containing /istʀstʀy/, started the lip rounding for /y/ during the first [s] or even during [i], which suggests the strategy “as early as allowed”, i.e. as soon as the gesture does not conflict with the specifications of the current segment. Most of the authors cited in this section refer to ***articulatory*** specifications: since rounding does not conflict with the articulatory specifications for [s], but does conflict with those for [i], the rounding will start in [s]. As far as motor planning is concerned, such descriptions may be realistic, but for purposes of explanation, I would rather talk about the linguistically more relevant ***perceptual*** specifications: rounding hardly conflicts with the perceptual specifications of |s| (sibilant noise), but does conflict with those of |i| (maximum F_2). In this respect, I would like to quote the pre-OT account by Perkell & Matthies (1992: 2911), who propose that the /iC(C)(C)u/ phenomena show the “simultaneous and variable expression of three competing constraints”, among which a constraint to “begin the /u/-related protrusion movement when permitted by relaxation of the perceptually motivated constraint that the preceding /i/ be unrounded.” In the current section, I show how we can formalize such accounts.

In contradiction with this ***feature-spreading*** model, Bell-Berti & Harris (1979) found that lip rounding started at a fixed time before the coronal release in sequences as [patup] and [pastup] (in their own speech). Bell-Berti & Krakow (1991) found a comparable result for the timing of the velar gesture in |an|: the timing difference between velum lowering and the coronal closure did not depend on the material that preceded [an].

I will now show that these conflicting *feature-spreading* and ***coproduction*** models both turn out to be expected in a typology of strictly ranked phonetic-implementation constraints. Consider the specification |kan|. The plosive is strongly specified for being plosive, because that is its primary specification; I will express this circular statement tautologically as a high-ranked MAXPLOS. The vowel is weakly specified for being non-nasal, because its primary specifications are sonorance and lowness, both of which are not seriously injured by nasalization; I will express this as a constraint family *INSERT (nasal / V / duration), in which I make explicit the dependence of its ranking on the degree of overlap between the lowered velum and the vowel. The nasal specification of |n| wants to make itself heard as early as possible; the ranking of the MAXNAS constraint depends on the duration of nasality: the shorter its duration, the stronger the violation of MAXNAS. Finally, we have a synchronization-and-precision constraint, whose ranking is determined by the working point established in §19.1.2; for a given timing difference Δt, the ranking of this *NASALBURST constraint is the minimum of the rankings of *INSERT (nasal burst

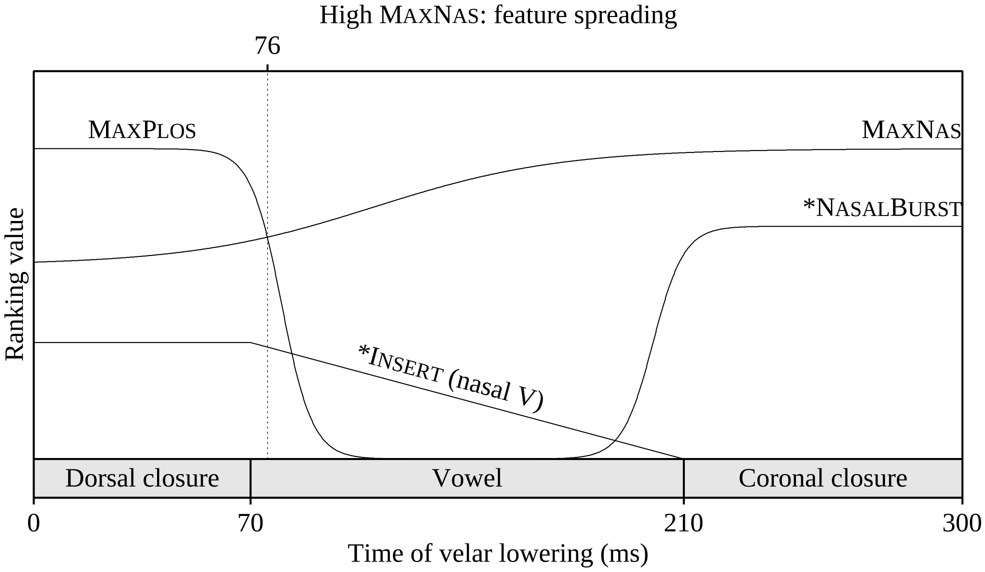


(19.19)

/ Δt / $\sigma = x$) and *SYNC (velum, blade / $\sigma < x$) as functions of x. For instance, for Δt = 20 ms, it is the ranking value associated with the leftmost cutting point in figure 19.10. We can now make the continuous tableau (19.19) of the violated constraints as a function of the moment of velum lowering in [kan].

Optimality Theory is about minimizing the maximum problem. The 188-ms candidate in (19.19) is the most harmonic: this working point is determined by the interaction of the synchronization constraint *NASALBURST and the orality specification for the vowel. If we lengthen the vowel, giving |kaːn|, the curve of *INSERT (nasal V) may lower somewhat (because most of the vowel will be oral), so that the working point will shift a little bit to the left; if we replace the plosive with a glide, however, giving |jaːn|, the working point will not change. Basically, therefore, the constraint rankings in (19.19) are compatible with the coproduction hypothesis.

But we have the freedom of ranking the MAXNAS constraint higher than in (19.19):



(19.20)

The working point has shifted to 76 milliseconds, which is where we find the minimal maximum problem. If we lengthen the utterance to |kajan|, the MAXNAS constraint will dominate the non-nasal specifications of the complete |aja| sequence, and the working point will again be determined by the interaction of MAXNAS with the plosive specification. The rankings in (19.20), therefore, are compatible with the feature-spreading hypothesis.

19.2 An example: nasal harmony

To show that the above account is not a mere restatement of the facts, we must first note that it actually makes predictions about possible languages, and then that these predictions are borne out by the facts.

The proposal that articulatory spreading can be blocked by perceptual specifications, i.e. by protesting *REPLACE constraints, predicts that the degree of opaqueness of the specified segment to spreading must depend on the height of the *REPLACE constraint, and, therefore, on the perceptual difference between the specified and the assimilated segment. We will see that the resulting universal *REPLACE hierarchy accurately predicts the typology of opaqueness to nasal spreading.

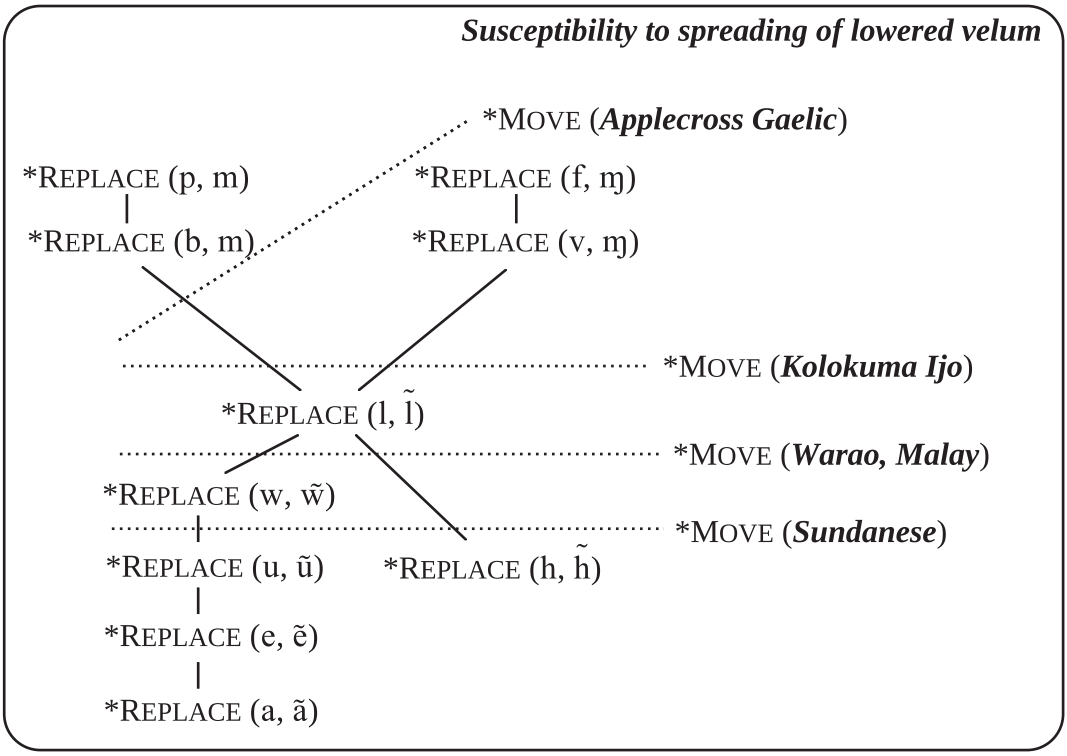
The second prediction is that in so-called perceptual spreading, segments are more transparent as their perceptual specifications are more different from their assimilated counterparts. We will see that this is also borne out for nasal harmony systems.

19.2.1 Functional explanation and description

In nasal-harmony systems, the [lowered velum] gesture is incompatible with the perceptual specifications of most consonants: in decreasing order of perceptual incompatibility, we find plosives, fricatives, liquids, oral glides, and laryngeal glides; this order reflects implicational universals of transparency of consonants to nasal harmony.

For instance, nasality spreads rightward through a glide in Malay [mãj̃ãn] 'stalk' but not through a plosive in [mãkan] 'eat' (Piggott 1992). The phonetic explanation is obvious again. In [mãj̃ãn], the glide becomes nasalized, which hardly makes it less of a glide; for [mãkan], by contrast, spreading would give *[mãŋãn], which replaces an underlying plosive with a nasal, clearly a perceptually much more drastic perturbation. We can rank the offensiveness of nasalization for any segment in the *REPLACE constraint family (19.21), noting that lowering the velum on a fricative will almost certainly produce a plain nasal, though a nasal fricative in Applecross Gaelic is reported not to lose its frication (Van der Hulst & Smith 1982).

The hierarchy is mainly based on the degree of constriction of the oral cavity: the narrower this constriction, the more the sound will be influenced by a lowering of the velum. The location of the constraint for /h/ is based on the perceptual distance between [h] and [h̃], which will also depend on the degree of mouth opening; the difference between a non-nasal and a nasal [h] will not be much different from the difference between a non-nasal and a nasal vowel with the same degree of oral constriction. As for



(19.21)

plosives and fricatives, it is hard to say a priori which of these groups will suffer the most from nasality, i.e. whether it is worse to lose plosiveness or to lose frication.

The typological predictions from (19.21) follow when we cross the *REPLACE hierarchy with the appropriate family of *MOVE (velum) constraints. All replacements whose offensiveness lies below *MOVE, will be implemented, and all those above will not. This will lead to the following implicational universals:

1. If glides can be nasalized, so can vowels and laryngeals.
2. If liquids can be nasalized, so can glides.
3. If plosives or fricatives can be nasalized, so can liquids.

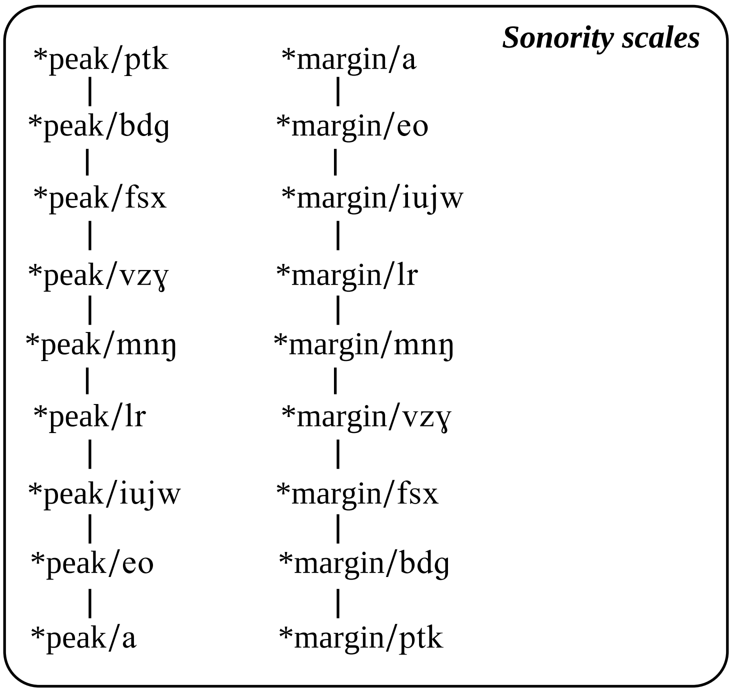
(19.22)

These predicted universals produce exactly the possible sets of nasalization targets identified in Piggott (1992:62) for “Type A” nasal-harmony systems, except that Piggott says that plosives never join in. Five of Piggott’s nasal-spreading systems are shown in (19.21): they all fit into the functional hierarchy that we derived.

19.2.2 Nasal spreading and the sonority hierarchy?

While our functional account may be descriptively adequate, its acceptance in the linguistic community will depend on how its results compare to traditional generative accounts of the same phenomena. I will discuss two previous accounts of nasal spreading. In this section, I will discuss Gnanadesikan’s (1995) idea of coupling the attested hierarchy of susceptibility of nasalization to the ***sonority hierarchy***.

The sonority hierarchy ranks speech sounds according to their suitability to form syllable margins (onsets and codas) and nuclei. Prince & Smolensky’s (1993) account of syllabification in Imdlawn Tashlhiyt Berber, which allows any segment in nucleus position and any segment except /a/ in onset position, provides the following universal hierarchies for margin avoidance and peak (nucleus) avoidance:



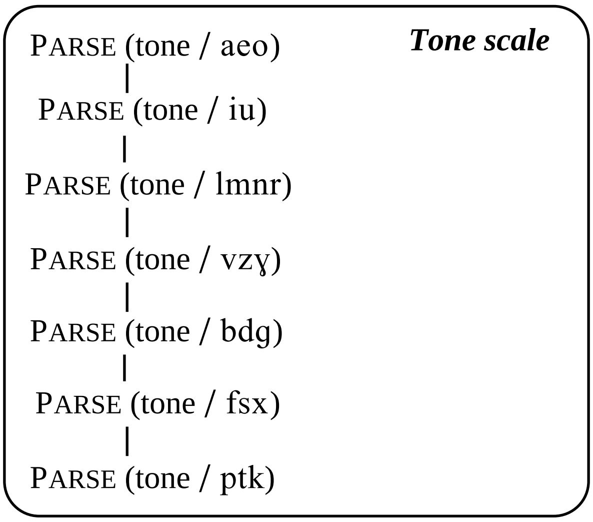
(19.23)

The rankings within these two families are thought to be universal, but the two families can be ranked with respect to one another in a language-specific way: Imdlawn Tashlhiyt chooses the wild ranking *margin/ptk >> *peak/ptk (with undominated PARSE and FILL, and ONSET just above *margin/iujw), while in Dutch the two families are joined somewhere between lr and iu.

Apparently, the rankings in (19.23) are based on several requirements for nuclei. Nuclei like to be continuous sounds, so that they can be lengthened; this moves the plosives /ptkbdg/ in (19.23) to the bottom of the nucleus-affinity hierarchy. Nuclei like to be voiced, so that they can bear tone; this leads to the subdisions of the fricatives and the plosives. And nuclei like to be loud, so that they contribute to the rhythm of the utterance; this leads to the subhierarchy based on the degree of supralaryngeal opening: a > e > i > l > m > v. Now, these phonetic explanations are admittedly *post hoc*, but a similar explanation would even be needed to explain the sonority hierarchy if it were an innate device. After all, natural selection tends to have the effect of improving the fitness of the organism to its environment (Darwin 1859), which in our case would mean that an innate sonority hierarchy would contribute to efficient communication.

But there are ways to determine whether a human property is innate or not. Humans have flexible fingers. We know that these were a result of natural selection (the races who could not make tools, produced fewer grandchildren), because the properties of fingers are hereditary: no infant swimming practice will create webs between the fingers. Now, we can still swim more or less with our innate maladapted peripherals, and the description of the use of the fingers in the art of swimming does not have to refer at all to their original function. If the sonority hierarchy were an innate device as well, likewise separated from its origin, we would expect it, too, to be used unchanged for things other than syllable structure. If, however, the sonority hierarchy is the result of language-specific learning, we expect that there can be hierarchies that look like sonority hierarchies but are just that little different, in line with their current function (they may have webs). We will see that the latter seems to be the case.

First, we note that the subhierarchy that tells us that voiceless fricatives are better nuclei than voiced plosives (used productively in Imdlawn Tashlhiyt), is based on the primacy of the continuity of the sound. If we steer away from syllable positions, and consider the suitability of segments to bear tone, we must conclude that the primary condition for tone is voicing, not continuity. The hierarchy for tone faithfulness can be expressed as the family *REPLACE (tone: H, L / *env*) etc, or loosely as PARSE (tone / *env*), with a fixed ranking by degree of voicing:



(19.24)

This ranking tells us that the higher we are in this scale, the lower we expect the perceptual confusion between high and low tones to be. The hierarchy is supported by some facts: Limburgian and Lithuanian sequences of a short vowel and a consonant can only exhibit a tone contrast if that consonant is a sonorant (lmnr); Limburgian (except Venlo) allows more tone contrasts in /aC/ sequences than in /iC/. The difference between (19.23) and (19.24) is the ranking of voiced plosives and voiceless fricatives. It predicts that there could be languages with voicing contrasts on /bdg/ but not on /fsx/, and no languages with the reverse. Unfortunately, I know of no data that bear on this matter.

More promising would be an investigation into the hierarchies of the susceptibility of segments to perturbations, as long as these hierarchies are expected to be close, though not equal, to the sonority scale. As an example, take the behaviour of [h] in syllabification and in harmony processes. Gnanadesikan (1995: 21) reports on a child that replaces unstressed initial syllables with [fi]: [fimawo] ‘tomorrow’, [fiteɾo] ‘potato’, [fimon] ‘Simone’; however, if the initial consonant of the final, stressed, syllable is a glide or liquid, the child replaces it by the initial consonant of the initial syllable, if that is less sonorous: [fibun] ‘balloon’, [fipis] ‘police’. Gnanadesikan rightly concludes that the sonority scale is involved, though she sees a problem in the behaviour of /h/, which patterns with the less sonorous segments: [fihajn] ‘behind’. However, this is exactly as we would expect in (19.23): [h] is voiceless and, therefore, not very suitable for a nucleus; phonetically, it is a voiceless fricative whose noise stems from the glottal constriction and from any other places in the vocal tract that happen to be narrowed; though its spectral properties depend strongly on the shape of the supralaryngeal cavities, we would be inclined to classify it with the low-sonority voiceless fricatives /fsx/ in the hierarchy (19.23). Gnanadesikan, however, states that “*h* is arguably more sonorous than liquids since it patterns with the more sonorous glides in processes such as nasal harmony”.

The special place of /h/ in (19.21) as compared to (19.23) is completely due to the fact that [h] is the only sound (of the ones considered) that gets it voicelessness from a glottal gesture instead of from an oral constriction: it violates the complementarity of sonorants and obstruents, since it is not a sonorant (i.e., there is no perception of voicing) and it is not an obstruent either (i.e., there is no strong supralaryngeal constriction). Thus, the hierarchy of transparency to nasal spreading follows the appropriate phonetic principle of perceptual contrast, not the allegedly innate sonority scale.

We must conclude that there is no evidence for the innateness of the sonority scale, and that the scales are equal to what they would look like if they were invented afresh by every language learner. What can be considered innate, is the ability to rank faithfulness constraints by degree of contrastivity, i.e. to rank highly what is useful and lowly what is superfluous; this ability may well have had an influence on the number of grandchildren that our forbears managed to put on the earth.

19.2.3 Nasal spreading in feature geometry?

The second generative account of nasal spreading that we will discuss is Piggott (1992). He casts the problem in feature-geometric terms, proposing that “the feature [nasal] is organized as a dependent of the Soft Palate node” (p. 34). Any interpretation of this in functional terms (the perceptual feature [nasal] depends on a soft-palate gesture for its implemantation) is ruled out by Piggott’s subsequent statement that “[s]preading is blocked in this pattern by segments specified for the Soft Palate node”. As we now know, it is the perceptual feature [nasal], not the soft-palate gesture, that is specified, and it is this perceptual specification that blocks the spreading.

Piggott’s basic idea is that segments that are opaque to nasal spreading have an underlying nasal specification, i.e. instead of the functional hierarchy of varying ***degrees*** of specification, Piggott subscribes to an all-or-none representational solution. In Malay, for instance, glides are targets for nasalization, so that they must be underlyingly unspecified for nasality. In Sundanese, glides are opaque to nasal spreading, which Piggott ascribes to a language-specific specification of these glides as [+consonantal]. The difference between Malay and Sundanese follows, then, from Piggott’s following assumption for Universal Grammar (my numbering):

(UG3819a) “If [+nasal] is an underlying property of [+consonantal] segments, then other segments specified underlying [sic] for a Soft Palate node must also be [+consonantal].”

This assumption refers to glides and laryngeals: if glides are [–consonantal], they cannot be opaque to nasal spreading; laryngeals (/h/ and /ʔ/) are assumed to be always [–consonantal], hence not opaque.

Piggott thus considers the laryngeal segments /h/ and /ʔ/ targets for nasal spreading, because they cannot be specified for the Soft Palate node. Now, nasalizing /h/ gives an articulatory coordination that we can describe as [h̃], which results in an auditory perception that we can describe as /h̃/, because some nasality will be heard in the friction noise; but nasalizing /ʔ/ gives an articulation that we can describe with the shorthand [ʔ̃], which will be perceived as /ʔ/, because no nasality will be heard during the closure (though perhaps it will during the glottal burst). Piggott goes into some lengths explaining that *phonologically*, the glottal stop is nasalized, though *phonetically*, it isn't. This is another example of the confusion of articulation and perception, which follows automatically from forcing phonology into the straightjacket of a hybrid feature system.

Note that Piggott's account does not yet predict that in Sundanese all non-glide, non-laryngeal consonants must be opaque, like the glides. In Kolokuma Ijo, the liquid /r/ is subject to nasalization. According to Piggott, /r/ must be unspecified for nasality in this language. Again, this account does not yet predict that the glides /w/ and /j/ are also subject to nasal spreading. In Applecross Gaelic, fricatives are targets of nasal spreading, and must be unspecified for nasality. Again, this does not predict the fact that liquids and glides are also subject to nasalization. To account for the hierarchies not explained by the representations, Piggott introduces a second assumption into Universal Grammar:

(UG3819b) "The segments specified for the Soft Palate node must otherwise constitute a natural class that is not limited to sonorants."

This statement probably requires some exegesis. The class of "segments specified for the Soft Palate node" always includes the nasal stops (/m/ and /n/); any other segments in this class must be opaque to nasal spreading, since they are specified for [–nasal]. Now let's see to what natural classes the nasal stops can belong.

- First, there is the class of stops ([–continuant] segments); this class contains the nasals plus the plosives, so that the plosives must form a possible class of opaque segments.
- Then there is the class of all [+consonantal] segments. This predicts that the set of all non-nasal consonants (with the glides optionally included) can be opaque to nasal spreading.
- The nasals also belong to the class of [+sonorant] segments. This set is ruled out from relevance by the ad-hoc condition "not limited to sonorants" in (UG3819b).
- Piggott comes up with the 'natural class' of ***non-approximant consonants***. Besides the nasals, this class comprises the fricatives and plosives, so that the fricatives and plosives together must form a possible class of opaque segments.

The attested typology, now, can be generated by two parameters: a binary parameter that determines whether glides are consonantal, and a ternary parameter that determines whether the set of segments specified for the Soft Palate node comprises all consonants, or just the non-approximants, or only the stops.

The problem with Piggott's approach is that his assumptions are completely arbitrary and ad hoc, especially the "limitation to sonorants". Without this last condition, only liquids (and sometimes glides) would be opaque, and fricatives and plosives would be targets for nasal spreading, clearly an impossible situation on simple functional grounds. This move makes Piggott’s account hardly acceptable even for a large part of the generative community, but it is hard to see what could be done to save the feature-geometric approach with its hybrid representations of phonological features. The reader is invited to compare this to the functional account, which makes no assumptions beyond the one that phonology adheres to common principles of human motor behaviour and perception.

19.2.4 An empirical difference: nasalization of plosives

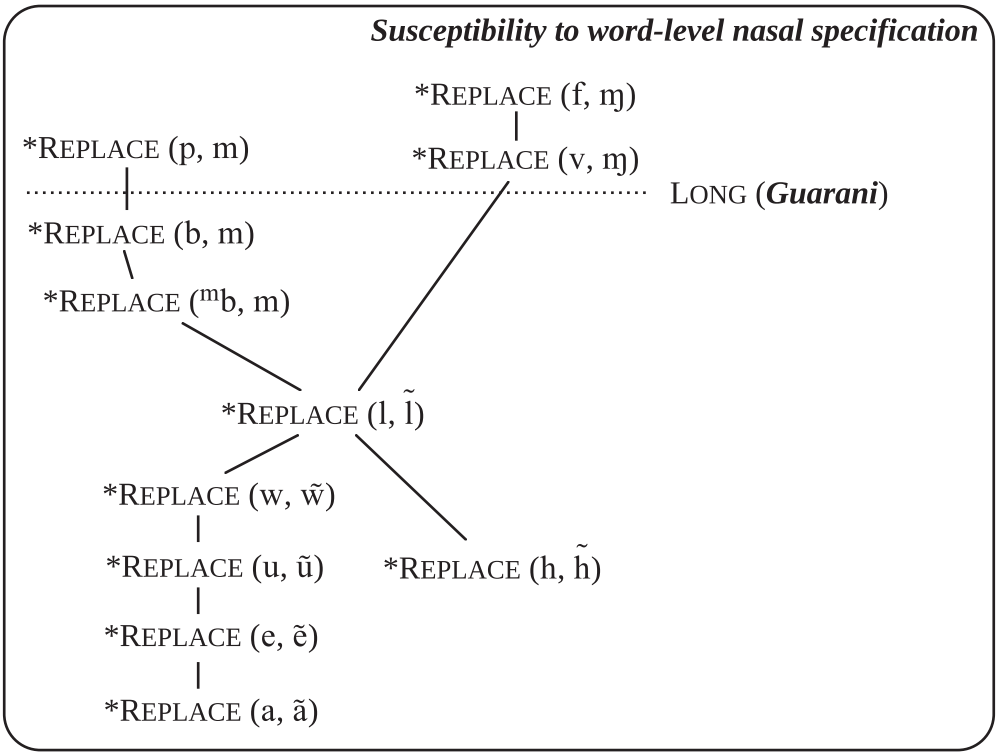
In Piggott’s account, it is impossible that plosives are targets for nasal spreading: the class of segments specified for the Soft Palate node would have to consist of the set of nasal stops alone, and this is ruled out by the famous condition in (UG3819b). The functionally derived hierarchy (19.21), on the other hand, would predict that plosives can also be nasalized, namely, if the *MOVE family is ranked high enough. Of course, the position of *MOVE becomes more rare as it is farther away from the crosslinguistically average position, but a small amount of plosive nasalization should be expected.

While I know of no systematic harmony-like spreading involving plosives, we find a relevant example of sandhi in Sanskrit, where every word-final plosive becomes a nasal if the following word starts with a nasal; unfortunately, we cannot tell what word-final fricatives would do, since these do not exist in Sanskrit. In the Dutch dialect of Bemmel, the nasal sandhi in |ɑ̈k| ‘if I’ + |min| ‘my’, which may surface as /ɑ̈ŋmin/ ‘if I my’, may extend to a prepended |ɔ̈k| ‘also’, giving /ɔ̈ŋɑ̈ŋmin/ ‘even if I my’ (with nasalized vowels); however, this process seems not to be allowed to occur even in a sequence like |ɔ̈k+ɑ̈j+min| ‘even if you me’, which is realized as /ɔ̈kɑ̈jmin/, so we may not be able to draw any conclusions from these data.

19.2.5 Morpheme-level nasal specifcations

The other type of nasal harmony, coined “type B” by Piggott (1992), shows ***transparency*** of obstruents, as in the Guarani example of §19.1.7. Functionally, we expect exactly the same hierarchy as in (19.21), as is shown in (19.25). The *REPLACE constraints have to compete with constraints that try to make every segment in the word nasal. Only those segments that would not lose their main perceptual specifications, are allowed to become nasalized. Fricatives and voiceless plosives generally seem to be belong to the transparent class. Voiced plosives, however, may become nasals: surely the perceptual distance between /b/ and /m/ is less than the distance between /p/ and /m/, because /b/ and /m/ share at least their specification for voicedness.

The fact that the voiced plosives are often /^{m}b/ instead of /b/, leads Piggott to the proposal that voiced stops are specified for the ***Spontaneous Voicing node***. Piggott’s generalization is that only segments specified for Spontaneous Voicing are targets for nasalization. There is, however, an interesting move that Piggott has to make in order to defend his Spontaneous Voicing hypothesis. In his discussion of Type A nasal harmony, Piggott considered the laryngeal segments /h/ and /ʔ/ targets for nasal spreading; in his



(19.25)

discussion of Type B harmony, these laryngeal segments suddenly turn up as *transparent*. This is necessary because according to theories of feature geometry, laryngeal consonants cannot be specified for Spontaneous Voicing. This means that Piggott holds that /h/ is not nasalized in Type B nasal harmony, and that /ʔ/ is not even just “phonologically” nasalized. This is a clear prediction, and it is completely contrary to the ‘functional’ prediction from (19.25), which must hold that |h| and |ʔ| are nasalized.

Thus, we are left with an empirical question: are the laryngeals in Guarani-type nasal-harmony systems pronounced with a lowered velum or not? Contra Piggott, I predict that they are.

19.3 Conclusion

In this chapter, I argued that in articulatory spreading, strong perceptual specifications may produce opacity, and that in perceptual ‘spreading’, strong perceptual specifications may produce transparency.

From the functional standpoint, it is difficult to share Gnanadesikan’s surprise that /h/ turns up in two different places in the two otherwise similar hierarchies (19.21) and (19.23); we should be surprised if it didn’t.

Compared with Piggott’s carefully contrived representational solution, the functional approach needs no recourse to far-fetched assumptions for accurately predicting the attested typology of opacity to nasal spreading.

20 Conclusion

This book established the incorporation of phonetic and functional principles directly into the grammar, putting elements of explanation into a form suitable for actual description. This move replaces the data-driven theories of generative phonology with a theory that predicts what languages would look like if they adhered to principles of effective and efficient communication, and showed common properties of human motor behaviour and perception. Of course, these predictions must be tested with the data.

20.1 Tenets

The following are the basic standpoints defended in this book.

20.1.1 Representations

The single most perspicuous aspect of the functional approach to phonology is its principled distinction between articulatory gestures and perceptual features. This is quite different from the generative approach with its “hybrid” (or “cognitive”) features.

20.1.2 Innateness

The functional standpoint involves not invoking innateness as an explanatory device: while innate coordinations include breathing, sucking and swallowing, those of speech must, can, and will be ***learned***. For instance, the need for perceptual contrast requires /u/ to be labial and velar; these two articulations are subsequently learned as a fixed coordination ([back] → [round]), which is arbitrary from the point of speech production: any other coordination is equally learnable, but may be less suitable for use in a system for human communication.

Constraints are learned (ch. 14), not innate. Children start with empty grammars. Each time a perceptual category emerges, the relevant faithfulness constraints come into being; each time that the child learns to connect an articulatory gesture to a perceptual result, constraints against such gestures come into the picture.

A theory that challenges the innateness assumptions of generative grammar must prove that a human child without the advantages of innate universals can still learn the grammar within a realistic period of time. While I have proved the learnability of the ***structure*** of the grammar (the constraint rankings) in chapter 14, with the first constraint-sorting algorithm that can work with actual acquisition data, the model would put on credibility if it could solve the bootstrapping problem of learning the ***contents*** (the constraints themselves). I expect that an adequate model will result from marrying the gradual learning algorithm with a neural categorization model that is supervised by the semantics and pragmatics of the communicative situation.

20.1.3 Functional constraints

Starting from the functional principles of minimization of effort and confusion, we identified several articulatory constraint families (ch. 7) and perceptual (faithfulness) constraint families (ch. 9), and there are probably some more of them. We also developed a strategy for finding universal rankings and predicting at what points languages are allowed freedom (ch. 11). Optimality Theory seems to be very suitable for expressing function.

20.1.4 Continuous constraint families

We can describe phonetic implementation and perceptual categorization with an Optimality-Theoretic interaction of continuous constraint families (chs. 8 and 10; §19.1). The interaction between two continuous families often leads to an optimal working point, whose position depends on many internal or external factors like adjacent segments, stress versus unstressed position, pragmatically determined reranking of faithfulness, or the crowding of the phonological space. A transition to a discrete phenomenon (e.g., lexical reduction) results if the two interacting constraint families do not intersect. Thus, strictly ranked constraints are the natural devices for the robust description of the gradient phenomena of phonetic implementation: while there is function in phonology, there is also organization in phonetics. It seems that postlexical phonology and phonetic implementation can be described together by a one-level constraint grammar.

20.1.5 What is optimized?

Sound systems are not structured directly according to principles of maximization of *acoustic contrast* and minimization of *motor effort*, but rather on the minimization of ***perceptual confusion***, which involves acoustic contrast and categorization, and on the minimization of ***production effort***, which involves motor effort and coordination. The organizational principles (categorization, coordination) seem to be more important than the more peripheral principles (energy, auditory distinctivity).

20.2 Comparison to generative phonology

The theory of functional phonology shows that several devices that were posited in generative phonology, actually result from the interactions of more fundamental principles of human communication.

20.2.1 Spreading

Spreading is not a separate phonological device (ch. 19). Assimilation effects result from the interaction between articulatory and perceptual constraints. Perceptual features can block articulatory spreading, which leads to a hierarchy of susceptibility to feature change that is based on the perceptual distance between the specified and the changed segment.

20.2.2 Obligatory Contour Principle

The OCP is not a separate phonological device (ch. 18). As a constraint, it belongs in the perception grammar. Its effects in the production grammar are indirect: they result from the interaction of a constraint against loss of perceptual identity with articulatory and perceptual constraints.

20.2.3 Feature geometry

Feature Geometry is not a separate phonological device (§1.2.7, §19.1.1). The alleged ***nodes*** are illusions: they are the results of combining articulatory gestures that have cancelling perceptual results (e.g., spread glottis and constricted glottis) or similar perceptual results (e.g., lip, blade, or tongue-body closures implement nasal stops). In §19.1.1, I showed that the ***place node*** in feature geometry is an illusion that results from interpreting a common perceptual result of the labial, coronal, and dorsal articulators (namely, the [+nasal, –oral] combination) as a built-in phonological device. In §19.2.3, I showed that the ***soft palate node*** is an illusion that results from not distinguishing the articulatory gesture of velum lowering from the perceptual feature of nasality. The third commonly accepted node is the ***laryngeal node***; its establishment is a result of the simple fact that the gestures of glottal narrowing and widening have cancelling perceptual effects.

The only geometries that remain are of the implicational type (§1.2.7); e.g., in order to have any value along the tone scale, a sound must be voiced.

20.2.4 Underspecification

Underspecification is an overused phonological device (ch. 13). Specifications are constraints, and as all other constraints, some are strong and some are weak (ch. 17, 19).

20.2.5 Limitations of the subject matter

The theory as developed in this book is limited to the "melodic" phenomena of segmental and autosegmental phonology, and can be seen as a replacement for the theories of autosegental phonology and feature geometry. Complementary subjects are metrical phonology and prosody, and the description of these will need its own set of constraints, perhaps rooted in universal properties of the production and perception of rhythm.

20.2.6 Empirical adequacy

A theory that offers itself as a new linguistic theory, should be able to describe the actual language data with at least the same amount of insight, generalizability, heuristic power, and explanatory force. While I think that I have succeeded in showing that the theory of functional phonology lives up to these requirements in its handling of the cases of nasal place assimilation and nasal harmony, many other issues in autosegmental phonology and feature geometry may have to be tackled in a similar way before the theory can be expected to receive general acceptance.

20.3 Grammar model

In a theory that takes into account properties of the speaker as well as the listener, we need a ***production grammar*** and a ***perception grammar*** (ch. 6). Given a perceptual specification (an "underlying form"), the production grammar chooses the best from a possibly large number of candidate articulations, by evaluating the articulatory effort of each candidate with gestural constraints (ch. 7), and the faithfulness of the perceptual result of each candidate with faithfulness constraints (ch. 9). Given an acoustic input, the perception grammar categorizes the acoustic events into language-specific perceptual classes.

20.3.1 Optionality

To model learning curves in a realistic way, the learning algorithm must be gradual and the evaluation of each constraint must contain a certain amount of noise (ch. 14). This actually renders the algorithm capable of learning stochastically evaluating grammars: the learner will reproduce the adult degree of optionality (ch. 15). The empirical prediction from this model is that optionality is restricted to grammars compatible with the proposal that each constraint has a fixed mean ranking value and all constraints have the same ranking noise. For instance, an unrestrained output space with 25 candidates has 24 degrees of freedom; if the attested output can be described with a stochastic grammar of 11 constraints, the actual output space must have 10 degrees of freedom. Therefore, we should analyse actual optionality data and see whether the attested grammars fall inside these restricted subspaces. If so, this would constitute an empirical corroboration of the proposal of stochastic evaluation.

20.3.2 Inventories of segments

In segment inventories, symmetries and gaps are predicted by the two constraint-ranking systems of the production and perception grammars (ch. 16). "Poverty of the base" is achieved by the finiteness of the number of categories allowed by the perception grammar.

20.3.3 Sound change

Randomly varying constraint ranking produces a pressure in the direction of preferred sound change (ch. 17). An eternally optimizing sequence of sound change can be circular.

20.3.4 Heuristic power

I explained some language-independent constraint rankings with phonetic principles, but other principles may have to be derived from the data of the languages of the world, especially in the realm of metrical phonology. This situation may be less than ideal, but the possibility of bridging the gap between phonology and phonetics at all is such a good prospect that we should not be afraid of a few initial holes in our knowledge. More positively: if more than one phonetic explanation for a given language fact has been advanced (as is often the case), the phonology may well tell us which of them is correct. Within the realm of phonetic implementation (§10.5.5, §19.1.8), we have seen that evaluation with strict ranking of continuous constraints is compatible with most theories that have been advanced for the explanation of vowel reduction and coarticulation.

20.3.5 Grammar levels

Phonetics and phonology can be described within the same formalism. Postlexical phonology and phonetic implementation can be described together as a single level of constraints, to be evaluated in parallel; exceptionlessness and optionality are perspicuous properties of the phenomena here. Lexical phonology represents the fossilization of such constraints into a more arbitrary system, which may or may not reflect the parallelism typical of the more superficial system. After all, the historical changes that led to the lexical phonology of our language were ordered in time; this may create systems of crucially sequentially ordered rules, reflecting the historical order. Thus, the need of derivation in the more abstract strata of phonology mirrors the arbitrary order in which sound changes took place.

If we realize that all phonological constraints discussed in this book are constraints on the articulatory or perceptual ***output***, we must conclude that a description of the lexical phonology will often involve a serial stratification of the grammar, though crucial seriality would be limited to sequences that used to be described with counterfeeding or counterbleeding rule ordering.

20.3.6 Unnatural phonology

Many a process in morphology, e.g. the English /k/-/s/ alternation in *electric-electricity*, cannot be expressed as the result of an interaction between functional principles. If phenomena such as these are to be expressed in a constraint-ranking grammar, they will have to be handled by ***language-specific constraints***, regardless of whether one works within a functional theory of phonology or within the original generative Optimality Theory. Since these constraints must be *learned*, they pose as much of a problem to the generative maxim of the innateness of the constraint set as does the language-specific acquisition of perceptual categories and articulatory gestures. This is independent evidence for the functional attitude to the innateness question: since some constraints must be learned anyway, the language user is evidently capable of learning constraints, including those sometimes attributed to an innate constraint set.

It should be clear, now, what it is that the constraint sets advocated in this book try to replace: the allegedly innate set of constraints proposed by generative Optimality Theory. I have not proposed that phonology is all phonetics.

20.3.7 How does the speaker speak?

This book treats the explanation and the description of phonological phenomena, but has little to say about their mental implementation. In other words, the ***what*** and the ***why*** are accounted for, but the ***how*** is not.

Consider the English utterance [t^{h}ɛɛ̃n_^{t}s]. We handled its description (velum lowering before blade closure, velum raising before blade release, insertion of a nasal-oral path, insertion of a silence and a release burst) and its explanation (the articulatory problem of synchronization, the perceptual offensiveness of a nasal burst), but we failed to discuss whether the articulatory gestures are intrinsically or extrinsically timed, whether all thinkable output candidates must be generated at production time, or how perceptual expectations (e.g. of the inserted [ẽ] or silence) contribute to speech recognition. A comprehensive theory should handle these things.

The most striking unimplemented device in the theory is the function that generates the output candidates for the given specification; this corresponds to the function GEN of Optimality Theory. In answer to the question what candidates should be evaluated, I would have to say: all the candidates that are even remotely relevant, where *relevant* means: interesting to the linguists that join the discussion about the phonological phenomenon at hand. Quite clearly, this has nothing to do with what the speaker actually does when speaking. The speaker may be very good at evaluating a number of competing candidates: a ranked constraint system can easily and quickly determine the optimal candidate; but the constraint system (which OT identifies with the grammar) cannot generate the candidates. I hypothesize that the speaker only generates the actually possible candidates, i.e. enough candidates to be able to implement the variation that results from pragmatic and stochastic reranking, for instance, for |aːn+pɑsə| only the candidates [aːnpɑsə] and [aːmpɑsə]. The speaker would add a candidate to such a small set if someone else comes up with a better candidate, given the idea that the speaker can easily evaluate any candidates, including those that she did not come up with by herself. The result of this procedure is that if *anyone* comes up with what must be the best candidate according to the shared constraint system, this candidate will diffuse throughout the language community and will ultimately be accepted by every speaker as the optimal form.

If the constraint system, therefore, only describes the speaker's ***competence*** in the sense of an accurate judgment of the competing forms, her speaking performance must be handled by a forward generation system, whose workings, who knows, might be described by such arbitrary rules like

$$\begin{bmatrix}+\text{nas}\\+\text{cor}\end{bmatrix}\rightarrow\begin{bmatrix}+\text{lab}\\-\text{cor}\end{bmatrix}\Big/\ \_\_\begin{bmatrix}\text{C}\\+\text{lab}\end{bmatrix}\qquad\text{(optional)}$$

20.4 Conclusion

A serious theory of sound patterns will require that functional principles like articulatory and perceptual ease be brought into the grammar, and, conversely, that phonetic explanation will include the organizational principles of coordination and categorization. I hope that this book will be the starting point for a functional theory of phonology; that future research into autosegmental phenomena will corroborate its empirical adequacy; and that it will contribute to modelling the language-dependent speech processing that takes place in the language user.

Summary

In this book, I showed that descriptions of the phenomena of phonology would be well served if they were based on accounts of articulatory and perceptual needs of speakers and listeners. For instance, the articulatory gain in pronouncing an underlying |n+k| as [ŋk] is the loss of a tongue-tip gesture. Languages that perform this assimilation apparently weigh this articulatory gain higher than the perceptual loss of the coronal place cues. This perceptual loss causes the listener to have more trouble in reconstructing the perceived /ŋ/ as an underlying |n|. This functionalist account is supported by the markedness relations that it predicts: the ranking of the faithfulness (anti-perceptual-loss) constraints depends on the perceptual distance between the underlying specification (/n/) and the perceptual result (/ŋ/) and on the commonness of the feature values (coronal is more common than dorsal), leading to more or less fixed local rankings as

“do not replace /t/ with /k/” >> “do not replace /n/ with /ŋ/”

and

“do not replace /ŋ/ with /n/” >> “do not replace /n/ with /ŋ/”

where the “ >>“ symbol means “is ranked higher than” or “is more important than”. The first of these two rankings is universal because plosives have better place cues than nasals, and the second is valid in those languages where coronals are more common than dorsals (ch. 9). These universal rankings lead again to near-universals (ch.11) like “if plosives assimilate, so do nasals (at the same place of articulation)” and “if dorsals assimilate, so do coronals (in languages where coronals are more common than dorsals)”.

The idea of constraint ranking is taken from Optimality Theory, which originated in the generative tradition (Prince & Smolensky 1993). The interesting thing of the optimality-theoretic approach to functional principles, is that phonetic explanations can be expressed directly in the production grammar as interactions of gestural and faithfulness constraints. This move makes phonetic explanation relevant for the phonological description of how a speaker generates the surface form from the underlying form. I have shown (chs. 13, 17, 18, 19) that this is not only a nice idea, but actually describes many phonological processes more adequately than the generative (nativist) approach does, at least those processes that have traditionally been handled with accounts that use the hybrid features of autosegmental phonology, underspecification theory, and feature geometry.

The model of a production grammar in functional phonology (ch. 6) starts with a ***perceptual specification***, which is an underlying form cast in perceptual features and their combinations. For each perceptual specification, a number of ***candidate articulations*** are evaluated for their articulatory effort and for the faithfulness of their ***perceptual results*** to the specification. This evaluation is performed by a grammar of many strictly ranked articulatory constraints (ch. 7) and faithfulness constraints (ch. 9), and the best candidate is chosen as the one that will be actually spoken.

There is also a perception grammar, which is a system that categorizes the acoustic input to the listener's ear into language-specific perceptual classes (ch. 8). The listener uses the perception grammar as an input to her speech-recognition system, and the speaker uses the perception grammar to monitor her own speech: in the production grammar, a faithfulness constraint is violated if the output, **as perceived by the speaker**, is different from the specification.

In the language-learning child (ch. 14), the production and perception grammars are empty: they contain no constraints at all. As soon as the child acquires the categorization of acoustic events into communicatively relevant classes, the perception grammar comes into being, and as soon as the child decides that she wants to use the acquired categories to convey semantic and pragmatic content, faithfulness constraints arise in the production grammar. As soon as the child has learned (by play) how to produce the required sounds, constraints against the relevant articulations enter the production grammar. These constraints lower as the child becomes more proficient (by play and imitation), thus leading to more faithful utterances. A general gradual learning algorithm hypothesizes that the child will change her constraint rankings (by a small amount) if her own utterance, **as perceived by herself**, is different from the adult utterance, **as perceived by the child** (the bold phrases on this page stress the prominent role for perception in a functional theory of phonology, as opposed to theories that maintain hybrid phonological representations). This learning algorithm, by the way, is capable of learning ***stochastic grammars***, i.e. the child will learn to show the same degree of variation and optionality as she hears in her language environment (ch. 15).

The original aim of this book was to propose a model for inventories of consonants, based on functional principles of human communication, like minimization of articulatory effort and minimization of perceptual confusion. The symmetry that phonologists see in these inventories follows from the finiteness of the number of perceptual categories and the finiteness of the number of acquired articulatory gestures. The gaps that phoneticians see in these inventories follow from asymmetries in the context dependence of articulatory effort and perceptual contrast. This functional approach to inventories (ch. 16) and phonological phenomena in general marries the linguist's preference for description with the speech scientist's preference for explanation, in a way that, I hope, will eventually appeal to both convictions.

Samenvatting

In dit boek verdedig ik de stelling dat de organisatie van klanken in gesproken taal bepaald wordt door een wisselwerking tussen een aantal algemene principes van efficiënte en effectieve communicatie.

Neem bijvoorbeeld het Nederlandse woord *aanpassen*. De meeste Nederlanders spreken dit uit als *aampassen* (ook al denken ze van niet). Waarom zou iemand dit willen doen? De verklaring luidt: als je de *n* van *aanpassen* als een *n* wilt laten klinken, moet je tijdens de *n* de tongpunt tegen het stuk van de bovenkaak vlak achter de boventanden aan houden, maar als je het uitspreekt als *aampassen*, spaar je deze hele tongpuntbeweging uit. In plaats daarvan moet je nu wel een *m* maken, en daarvoor moet je de lippen dichthouden. Maar die lippen moesten sowieso al dicht om de *p* te kunnen maken, dus dat kan nu fraai in één moeite door. De netto winst van *aampassen* zeggen in plaats van *aanpassen*, is dus dat ene tongpuntgebaar. Dat is dus winst voor het principe van ***minimalisatie van articulatorische moeite***, zeg maar de luiheid van de spreker.

Tegenover deze winst staat wel een verlies. De luisteraar zal bij de uitspraak *aampassen* meer moeite hebben om te reconstrueren wat de spreker eigenlijk bedoelde, dan bij de uitspraak *aanpassen*. Dat is dus verlies voor het principe van ***minimalisatie van perceptieve verwarring***.

Er zijn dus twee functionele principes die samen bepalen of we *aanpassen* dan wel *aampassen* zeggen, namelijk de regel “maak geen tongpuntgebaar” en de regel “vervang een *n* niet door een *m*”. Welke van de twee uitspraken de spreker kiest, hangt af van de relatieve belangrijkheid van deze regels. Voor de meeste Nederlandstaligen wint de anti-tongpuntregel het van de *n*-getrouwheidsregel, zodat de kandidaat *aampassen* het wint.

De rangschikking van de functionele regels (“constraints”) is niet voor elke taal hetzelfde. In het Limburgs is bijvoorbeeld de *n*-getrouwheid hoger gerangschikt dan de anti-tongpuntregel. Dit kun je het duidelijkst horen in de uitspraak van een woord als *menke* (‘mannetje’), dat absoluut niet mag rijmen op het woord voor *tanken* (nog afgezien van het toonverschil). Hieruit moet je overigens niet concluderen dat Limburgers minder luie sprekers zijn dan Nederlandstaligen: vergelijk maar eens het Nederlandse *komt* (lip- plus tongpuntgebaar) met het Limburgse *kump* (alleen een lipgebaar). Blijkbaar heeft elke taal een groot aantal anti-moeiteregels en anti-verwarringsregels, die op een ingewikkelde manier in elkaar grijpen.

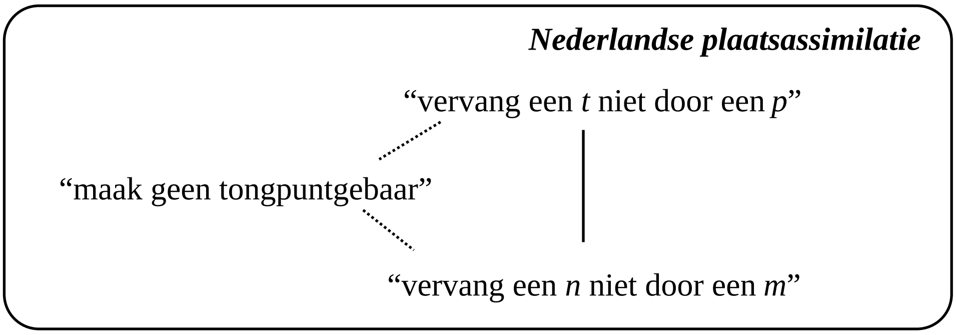
Een van de taken van de taalkunde is het om te voorspellen wat mogelijke talen zijn en wat onmogelijke talen zijn. Als het om het rangschikken van constraints gaat, willen we dus kunnen voorspellen welke rangschikkingen universeel (dus voor alle talen) vastliggen, en welke rangschikkingen door elke taal apart kunnen worden geregeld. Ik zal dit duidelijk maken met het voorbeeld van voorrangsregels in het verkeer.

Er is een regel die zegt “wie groen licht heeft, mag eerst” en een regel die zegt “wie op een voorrangsweg rijdt, mag eerst”. Maar als ik op een voorrangsweg rijdt en ik heb rood licht, mag die ander, die groen heeft, toch eerst. Blijkbaar telt de stoplichtregel zwaarder dan de voorrangswegregel. Ik denk dat deze rangschikking universeel is: iedereen zou het gek vinden als je door rood mocht als je op een voorrangsweg reed (maar waarom eigenlijk?).

Er zijn ook andere voorrangsregels, bijvoorbeeld “rechts gaat voor links” en “auto gaat voor fiets”. Maar als ik op mijn fiets van rechts kom, heeft die auto toch voorrang; blijkbaar telt de auto-fietsregel zwaarder dan de rechts-linksregel. Maar in dit geval zou niemand het erg gek vinden als de volgorde omgekeerd was en fietsers van rechts dus voorrang hadden, en in sommige landen is dit inderdaad het geval. We hebben hier dus niet te maken met een universele rangschikking van regels, maar met een land-afhankelijke.

Iets dergelijks kunnen we doen met het voorbeeld van *aampassen*. We weten al dat de keuze tussen *n* en *m* bepaald wordt door een taalafhankelijke ordening van regels. Maar neem nu het woordje *uitmaken*. Ook hier zouden we aan de regel “maak geen tongpuntgebaar” kunnen voldoen, namelijk door het uit te spreken als *uipmaken*. Er zijn wel talen die zoiets doen, maar het Nederlands is daar niet bij. De kandidaat *uipmaken* schendt namelijk de regel “vervang een *t* niet door een *p*”, en deze regel telt altijd zwaarder dan “vervang een *n* niet door een *m*”, omdat een luisteraar bij het horen van een *m* er nog altijd wel rekening mee zal houden dat het misschien een *n* was, terwijl zoiets bij het horen van een *p* (die misschien eigenlijk een *t* was) veel minder het geval is. Dit komt doordat *p* en *t* zo op het oor veel meer van elkaar verschillen dan *m* en *n*, zoals iedereen weet die wel eens door de telefoon letters heeft moeten spellen: daarvoor moeten om de haverklap Marie en Nico op komen draven.

We hebben nu dus naast een taalafhankelijke rangschikking ook een universele rangschikking gevonden. Het Nederlands heeft de drie regels op de volgende manier geordend (hoe hoger in het plaatje, hoe belangrijker):


Nederlandse plaatsassimilatie
“vervang een *t* niet door een *p*”
“maak geen tongpuntgebaar”
“vervang een *n* niet door een *m*”

De doorgetrokken lijn geeft een universele rangschikking aan, de stippellijntjes een rangschikking die specifiek is voor het Nederlands. Afhankelijk van de hoogte van de anti-tongpuntregel verwachten we dat er drie mogelijke soorten talen zijn: die waarin zowel *n* als *t* kunnen veranderen, die waarin alleen n verandert, en die waarin geen van beiden verandert; een taal waarin *t* wel verandert maar *n* niet, zou niet mogen bestaan. Als we veel talen onderzoeken, blijkt deze voorspelling te kloppen.

In de taalkunde betekent “grammatica”: het geheel van taalregels dat elke spreker, meestal zonder er zich van bewust te zijn, gebruikt bij het spreken en verstaan. Een nieuwe theorie van de grammatica heeft slechts bestaansrecht als ze kan verantwoorden hoe een kind de taal leert. In het geval van de ***functionele fonologie***, die ik in dit boek verdedig, moet ik laten zien hoe een kind zowel de regels zelf kan leren, als de rangschikking van de regels. Welnu, het kind vangt aan met een lege grammatica: getrouwheids-regels komen daar pas in zodra het kind weet welke klanken ze uit elkaar moet kunnen houden om mensen te verstaan, en anti-moeiteregels ontstaan pas zodra het kind weet op welke manier ze de klanken moet produceren. Vervolgens bepaalt het kind in kleine stapjes de rangschikking van al die regels, door haar eigen uitingen met die van de mensen in haar omgeving te vergelijken. Na een jaar of vier met taal bezig zijn, heeft ze die rangschikking bijna helemaal uitgekristalliseerd. Voor het laatste stapje is nog eens een jaar of vier nodig, net als voor het gereedkomen van dit boek.

Uithoorn, 8 maart 1998.

References

Abramson, Arthur S. & Leigh Lisker (1970): “Discriminability across the voicing continuum: cross-language tests.” *Proceedings of the 6th International Congress of Phonetic Sciences*, edited by Bohuslav Hála, Milan Romportl & Přemysl Janota, 569–573. Prague: Academia.

Abry, Christian & Pierre Badin (1996): “Speech Mapping as a framework for an integrated approach to the sensori-motor foundations of language.” *Proceedings 1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, 175–184.

Akamatsu, Tsutomu (1997): *Japanese phonetics: theory and practice*. München: LINCOM Europa.

Allwood, E. & Celia Scully (1982): “A composite model of speech production.” *Proceedings IEEE International Congress on Acoustics, Speech and Signal Processing ’82*, Vol. 2: 932–935.

Andersen, Henning (1969): “Lenition in Common Slavic.” *Language* **45**: 553–574.

Anderson, J.M. & Colin Ewen (1987): *Principles of Dependency Phonology*. Cambridge University Press.

Andrews, Henrietta (1949): “Phonemes and morphophonemes of Temoayan Otomi.” *International Journal of American Linguistics* **15**: 213–222.

Anttila, Arto (1995): *Deriving variation from grammar: A study of Finnish genitives*. Manuscript, Stanford University. [Rutgers Optimality Archive **63**, http://ruccs.rutgers.edu/ roa.html]

Archangeli, Diana (1984): *Underspecification in Yawelmani phonology and morphology*. Doctoral thesis, Massachusetts Institute of Technology, Cambridge. [New York: Garland Press, 1988]

Archangeli, Diana (1988): “Aspects of underspecification theory.” *Phonology* **5**: 183–207.

Archangeli, Diana & Douglas Pulleyblank (1994): *Grounded phonology*. Cambridge: Massachusetts Institute of Technology Press.

Avanesov, P.I. (1949): *Očerki russkoj dialektologii*. Moscow: Učpedgiz.

Avery, P. & Keren Rice (1989): “Segment structure and coronal underspecification.” *Phonology* **6**: 179–200.

Badin, Pierre (1989): “Acoustics of voiceless fricatives: production theory and data.” *Quarterly Progress and Status Report*, *STL-QPSR* ’89:**3**: 33–55. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology (KTH).

Baer, Thomas, J.C. Gore, L.C. Gracco & P.W. Nye (1991): “Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels.” *Journal of the Acoustical Society of America* **90**: 799–828.

Bailly, Gérard (1997): “Learning to speak. Sensori-motor control of speech movements.” *Speech Communication* **22**: 251–267.

Barker, M.A.R. (1964): *Klamath grammar*. Berkeley and Los Angeles: University of California Press.

Beach, D.M. (1938): *The phonetics of the Hottentot language*. Cambridge: Heffer.

Beekes, Robert S.P. (1990): *Vergelijkende taalwetenschap*. Utrecht: Het Spectrum.

Behnke, Kay (1998): *The acquisition of phonetic categories in young infants: a self-organising artificial neural network approach*. Doctoral thesis, Universiteit Twente. [*Max Planck Institute Series in Psycholinguistics* **5**]

Bell-Berti, Fredericka & Katherine S. Harris (1979): “Anticipating coarticulation: Some implications from a study of liprounding.” *Journal of the Acoustical Society of America* **65**: 1268–1270.

Bell-Berti, Fredericka & Rena Arens Krakow (1991): “Anticipatory velar lowering: a co-production account.” *Journal of the Acoustical Society of America* **90**: 112–123.

Benguérel, André-Pierre & Helen A. Cowan (1974): “Coarticulation of upper lip protrusion in French.” *Phonetica* **30**: 41–55.

Berg, Jw. van den, J.T. Zantema & P. Doornenbal (1957): “On the air resistance and the Bernoulli effect of the human larynx.” *Journal of the Acoustical Society of America* **29**: 626–631.

Bergem, Dick R. van (1993): “Acoustic vowel reduction as a function of sentence accent, word stress and word class.” *Speech Communication* **12**: 1–23.

Berko, Jean & Roger Brown (1960): “Psycholinguistic research methods.” *Handbook of research methods in child development*, edited by Paul Mussen, 517–557. New York: Wiley.

Berwick, Robert C. (1985): *The acquisition of syntactic knowledge*. Cambridge: MIT Press.

Berwick, Robert C. & Partha Niyogi (1996): “Learning from triggers.” *Linguistic Inquiry* **27**: 605–622.

Best, C.T., G.W. McRoberts & N.M. Sithole (1988): “Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English speaking adults and infants.” *Journal of Experimental Psychology: Human Perception and Performance* **14**: 345–60.

Bhaskararao, Peri & Peter Ladefoged (1991): “Two types of voiceless nasals.” *Journal of the International Phonetic Association* 80–88.

Bird, Steven & T. Mark Ellison (1994): “One-Level Phonology: autosegmental representations and rules as finite automata.” *Computational Linguistics* **20-1**: 55–90.

Bird, Steven & Ewan Klein (1990): “Phonological events.” *Journal of Linguistics* **26**: 33–56.

Bladon, R.A.W. & Björn Lindblom (1981): “Modelling the Judgement of Vowel Quality Differences.” *Journal of the Acoustical Society of America* **69**: 1414–1422.

Bliese, Loren F. (1981): *A generative grammar of Afar*. Arlington: Summer Institute of Linguistics, University of Texas.

Blight, Richard C. & Eunice V. Pike (1976): “The phonology of Tenango Otomi.” *International Journal of American Linguistics* **42**: 51–57.

Boë, Louis-Jean, Pascal Perrier, Bernard Guérin & Jean-Luc Schwartz (1989): “Maximal vowel space.” *Proceedings Eurospeech ’89*, **2**: 281–284.

Boë, Louis-Jean, Jean-Luc Schwartz, Rafael Laboissière & Nathalie Vallée (1996): “Integrating articulatory-acoustic ocnstraints in the prediction of sound structures.” *Proceedings 1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, 163–166.

Boë, Louis-Jean, Jean-Luc Schwartz & Nathalie Vallée (1994): “The prediction of vowel systems: perceptual contrast and stability.” *Fundamentals of speech synthesis and speech recognition*, edited by Eric Keller, 185–213. London & New York: Wiley.

Boersma, Paul (1989): “Modelling the distribution of consonant inventories by taking a functionalist approach to sound change.” *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **13**: 107–123.

Boersma, Paul (1990): “Modelling the distribution of consonant inventories.” *Proceedings Linguistics and Phonetics ’90*, Prague.

Boersma, Paul (1991): “Synthesis of speech sounds from a multi-mass model of the lungs, vocal tract, and glottis.” *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* **15**: 79–108.

Boersma, Paul (1993a): “An articulatory synthesizer for the simulation of consonants.” *Proceedings Eurospeech ’93*: 1907–1910.

Boersma, Paul (1993b): “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 of the University of Amsterdam* **17**: 97–110.

Boersma, Paul (1995): “Interaction between glottal and vocal-tract aerodynamics in a comprehensive model of the speech apparatus.” *Proceedings of the International Congress of Phonetic Sciences* **2**: 430–433.

Boersma, Paul (1997a): *The elements of Functional Phonology*. Manuscript, University of Amsterdam. [Rutgers Optimality Archive **173**, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (1997b): “How we learn variation, optionality, and probability.” *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**: 43–58. [an earlier version in: Rutgers Optimality Archive **221**, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (1997c): “Inventories in Functional Phonology.” *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**: 59–90. [Rutgers Optimality Archive **232**, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (1997d): *Sound change in Functional Phonology*. Manuscript, University of Amsterdam. [Rutgers Optimality Archive **237**, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (1997e): “Functional Optimality Theory.” *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**: 37–42.

Boersma, Paul (1998a): *Spreading in functional phonology*. Manuscript, University of Amsterdam. [Rutgers Optimality Archive, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (1998b): *The OCP in functional phonology*. Manuscript, University of Amsterdam. [Rutgers Optimality Archive, http://ruccs.rutgers.edu/roa.html]

Boersma, Paul (to appear): “Learning a grammar in Functional Phonology.” *Optimality Theory: Phonology, syntax, and acquisition*, edited by Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer. Oxford University Press.

Boersma, Paul & David Weenink (1996): *Praat, a system for doing phonetics by computer*. Institute of Phonetic Sciences, University of Amsterdam, report **132**. [http://fonsg3.hum.uva.nl/praat/]

Bosch, Louis ten (1991): *On the structure of vowel systems. Aspects of an extended vowel model using effort and contrast*. Doctoral thesis, Universiteit van Amsterdam.

Braine, M.D.S. (1976): “Review of N.V. Smith, The acquisition of phonology.” *Language* **52**: 489–98.

Brentari, Diane (1995): “Sign language phonology: ASL.” *The handbook of phonological theory*, edited by John A. Goldsmith, 615–639. Oxford: Blackwell.

Browman, Catherine P. & Louis Goldstein (1984): “Dynamic modeling of phonetic structure.” *Status Report on Speech Research, Haskins Laboratories, New Haven* **79/80**: 1–17. [Also in *Phonetic linguistics: Essays in honour of Peter Ladefoged*, edited by Victoria A. Fromkin, 35–53. Academic Press, 1985]

Browman, Catherine P. & Louis Goldstein (1986): “Towards an articulatory phonology.” *Phonology Yearbook* **3**, edited by Colin Ewen & John Anderson, 219–252. Cambridge University Press.

Browman, Catherine P. & Louis Goldstein (1989): “Articulatory gestures as phonological units.” *Phonology* **6**: 201–251.

Browman, Catherine P. & Louis Goldstein (1990a): “Tiers in articulatory phonology, with some applications for casual speech.” *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, edited by John Kingston & Mary Beckman, 341–376. Cambridge University Press.

Browman, Catherine P. & Louis Goldstein (1990b): “Gestural specification using dynamically-defined articulatory structures.” *Journal of Phonetics* **18**: 299–320.

Browman, Catherine P. & Louis Goldstein (1992): “Articulatory Phonology: an overview.” *Phonetica* **49**: 155–180.

Browman, Catherine P. & Louis Goldstein (1993): “Dynamics and articulatory phonology.” *Status Report on Speech Research, Haskins Laboratories, New Haven* **113**: 51–62.

Brugmann, K. & B. Delbrück (1897): *Grundriß der vergleichenden Grammatik der indogermanischen Sprachen*. Vol. I, *Einleitung und Lautlehre*. Strasbourg: Trübner.

Carpenter, Gail A. & Stephen Grossberg (eds., 1991): *Pattern recognition by self-organizing neural networks*. Cambridge: MIT Press.

Catford, J.C. (1977): *Fundamental problems in phonetics*. Edinburgh University Press.

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

Cho, Taehong & Peter Ladefoged (1997): “Variations and universals in VOT: evidence from 17 endangered languages.” *UCLA Working Papers in Phonetics* **95**: 18–40.

Chomsky, Noam & Morris Halle (1968): *The sound pattern of English*. New York: Harper and Row.

Clark, Robin & Ian Roberts (1993): “A computational model of language learnability and language change.” *Linguistic Inquiry* **24**: 299–345.

Clements, G. Nick (1985): “The geometry of phonological features.” *Phonology Yearbook* **2**: 225–252.

Clements, G. Nick (1987): “Phonological feature representation and the description of intrusive stops.” *Chicago Linguistic Society Parasession* **23**: 29–50.

Clements, G. Nick & Elizabeth V. Hume (1995): “The internal organization of speech sounds.” *The handbook of phonological theory*, edited by John A. Goldsmith, 245–306. Oxford: Blackwell.

Coker, Cecil H. (1968): “Speech synthesis with a parametric articulatory model.” *Speech Symposium Kyoto*, paper A4. [Reprinted in *Speech synthesis*, edited by James L. Flanagan & Lawrence R. Rabiner, 135–139. Stroudsburg, Pa.: Dowden, Hutchinson & Ross, 1973]

Cole, Jennifer S. & Charles W. Kisseberth (1994): “An optimal domains theory of harmony.” *Studies in the Linguistic Sciences* **24-2**. [Rutgers Optimality Archive **22**, http://ruccs. rutgers.edu/roa.html]

Cornyn, William S. (1944): *Outline of Burmese grammar*. *Language*, Supplement **20-4**.

Cranen, Bert (1987): *The acoustic impedance of the glottis*. Doctoral thesis, Katholieke Universiteit Nijmegen.

Cranen, Bert & Juergen Schroeter (1996): “Physiologically motivated modelling of the voice source in articulatory analysis/synthesis.” *Speech Communication* **19**: 1–19.

Crothers, John (1978): “Typology and universals of vowel systems.” *Universals of human language*, Vol. 2, *Phonology*, edited by Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik, 93–152. Stanford, Calif.: Stanford University Press.

Dantsuji, Masatake (1982): “An acoustic study on glottalized vowels in the Yi (Lolo) language – a preliminary report –.” *Studia Phonologica* **XVI**: 1–11.

Dantsuji, Masatake (1984): “A study on voiceless nasals in Burmese.” *Studia Phonologica* **XVIII**: 1–14.

Dantsuji, Masatake (1986): “Some acoustic observations on the distinction of place of articulation for voiceless nasals in Burmese.” *Studia Phonologica* **XX**: 1–11.

Darwin, Charles (1859): *On the origin of species by means of natural selection, or The preservation of favoured races in the struggle for life*. London: John Murray.

Dik, Simon C. (1978): *Functional Grammar*. London: Academic Press.

Dik, Simon C. (1989): *The theory of Functional Grammar*. Part I, *The structure of the clause*. Dordrecht: Foris.

Dols, Willy (1944): *Sittardse diftongering*. [Sittard: Alberts’ Drukkerijen, 1953]

Eimas, P.D., E.R. Siqueland, Peter W. Jusczyk & J. Vigorito (1971): “Speech perception in infants.” *Science* **171**: 303–6.

Ellison, T. Mark (to appear): “The universal constraint set: convention not fact.” *Optimality Theory: phonology, syntax, and acquisition*, edited by Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer. Oxford University Press.

Ewen, Colin & Harry G. van der Hulst (1987): “Single-valued features and the distinction between [–F] and [ØF].” *Linguistics in the Netherlands 1987*, edited by F. Beukema & Peter Coopmans, 51–60. Dordrecht: Foris.

Fant, Gunnar (1960): *Acoustic theory of speech production*. The Hague: Mouton.

Fant, Gunnar (1968): “Analysis and synthesis of speech processes.” *Manual of phonetics*, edited by Bertil Malmberg, 173–277. Amsterdam: North Holland.

Fink, B. Raymond & Robert J. Demarest (1978): *Laryngeal biomechanics*. Cambridge, Mass.: Harvard University Press.

Flanagan, James L. (1955): “A difference limen for vowel formant frequency.” *Journal of the Acoustical Society of America* **27**: 1223–1225.

Flanagan, James L. (1972): *Speech analysis synthesis and perception*. Second, Expanded Edition. Berlin: Springer.

Flanagan, James L. & L. Cherry (1969): “Excitation of vocal-tract synthesizers.” *Journal of the Acoustical Society of America* **45**: 764–769.

Flanagan, James L. & Kenzo Ishizaka (1977): “Acoustic characterization and computer simulation of the air volume displaced by the vibrating vocal cords: lateral and longitudinal motion.” *Articulatory modeling and phonetics*, Proceedings of a symposium held at Grenoble, edited by R. Carré, R. Descout & M. Wajskop. GALF Groupe de la Communication Parlée.

Flanagan, James L., Kenzo Ishizaka & K.L. Shipley (1975): “Synthesis of speech from a dynamic model of the vocal cords and vocal tract.” *Bell System Technical Journal* **54**: 485–506.

Flanagan, James L., Kenzo Ishizaka & K.L. Shipley (1980): “Signal models for low bit-rate coding of speech.” *Journal of the Acoustical Society of America* **68**: 780–791.

Flanagan, James L. & Lorinda L. Landgraf (1968): “Self-oscillating source for vocal-tract synthesizers.” *IEEE Transactions on Audio and Electroacoustics* **AU-16**: 57–64. [Reprinted in *Speech synthesis*, edited by James L. Flanagan & Lawrence R. Rabiner, 140–147. Stroudsburg, Pa.: Dowden, Hutchinson & Ross, 1973]

Flemming, Edward (1995): *Auditory representations in phonology*. Doctoral thesis, University of California at Los Angeles.

Fletcher, Harvey & W.A. Munson (1933): “Loudness, its definition, measurement, and calculation.” *Journal of the Acoustical Society of America* **5**: 82–108.

Fourakis, Marios & Robert Port (1986): “Stop epenthesis in English.” *Journal of Phonetics* **14**: 197–221.

Fowler, Carol A. (1980): “Coarticulation and theories of extrinsic timing.” *Journal of Phonetics* **8**: 113–133.

Frisch, Stefan, Michael Broe & Janet Pierrehumbert (1997): *Similarity and phonotactics in Arabic*. Rutgers Optimality Archive **223**, http://ruccs.rutgers.edu/roa.html.

Fry, D.B., Arthur S. Abramson, Peter D. Eimas & Alvin M. Liberman (1962): “The identification and discrimination of synthetic vowels.” *Language and Speech* **5**: 171–179.

Gamkrelidze, T. & V. Ivanov (1973): “Sprachtypologie und die Rekonstruktion der gemeinindogermanischen Verschlüsse.” *Phonetica* **27**: 150–156.

Gandour, Jack (1974): “The features of the larynx: n-ary or binary?” *UCLA Working Papers in Phonetics* **27**: 147–159.

Garrett, Andrew (1991): “Indo-European reconstruction and historical methodologies.” *Language* **67**: 790–804.

Gentil, Michèle (1990): “Organization of the articulatory system: peripheral mechanisms and central coordination.” *Speech production and speech modelling,* edited by William J. Hardcastle & Alain Marchal, 1–22. Dordrecht: Kluwer.

Gibson, Edward & Kenneth Wexler (1994): “Triggers.” *Linguistic Inquiry* **25**: 407–454.

Gnanadesikan, Amalia (1995): *Markedness and faithfulness constraints in child phonology*. Manuscript, University of Massachusetts. [Rutgers Optimality Archive **67**, http://ruccs.rutgers. edu/roa.html]

Gnanadesikan, Amalia (1997): *Phonology with ternary scales*. Doctoral thesis, University of Massachusetts.

Gold, E. Mark (1967): “Language identification in the limit.” *Information and Control* **10**: 447–474.

Goldsmith, John (1976): *Autosegmental phonology*. Doctoral thesis, Massachusetts Institute of Technology, Cambridge. [Indiana University Linguistics Club, Bloomington. New York: Garland Press, 1979]

Goldsmith, John (1993): “Harmonic phonology.” *The last phonological rule: Reflections on constraints and derivations*, edited by John A. Goldsmith, 21–60. University of Chicago Press.

Grossberg, Stephen (1976): “Adaptive pattern classification and universal recoding: a parallel development and coding of neural feature detectors.” *Biological Cybernetics* **23**: 121–34.

Gupta, J.P., S.S. Agrawal & Rais Ahmed (1968): “Perception of (Hindi) consonants in clipped speech.” *Journal of the Acoustical Society of America* **45**: 770–773.

Hale, Mark & Charles Reiss (1996): *The initial ranking of faithfulness constraints in UG*. Manuscript, Concordia University. [Rutgers Optimality Archive **104**, http://ruccs.rutgers.edu/ roa.html]

Hale, Mark & Charles Reiss (to appear): “Formal and empirical arguments concerning phonological acquisition.” *Linguistic Inquiry*. [Rutgers Optimality Archive, http://ruccs.rutgers.edu/roa.html]

Halle, Morris & Kenneth N. Stevens (1971): “A note on laryngeal features.” *Quarterly Progress Report*, Research Laboratory of Electronics, Massachusetts Institute of Technology **101**: 198–213.

Hammarström, Göran (1973): “Generative phonology: a critical appraisal.” *Phonetica* **27**: 157–184.

Hammond, Michael (1995): *There is no lexicon*! Manuscript, University of Arizona, Tucson. [Rutgers Optimality Archive **43**, http://ruccs.rutgers.edu/roa.html]

Hanoteau, Adolphe (1858): *Essai de grammaire kabyle*. Second edition, 1906. Reprinted 1976 by Philo Press, Amsterdam.

Hardcastle, William J. (1976): *Physiology of speech production. An introduction for speech scientists*. London: Academic Press.

Hayes, Bruce (1989): “Compensatory lengthening in moraic phonology.” *Linguistic Inquiry* **20**: 253–306.

Hayes, Bruce (1995): “A phonetically-driven, optimality-theoretic account of post-nasal voicing.” Handout Tilburg Derivational Residue Conference. [Rutgers Optimality Archive **93g**, http://ruccs.rutgers.edu/roa.html]

Hayes, Bruce (1996a): “Phonetically driven optimality-theoretic phonology.” Handout of a LOT course, Utrecht.

Hayes, Bruce (1996b): “Phonetically driven phonology: the role of Optimality Theory and inductive grounding.” *Proceedings of the 1996 Milwaukee Conference on Formalism and Functionalism in Linguistics*. [Rutgers Optimality Archive **158**, http://ruccs.rutgers. edu/roa.html]

Hayes, Bruce & Margaret MacEachern (to appear): “Folk verse form in English.” *Language*.

Hayward, Katherine M. (1989): “The Indo-European language and the history of its speakers: The theories of Gamkrelidze and Ivanov.” *Lingua* **78**: 37–86.

Hayward, Katherine (1993): A review of the 1990 reprint of Westermann & Ward (1933). *Journal of Phonetics* **21**: 497–499.

Hegedüs, L. (1959): “Beitrag zur Frage der Geminaten.” *Zeitschrift für Phonetik* **12**: 68–106.

Hirano, M., W. Vennard & John J. Ohala (1970): “Regulation of register, pitch and intensity of voice. An electromyographic investigation of intrinsic laryngeal muscles.” *Folia Phoniatrica* **22**: 1–20.

Hirose, Hajime (1997): “Investigating the physiology of laryngeal structures.” *The handbook of phonetic sciences*, edited by William J. Hardcastle & John Laver, 116–136. Oxford: Blackwell.

Hirose, Hajime & Thomas Gay (1972): “The activity of the intrinsic laryngeal muscles in voicing control.” *Phonetica* **25**: 140–164.

Hirsch, Charles (1988): *Numerical computation of internal and external flows*. Vol. 1, *Fundamentals of numerical discretization*. Chichester: Wiley.

Hirsch, Charles (1990): *Numerical computation of internal and external flows*. Vol. 2, *Computational methods for inviscid and viscous flows*. Chichester: Wiley.

Hixon, Thomas J. (1987): “Respiratory function in speech.” *Respiratory function in speech and song*, Thomas J. Hixon and collaborators, 1–54. London: Taylor & Francis.

Hixon, Thomas J., Dennis H. Klatt & Jere Mead (1971): “Influence of forced transglottal pressure changes on vocal fundamental frequency.” *Journal of the Acoustical Society of America* **49**: 105 (abstract).

Hulst, Harry van der (1989): “Atoms of segmental structure: components, gestures, and dependency.” *Phonology* **6**: 253–284.

Hulst, Harry van der & Norval Smith (1982): “Prosodic domains and opaque segments in autosegmental theory.” *The structure of phonological representations*, vol. 2, edited by Harry van der Hulst & Norval Smith, 311–336. Dordrecht: Foris.

Honda, Kiyoshi (1996): “Organization of tongue articulation for vowels.” *Journal of Phonetics* **24**: 39–52.

Hyman, Larry (1985): *A theory of phonological weight*. Dordrecht: Foris.

Ishizaka, Kenzo & James L. Flanagan (1972): “Synthesis of voiced sounds from a two-mass model of the vocal cords.” *Bell System Technical Journal* **51**: 1233–1268. [Reprinted in *Speech synthesis*, edited by James L. Flanagan & Lawrence R. Rabiner, 148–183. Stroudsburg, Pa.: Dowden, Hutchinson & Ross, 1973]

Ishizaka, Kenzo & Masatoshi Matsudaira (1972): *Fluid mechanical considerations of vocal cord vibration*. Monograph 8, Speech Communications Research Laboratory, Santa Barbara.

Itô, Junko & R. Armin Mester (1986): “The phonology of voicing in Japanese.” *Linguistic Inquiry* **17**: 49–73.

Itô, Junko, Armin Mester & Jaye Padgett (1995): “NC: licensing and underspecification in Optimality Theory.” *Linguistic Inquiry* **26**: 571–613.

Jacobs, Haike & Carlos Gussenhoven (to appear): “Loan phonology: perception, salience, the lexicon and OT.” *Optimality Theory: Phonology, syntax, and acquisition*, edited by Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer.

Jakobson, Roman (1941): *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala.

Jakobson, Roman, E.C. Cherry & Morris Halle (1953): “Toward the logical description of languages in their phonemic aspect.” *Language* **29**: 34–46.

Jesteadt, Walt, Craig C. Wier & David M. Green (1977): “Intensity discrimination as a function of frequency and sensation level.” *Journal of the Acoustical Society of America* **61**: 169–177.

Jun, Jongho (1995): “Place assimilation as the result of conflicting perceptual and articulatory constraints.” *West Coast Conference on Formal Linguistics* **14**.

Jusczyk, Peter W. (1986): “Toward a model of the development of speech perception.” *Invariance and variability in speech processes*, edited by Joseph S. Perkell & Dennis H. Klatt, 1–35. Hillsdale, N.J.: Lawrence Erlbaum.

Jusczyk, Peter W. (1992): “Developing phonological categories from the speech signal.” *Phonological development: models, research, implications,* edited by Charles A. Ferguson, Lise Menn & Carol Stoel-Gammon. Timonium, Md.: York Press.

Kaisse, Ellen (1992): “Can [consonantal] spread?” *Language* **68**: 313–332.

Kats, J.C.P. (1985): *Remunjs waordebook*. Roermond: H. van der Marck.

Kawasaki, Haruko (1982): *An acoustical basis for universal constraints on sound sequences*. Doctoral thesis, University of California at Berkeley.

Kawato, M., Y. Maeda, Y. Uno & R. Suzuki (1990): “Trajectory formation of arm movement by cascade meural network model based on minimum torque-change criterion.” *Biological Cybernetics* **62**: 275–288.

Keating, Patricia A. (1990): “The window model of coarticulation: articulatory evidence.” *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, edited by John Kingston & Mary Beckman, 451–470. Cambridge University Press.

Kelly, John L. & Carol C. Lochbaum (1962): “Speech synthesis.” *Proceedings Fourth International Congress of Acoustics*, paper G42, 1–4. [Reprinted in *Speech synthesis*, edited by James L. Flanagan & Lawrence R. Rabiner, 127–130. Stroudsburg, Pa.: Dowden, Hutchinson & Ross, 1973]

Kelso, J.A. Scott, Elliot L. Saltzman & Betty Tuller (1986): “The dynamical perspective on speech production: data and theory.” *Journal of Phonetics* **14**: 29–59.

Kenstowicz, Michael (1994): *Phonology in generative grammar*. Cambridge Mass. & Oxford: Blackwell.

Kent, R.D. & F.D. Minifie (1977): “Coarticulation in recent speech production models.” *Journal of Phonetics* **5**: 115–133.

Kewley-Port, Diane & Bishnu S. Atal (1989): “Perceptual differences between vowels in a limited phonetic space.” *Journal of the Acoustical Society of America* **85**: 1726–1740.

Kibrik, A.E. & Kodzasov, Sandro V. (1990): *Sopostavitel’noe izucenie dagestanskix jazykov. Imja. Fonetika.* Moscow University Press. ˇ

Kiparsky, Paul (1985): “Some consequences of lexical phonology.” *Phonology Yearbook* **2**: 85–138.

Kiparsky, Paul & Lise Menn (1977): “On the acquisition of phonology.” *Language learning and thought*, edited by J. Macnamara, 47–78. New York: Academic Press.

Kirchner, Robert (1998): *Geminate inalterability and lenition*. Manuscript, University of California at Los Angeles. [Rutgers Optimality Archive **252**, http://ruccs.rutgers.edu/roa.html]

Klatt, Dennis (1982): “Prediction of perceived phonetic distance from critical-band spectra: A first step.” *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing* 1278–1281.

Klein, W., Reinier Plomp & Louis C.W. Pols (1970): “Vowel spectra, vowel spaces and vowel identification.” *Journal of the Acoustical Society of America* **48**: 999–1009.

Kluender, K.R., Randy L. Diehl & P.R. Killeen (1987): “Japanese quail can learn phonetic categories.” *Science* **237**: 1195–7.

Koike, Yasuo & Satoshi Imaizumi (1988): “Objective evaluation of laryngostroboscopic findings.” *Vocal physiology: Voice production, mechanisms and functions,* edited by Osamu Fujimura, 433–442. New York: Raven Press.

Koopmans-van Beinum, Florien J. (1980): *Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions*. Doctoral thesis, University of Amsterdam.

Koopmans-Van Beinum, Florien J. & Jeannette M. van der Stelt (1986): “Early stages in the development of speech movements.” *Precursors of early speech*, edited by B. Lindblom & R. Zetterström. Basingstoke: Macmillan.

Koopmans-Van Beinum, Florien J. & Jeannette M. van der Stelt (to appear): "Early speech development in children acquiring Dutch." *The acquisition of Dutch*, edited by S. Gillis & A. de Houwer. Amsterdam and Philadelphia: John Benjamins.

Kröger, Bernd J. (1990): "A moving noise source and a tube bend in the reflection type line analog." *IPKöln-Berichte* **16**: 59–67.

Kuhl, Patricia K. (1979): "The perception of speech in early infancy." In Lass, N.J. (ed.): *Speech and language: Advances in basic research and practice*. Vol. 1, edited by N.J. Lass, 1–47. New York: Academic Press.

Kuhl, Patricia K. (1991): "Human adults and human infants show a "perceptual magnetic effect" for the prototypes of speech categories, monkeys do not." *Perception and Psychophysics* **50**: 93–107.

Kuhl, Patricia K. & J.D. Miller (1978): "Speech perception by the chinchilla: identification functions for synthetic VOT stimuli." *Journal of the Acoustical Society of America* **63**: 905–17.

Kuhl, Patricia K. & D.M. Padden (1982): "Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques." *Perception and Psychophysics* **32**: 542–50.

Laboissiére, Rafael & A. Galvan (1995): "Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences." *Proceedings of the XIIIth International Congress of Phonetic Sciences*. Vol. 1: 358–361.

Labov, William (1994): *Principles of linguistic change.* Vol. 1. *Internal factors*. Oxford: Blackwell.

Ladefoged, Peter (1971): *Preliminaries to linguistic phonetics*. University of Chicago Press.

Ladefoged, Peter (1973): "The features of the larynx." *Journal of Phonetics* **1**: 73–83.

Ladefoged, Peter (1990a): "On dividing phonetics and phonology: comments on the papers by Clements and by Browman and Goldstein." *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, edited by John Kingston & Mary Beckman, 398–405. Cambridge University Press.

Ladefoged, Peter (1990b): "Some reflections on the IPA." *UCLA Working Papers in Phonetics* **74**: 61–76.

Ladefoged, Peter (1995): "Voiceless approximants in Tee." *UCLA Working Papers in Phonetics* **91**: 85–88.

Ladefoged, Peter & Anthony Traill (1994): "Clicks and their accompaniments." *Journal of Phonetics* **22**: 33–64.

Lakoff, George (1993): "Cognitive Phonology." *The last phonological rule: Reflections on constraints and derivations*, edited by John A. Goldsmith, 117–145. University of Chicago Press.

Lass, Roger (1994): *Old English: A historical linguistic companion*. Cambridge University Press.

Landau, L.D. & E.M. Lifshitz (1953): *Gidrodinamika*. [French translation: *Mécanique des fluides*. Moscow: Editions Mir, 1971]

Leben, William (1973): *Suprasegmental phonology*. Doctoral thesis, Massachusetts Institute of Technology, Cambridge. [New York: Garland Press, 1980]

Leben, William (1978): "The representation of tone." *Tone: A linguistic survey*, edited by Victoria Fromkin, 177–219. New York: Academic Press.

Lehiste, Ilse, Katherine Morton & Mark A.A. Tatham (1973): "An instrumental study of consonant gemination." *Journal of Phonetics* **1**: 131–148.

Leoni, F.A., F. Cutugno & R. Savy (1995): "The vowel system of Italian connected speech." *Proceedings of the International Congress of Phonetic Sciences* **4**: 396–399.

LeVeque, Randall J. (1992): *Numerical methods for conservation laws*. Basle: Birkhäuser.

Levin, Juliette (1985): *A metrical theory of syllabicity*. Doctoral thesis, Massachusetts Institute of Technology, Cambridge.

Liberman, Alvin M., Katherine S. Harris, H.S. Hoffman & B.C. Griffith (1957): "The discrimination of speech sounds within and across phoneme boundaries." *Journal of Experimental Psychology* **54**: 358–368.

Liljencrants, Johan (1985): *Speech synthesis with a reflection-type line analog*. Doctoral thesis, Royal Institute of Technology, Stockholm.

Liljencrants, Johan & Björn Lindblom (1972): "Numerical simulation of vowel quality systems: the role of perceptual contrast." *Language* **48**: 839–862.

Lindau, Mona (1975): *[Features] for vowels* [*UCLA Working Papers in Phonetics* **30**].

Lindblom, Björn (1963): "Spectrographic study of vowel reduction." *Journal of the Acoustical Society of America* **35**: 1773–1781.

Lindblom, Björn (1986): "Phonetic universals in vowel systems." *Experimental Phonology*, edited by John J. Ohala & Jeri J. Jaeger, 13–44. Orlando: Academic Press.

Lindblom, Björn (1990a): "Models of phonetic variation and selection." *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm* **XI**: 65–100.

Lindblom, Björn (1990b): "Explaining phonetic variation: a sketch of the H&H theory." *Speech production and speech modelling*, edited by William J. Hardcastle & Alain Marchal, 403–439. Dordrecht: Kluwer.

Lindblom, Björn, James Lubker & Thomas Gay (1979): “Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation.” *Journal of Phonetics* **7**: 147–161.

Lindblom, Björn & Michael Studdert-Kennedy (1967): “On the rôle of formant transitions in vowel recognition.” *Journal of the Acoustical Society of America* **42**: 830–843.

Lindquist, J. (1969): “Laryngeal mechanisms in speech.” *Quarterly Progress and Status Report, STL-QPSR* ’69:**2-3**: 26–31. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology (KTH).

Lindquist, J. (1972): “Laryngeal articulation studies on Swedish subjects.” *Quarterly Progress and Status Report, STL-QPSR* ’72:**2-3**: 10–27. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology (KTH).

Lisker, Leigh & Arthur S. Abramson (1964): “A cross-language study of voicing in initial stops.” *Word* **20**: 384–422.

Lisker, Leigh & Arthur S. Abramson (1967): “The voicing dimension: some experiments in comparative phonetics.” *Proceedings of the Sixth International Congress of Phonetic Sciences*, 563–567.

Lombardi, Linda (1995): *Why Place and Voice are different: constraint interactions and featural faithfulness in Optimality Theory*. Manuscript, University of Maryland. [Rutgers Optimality Archive **105**, http://ruccs.rutgers.edu/roa.html]

McCarthy, John (1986): “OCP effects: gemination and antigemination.” *Linguistic Inquiry* **20**: 71–99.

McCarthy, John (1988): “Feature geometry and dependency: a review.” *Phonetica* **45**: 84–108.

McCarthy, John J. (1995): “Remarks on phonological opacity in Optimality Theory.” *Proceedings of the Second Colloquium on Afro-Asiatic Linguistics*, edited by Jacqueline Lecarme, Jean Lowenstamm & Ur Shlonsky. [Rutgers Optimality Archive **79**, http://ruccs.rutgers.edu/roa.html]

McCarthy, John J. (1998): *Sympathy & phonological opacity*. Manuscript, University of Massachusetts. [Rutgers Optimality Archive, http://ruccs.rutgers.edu/roa.html]

McCarthy, John J. & Prince, Alan (1986): *Prosodic morphology*. Manuscript, University of Massachusetts and Brandeis University.

McCarthy, John & Alan Prince (1993a): *Prosodic morphology I: Constraint interaction and satisfaction*. Manuscript, University of Massachusetts, Amherst, and Rutgers University, New Brunswick.

McCarthy, John & Alan Prince (1993b): “Generalized alignment.” *Yearbook of Morphology 1993*, edited by Geert Booij & Jaap van Marle, 79–153. Dordrecht: Kluwer. [Rutgers Optimality Archive **7**, http://ruccs.rutgers.edu/roa.html]

McCarthy, John & Alan Prince (1994): “The emergence of the unmarked: optimality in prosodic morphology.” *Papers of the 24th Annual Meeting of the North Eastern Linguistic Society*, edited by Mercè González, 333–379. Amherst: Graduate Linguistic Student Association. [Rutgers Optimality Archive **13**, http://ruccs.rutgers.edu/roa.html]

McCarthy, John & Alan Prince (1995): “Faithfulness and reduplicative identity.” *Papers in Optimality Theory* [*Occasional Papers* **18**], edited by J. Beckman, S. Urbanczyk & L. Walsh, 249–384. Amherst: University of Massachusetts. [Rutgers Optimality Archive **60**, http://ruccs.rutgers.edu/roa.html]

McGowan, Richard S. (1992): “Tongue-tip trills and vocal-tract wall impedance.” *Journal of the Acoustical Society of America* **91**: 2903–2910.

McGowan, Richard S. (1994): “Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests.” *Speech Communication* **14**: 19–48.

Macken, Marlys A. (1980): “The child”s lexical representation: The “puzzle-puddle-pickle” evidence.” *Journal of Linguistics* **16**: 1–17.

MacNeilage, Peter F. (1997): “Acquisition of speech.” *The handbook of phonetic sciences*, edited by William J. Hardcastle and John Laver, 303–332. Cambridge Mass. and Oxford: Blackwell.

Maddieson, Ian (1984): *Patterns of sounds*. Cambridge University Press.

Maddieson, Ian & Victoria Balboa Anderson (1994): “Phonetic structures of Iaai.” *UCLA Working Papers in Phonetics* **87**: 163–182.

Maeda, Shinji (1982): “A digital simulation method of the vocal-tract system.” *Speech Communication* **1**: 199–229.

Maeda, Shinji (1988): “Improved articulatory model.” *Journal of the Acoustical Society of America* **84**: S146 (abstract).

Maeda, Shinji (1990): “Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model.” *Speech production and modelling*, edited by William J. Hardcastle & Alain Marchal, 131–150. Dordrecht: Kluwer.

Marlett, S. & Joseph Stemberger (1983): “Empty consonants in Seri.” *Linguistic Inquiry* **5**: 617–639.

Martinet, André (1955): *Economie des changements phonétiques*. Bern: A. Francke.

Menn, Lise (1980): “Phonological theory and child phonology.” *Child phonology*. Vol. 1, *Production,* 23–41. New York & London: Academic Press.

Mermelstein, Paul (1973): “Articulatory model for the study of speech production.” *Journal of the Acoustical Society of America* **53**: 1070–1082.

Meyer-Eppler, W. (1953): “Zur Erzeugungsmechanismus der Geräuschlaute.” *Zeitschrift für Phonetik* **7**: 196–212.

Mitchell, A.R. (1969): *Computational methods in partial differential equations*. London: John Wiley.

Mohanan, K.P. (1986): *The theory of lexical phonology*. Dordrecht: Reidel.

Mohanan, K.P. (1993): “Fields of attraction in phonology.” *The last phonological rule: Reflections on constraints and derivations,* edited by John A. Goldsmith, 61–116. University of Chicago Press.

Mohanan, K.P. (1995): “Organization of the grammar.” *The handbook of phonological theory,* edited by John A. Goldsmith, 24–69. Oxford: Blackwell.

Mohanan, K.P. & Tara Mohanan (1984): “Lexical phonology of the consonant system in Malayalam.” *Linguistic Inquiry* **15**: 575–602.

Morse, Philip M. & K. Uno Ingard. (1968): *Theoretical acoustics*. New York: McGraw-Hill.

Moscati, Sabatino, Anton Spitaler, Edward Ullendorff & Wolfram von Soden (1964): *An introduction to the comparative grammar of the Semitic languages*. Wiesbaden: Otto Harrassowitz. [2nd printing, 1969]

Myers, J. Scott (1994): *OCP effects in Optimality Theory*. Rutgers Optimality Archive **6**, http://ruccs.rutgers.edu/roa.html.

Nelson, W.L. (1983): “Physical principles for economies of skilled movements.” *Biological Cybernetics* **46**: 135–147.

Newman, Stanley (1944): *Yokuts language of California*. New York: Viking Fund Publications in Anthropology **2**.

Odden, David (1986): “On the role of the Obligatory Contour Principle in phonological theory.” *Language* **62**: 353–383.

Odden, David (1988): “Anti antigemination and the OCP.” *Linguistic Inquiry* **19**: 451–475.

Odden, David (1995): “Tone: African languages.” *The handbook of phonological theory,* edited by John A. Goldsmith, 444–475. Oxford: Blackwell.

Ohala, John J. (1972): “How is pitch lowered?” *Journal of the Acoustical Society of America* **52**: 124 (abstract).

Ohala, John J. (1975): “Phonetic explanations for nasal sound patterns.” *Nasálfest*, edited by C.A. Ferguson, L.M. Hyman & J.J. Hyman, 289–316. Stanford University.

Ohala, John J. (1976): “A model of speech aerodynamics.” *Report of the Phonology Laboratory at Berkeley* **1**: 93–107.

Ohala, John J. (1993a): “The phonetics of sound change.” *Historical linguistics: Problems and perspectives,* edited by C. Jones, 237–278. London: Longman.

Ohala, John J. (1993b): “Sound change as nature’s speech perception experiment.” *Speech Communication* **13**: 155–161.

Ohala, John J. & Carol J. Riordan (1979): “Passive vocal tract enlargement during voiced stops.” *Speech Communication Papers*, edited by J.J. Wolf & D.H. Klatt, 89–92. New York: Acoustical Society of America. [*Journal of the Acoustical Society of America* **65 (S1)**: S23]

Okell, J. (1969): *A reference grammar of colloquial Burmese*. Oxford University Press.

Osborn, H. (1966): “Warao I: phonology and morphophonemics.” *International Journal of American Linguistics* **32**: 108–123.

Padgett, Jaye (1995): “Partial class behavior and nasal place assimilation.” *Proceedings of the Arizona Phonology Conference: Workshop on Features in Optimality Theory*, Coyote Working Papers, University of Arizona, Tucson. [Rutgers Optimality Archive **113**, http://ruccs.rutgers.edu/roa.html]

Passy, Paul (1891): *Etude sur les changements phonétiques et leurs caractères généraux*. Paris: Librairie Firmin - Didot.

Pater, Joe (1996): “Austronesian nasal substitution and other NC̥ effects.” *Proceedings of the Utrecht Prosodic Morphology Workshop*. [Rutgers Optimality Archive **160**, http://ruccs.rutgers.edu/roa.html]

Payan, Yohan & Pascal Perrier (1996): “Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis.” *Speech Communication* **22**: 185–205.

Peeters, Frans (1951): *Het klankkarakter van het Venloos*. Doctoral thesis, Katholieke Universiteit Nijmegen. Nijmegen: Centrale Drukkerij.

Perkell, Joseph S. (1996): “Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling.” *Journal of Phonetics* **24**: 3–22.

Perkell, Joseph S. & Melanie L. Matthies (1992): “Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability.” *Journal of the Acoustical Society of America* **91**: 2911–2925.

Perrier, Pascal, Christian Abry & Eric Keller (1989): “Vers une modélisation des mouvements du dos de la langue.” *Revue d’Acoustique* **2**: 69–77. [Also in *Bulletin Laboratoire de la Communication Parlée* **2**: 45–63, 1988]

Perrier, Pascal, Hélène Lœvenbruck & Yohan Payan (1996): “Control of tongue movements: the equilibrium-point hypothesis perspective.” *Journal of Phonetics* **24**: 53–75.

Perrier, Pacal, David J. Ostry & Rafael Laboissière (1996): “The Equilibrium Point Hypothesis and its application to speech motor control.” *Journal of Speech and Hearing Research* **39**: 365–378.

Peterson, Gordon E. & Harold L. Barney (1952): “Control methods used in a study of vowels.” *Journal of the Acoustical Society of America* **24**: 175–184.

Piggott, Glyne (1992): “Variability in feature dependency: The case of nasality.” *Natural Language and Linguistic Theory* **10**: 33–78.

Piggott, Glyne & Harry van der Hulst (1997): “Locality and the nature of nasal harmony.” *Lingua* **103**: 85–112.

Pinkerton, Sandra (1986): “Quichean (Mayan) glottalized and nonglottalized stops: A phonetic study with implications for phonological universals.” *Experimental phonology*, edited by John J. Ohala & Jeri J. Jaeger, 125–139. London: Academic Press.

Plomp, Reinier (1970): “Timbre as a multidimensional attribute of complex tones.” *Frequency analysis and periodicity detection in hearing*, ed. by R. Plomp & G.F. Smoorenburg, 397–414. Leiden: Sijthoff.

Pols, Louis C.W. (1983): “Three-mode principal component analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation.” *Speech Communication* **2**: 275–293.

Pols, Louis C.W., L.J.Th. van der Kamp & Reinier Plomp (1969): “Perceptual and physical space of vowel sounds.” *Journal of the Acoustical Society of America* **49**: 458–467.

Pols, Louis C.W., H.R.C. Tromp & Reinier Plomp (1973): “Frequency analysis of Dutch vowels from 50 male speakers.” *Journal of the Acoustical Society of America* **53**: 1093–1101.

Press, W.H., B.P. Flannery, S.A. Teukolsky & W.T. Vetterling (1989): *Numerical recipes*. Cambridge University Press.

Prince, Alan & Paul Smolensky (1993): *Optimality Theory: Constraint interaction in generative grammar*. Rutgers University Center for Cognitive Science Technical Report **2**.

Pulleyblank, Douglas (1993): “Vowel harmony and Optimality Theory.” *Actas do Workshop Sobre Fonologia*, University of Coimbra, 1–18.

Pulleyblank, Douglas (1996): “Neutral vowels in Optimality Theory: a comparison of Yoruba and Wolof.” *Canadian Journal of Linguistics* **41**: 295–347.

Pulleyblank, Douglas, Ping Jiang-King, Myles Leitch & Ọlanike Ọla (1995): “Typological variation through constraint rankings: low vowels in tongue root harmony.” *Proceedings of the Arizona Phonology Conference: Workshop on Features in Optimality Theory*. University of Arizona.

Pulleyblank, Douglas & William J. Turkel (1995): *Asymmetries in feature interaction. Learnability and constraint ranking*. Manuscript, University of British Columbia.

Pulleyblank, Douglas & William J. Turkel (1996): “Optimality Theory and learning algorithms: the representation of recurrent featural asymmetries.” *Current trends in phonology: Models and methods*, edited by J. Durand & B. Laks. University of Salford.

Queller, K. (1988): “Review of N. Waterson, *Prosodic phonology*.” *Journal of Child Language* **15**: 463–467.

Recasens, Daniel (1991): *Fonètica descriptiva del català* [*Biblioteca Filològica* **XXI**]. Barcelona: Institut d’Estudis Catalans.

Reenen, Pieter Thomas van (1981): *Phonetic feature definitions. Their integration into phonology and their relation to speech. A case study of the feature NASAL*. Doctoral thesis, Vrije Universiteit Amsterdam. Dordrecht: Foris.

Rivas, A. (1974): “Nasalization in Guaraní.” *Papers from the 5th Annual Meeting of the North Eastern Linguistic Society*, 134–143. Cambridge: Harvard University.

Rousselot, Jean (1891): *Les modifications phonétiques du langage*. Paris: Welter.

Rubin, Philip, Thomas Baer & Paul Mermelstein (1981): “An articulatory synthesizer for perceptual research.” *Journal of the Acoustical Society of America* **70**: 321–328.

Rubin, Philip, Elliot L. Saltzman, Louis Goldstein, Richard S. McGowan, Mark Tiede & Catherine P. Browman (1996): “CASY and extensions to the task-dynamic model.” *Proceedings 1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, 125–128.

Sagey, Elizabeth (1986): *The representation of features and relations in nonlinear phonology*. Doctoral thesis, Massachusetts Institute of Technology, Cambridge.

Schachter, Paul & Fe T. Otanes (1972): *Tagalog reference grammar*. Berkeley: University of California Press.

Schane, Sanford A. (1995): “Diphthongization in Particle Phonology.” *The handbook of phonological theory*, edited by John A. Goldsmith, 586–608. Oxford: Blackwell.

Schouten, M.E.H. (1981): “Het verschil tussen *bot* en *bod* — een vergeefse speurtocht.” *Nieuwe Taalgids* **74-6**: 537–546.

Schroeder, M.R., Bishnu S. Atal & J.L. Hall (1979): “Optimizing digital speech coders by exploiting masking properties of the human ear.” *Journal of the Acoustical Society of America* **66**: 1647–1652

Schwartz, Jean-Luc, Louis-Jean Boë, Pascal Perrier, Bernard Guérin & Pierre Escudier (1989): “Perceptual contrast and stability in vowel systems: a 3-D simulation study.” *Proc. Eurospeech ’89* **2**: 63–66.

Schwartz, Jean-Luc, Louis-Jean Boë & Nathalie Vallée (1995): “Testing the dispersion-focalization theory: phase spaces for vowel systems.” *Proceedings of the International Congress of Phonetic Sciences* **13-1**: 412–415.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée & Christian Abry (1997): “The Dispersion-Focalization Theory of vowel systems.” *Journal of Phonetics* **25**: 255–286.

Scobbie, James (1993): “Issues in constraint violation and conflict.” *Computational phonology* [*Edinburgh Working Papers in Cognitive Science* **8**], edited by T. Mark Ellison & James M. Scobbie, 37–54.

Scully, Celia, Eric Castelli, Eric Brearley & Marion Shirt (1992): “Analysis and simulation of a speaker’s aerodynamic and acoustic patterns for fricatives.” *Journal of Phonetics* **20**: 39–51.

Sievers, Eduard (1876): *Grundzüge der Lautphysiologie*. Leipzig: Breitkopf und Härtel.

Siewierska, Anna (1991): *Functional Grammar*. London: Routledge.

Sihler, Andrew L. (1995): *New comparative grammar of Greek and Latin*. New York: Oxford University Press.

Silverman, Daniel, Barbara Blankenship, Paul Kirk & Peter Ladefoged (1994): “Phonetic structures in Jalapa Mazatec.” *UCLA Working Papers in Phonetics* **87**: 113–130.

Sluijter, Agaath (1995): *Phonetic correlates of stress and accent*. Doctoral thesis, University of Leiden.

Smith, Neilson V. (1973): *The acquisition of phonology: A case study*. Cambridge University Press.

Smolensky, Paul (1996a): “On the comprehension/production dilemma in child language.” *Linguistic Inquiry* **27**: 720–731.

Smolensky, Paul (1996b): *The initial state and ‘richness of the base’ in Optimality Theory*. Technical Report 96-4, Department of Cognitive Science, Johns Hopkins University, Baltimore. [Rutgers Optimality Archive **154**, http://ruccs.rutgers.edu/roa.html]

Sommerfeld, Arnold (1964): *Mechanics of deformable bodies*. New York & London: Academic Press.

Son, Rob J.J.H. van (1993): *Spectro-temporal features of vowel segments* [*Studies in Language and Language Use* **3**]. Doctoral thesis, University of Amsterdam. Amsterdam: IFOTT.

Son, Rob J.J.H. van & Louis C.W. Pols (1992): “Formant movements of Dutch vowels in a text, read at normal and fast rate.” *Journal of the Acoustical Society of America* **92**: 121–127.

Sondhi, Man Mohan & J.R. Resnick (1983): “The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis.” *Journal of the Acoustical Society of America* **73**: 985–1002.

Sondhi, Man Mohan & Juergen Schroeter (1987): “A hybrid time-frequency domain articulatory speech synthesizer.” *IEEE Transactions on Acoustics, Speech and Signal Processing* **35**: 955–967.

Sprigg, R.K. (1965): “Prosodic analysis and Burmese syllable initial features.” *Anthropological Linguistics* **7-6**: 59–81.

Steenwijk, Han (1992): *The Slovene dialect of Resia. San Giorgio*. Doctoral thesis, Univ. of Amsterdam.

Steriade, Donca (1987): “Redundant values.” *Papers from the Parasession on Autosegmental and Metrical Phonology*, edited by A. Bosch, B. Need & E. Schiller, 339–362. Chicago Linguistic Society.

Steriade, Donca (1995): “Underspecification and markedness.” *The handbook of phonological theory*, edited by John A. Goldsmith, 114–174. Oxford: Blackwell.

Stetson, R.H. (1951): *Motor phonetics*. North Holland, Amsterdam.

Stevens, Kenneth N., (1971): “Airflow and turbulence noise for fricative and stop consonants: static considerations.” *Journal of the Acoustical Society of America* **50**: 1180–1192.

Stevens, Kenneth N. (1989): “On the quantal nature of speech.” *Journal of Phonetics* **17**: 3–45.

Stevens, Kenneth N. (1990): “Some factors influencing the precision required for articulatory targets: comments on Keating’s paper.” *Papers in Laboratory Phonology I: Between the grammar and physics of Speech*, ed. by John Kingston & Mary Beckman, 471–475. Cambridge University Press.

Stevens, Kenneth N. and Samuel Jay Keyser (1989): “Primary features and their enhancement in consonants.” *Language* **65**: 81–106.

Stevens, Kenneth N., Samuel Jay Keyser & Haruko Kawasaki (1986): “Toward a phonetic and phonological theory of redundant features.” *Invariance and variability in speech processes*, edited by Joseph S. Perkell & Dennis H. Klatt, 426–449. Hillsdale, N.J.: Lawrence Erlbaum.

Streeter, L.A. (1976): "Language perception of 2-month-old infants shows effects of both innate mechanisms and experience." *Nature* **259**: 39–41.

Streitberg, Wilhelm (1896): *Urgermanische Grammatik*. Carl Winter, Heidelberg.

Stuart-Smith, Jane (1995): "The role of phonetics in the evaluation of reconstructed sound change." *Proceedings of the International Congress of Phonetic Sciences* **4**: 682–685.

Švec, Jan G. & Josef Pešák (1994): "Vocal breaks from the modal to falsetto register." *Folia Phoniatrica et Logopedica* **46**: 97–103.

Tans, J.G.H. (1938): *Isoglossen rond Maastricht*. Doctoral thesis, Katholieke Universiteit Nijmegen.

Tesar, Bruce (1995): *Computational Optimality Theory*. Doctoral thesis, University of Colorado.

Tesar, Bruce (1996): "An iterative strategy for learning metrical stress in Optimality Theory." To appear in *Proceedings of the 21st Annual Boston University Conference on Language Development*.

Tesar, Bruce (1997): "An iterative strategy for language learning." *Lingua* **104**: 131–145.

Tesar, Bruce (to appear): "On the roles of optimality and strict domination in language learning." *Optimality Theory: Phonology, syntax, and acquisition*, edited by Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer.

Tesar, Bruce & Paul Smolensky (1993): *The learnability of Optimality Theory: an algorithm and some basic complexity results*. Manuscript, Department of Computer Science & Institute of Cognitive Science, University of Colorado at Boulder. [Rutgers Optimality Archive **2**, http://ruccs.rutgers.edu/roa.html]

Tesar, Bruce & Paul Smolensky (1996): *Learnability in Optimality Theory (long version)*. Technical Report 96-3, Department of Cognitive Science, Johns Hopkins University, Baltimore. [Rutgers Optimality Archive **156**, http://ruccs.rutgers.edu/roa.html]

Trask, R.L. (1996): *A dictionary of phonetics and phonology*. London: Routledge.

Turkel, William (1994): *The acquisition of Optimality Theoretic systems*. Manuscript, University of British Columbia. [Rutgers Optimality Archive **11**, http://ruccs.rutgers.edu/roa.html]

Vallée, Nathalie (1994): *Systèmes vocaliques: de la typologie aux prédictions*. Doctoral thesis, Institut de la Communication Parlée, Université Stendhal, Grenoble.

Vihman, Marilyn M. (1982): "A note on children's lexical representations." *Journal of Child Language* **9**: 249–53.

Vihman, Marilyn M. (1996): *Phonological development. The origins of language in the child.* Cambridge Mass. and Oxford: Blackwell.

Voegelin, C.F. (1956): "Phonemicizing for dialect study with reference to Hopi." *Language* **32**: 116–135.

Waterson, N. (1971): "Child phonology: A prosodic view." *Journal of Linguistics* **7**: 179–211.

Weidert, Alfons (1975): *Componential analysis of Lushai phonology*. Amsterdam: John Benjamins.

Werker, Janet (1991): "The ontogeny of speech perception." *Modularity and the motor theory of speech production*, edited by Ignatius G. Mattingly & Michael Studdert-Kennedy, 91–109. Hillsdale, N.J.: Lawrence Erlbaum.

Westbury, John R. (1983): "Enlargement of the supraglottal cavity and its relation to stop consonant voicing." *Journal of the Acoustical Society of America* **73**: 1322–1336.

Westbury, John R. & Patricia Keating (1986): "On the naturalness of stop consonant voicing." *Journal of Linguistics* **22**: 145–166. [Also in *UCLA Working Papers in Phonetics* **60**: 1–19, 1985]

Westermann, D. & Ida C. Ward (1933): *Practical phonetics for students of African languages*. Oxford University Press, London. [Reprinted with an Introduction by J. Kelly in 1990, by Kegan Paul International, London and New York]

Wexler, K. & M.R. Manzini (1987): "Parameters and learnability in binding theory." *Parameter setting*, edited by T. Roeper and E. Williams, 41–76. Dordecht: Reidel.

Wilhelms-Tricarico, Reiner (1995): "Physiological modeling of speech production: methods for modeling soft-tissue articulators." *Journal of the Acoustical Society of America* **95**: 3085–3098.

Wilhems-Tricarico, Reiner (1996): "A biomechanical and physiologically-based vocal tract model and its control." *Journal of Phonetics* **24**: 23–28.

Wolfram, Stephen (1991): *Mathematica, a system for doing mathematics by computer, second edition*. Redwood: Addison-Wesley.

Yip, Moira (1988): "The obligatory contour principle and phonological rules: a loss of identity." *Linguistic Inquiry* **19**: 65–100.

Zoll, Cheryl S. (1994): "Subsegmental parsing: floating features in Chaha and Yawelmani." *Phonology at Santa Cruz* **3**. [Rutgers Optimality Archive **29**, http://ruccs.rutgers.edu/ roa.html]

Zoll, Cheryl S. (1996): *Parsing below the segment in a constraint based framework*. Doctoral thesis, University of California at Berkeley. [Rutgers Optimality Archive **143**, http://ruccs.rutgers.edu/roa.html]

Zwicker, E. & R. Feldtkeller (1967): *Das Ohr als Nachrichtenempfänger*. Stuttgart: S. Hirzel.

Index