Lecture 28: Voice

1 Anatomy of the Voice

The voice is the oldest and in many ways most complicated of the musical instruments. We will discuss it in a series of lectures which will all be covered in this set of notes. They are broken up into a discussion of the anatomy of the vocal system, a discussion of vowel production, and a discussion of consonant production.

The vocal system is the upper part of the respiratory system. It has many parts, some completely familiar to everyone and some a little less so; we will spend more time describing the less familiar but important components.

The vocal tract is shown in figure 1. It begins at two openings to the outside, the nose and the lips. The nose leads into the **nasal cavity,** the mouth into the **oral cavity**. These meet at the back of the mouth and join into a single tube or cavity called the **pharynx**, which in turn divides into two tubes, the **esophagus**, which goes to the stomach, and the **trachea**, which goes to the lungs. The top of the trachea is a formation called the **larynx** or voice-box.

It should be emphasized that all of these anatomical features serve several roles, some of them more important than speech. As you know, the most important roles for the vocal tract are in breathing and in eating. The fact that the same features which are needed for these jobs (such as lips, teeth, and tongue, all essential in eating) can also be used so elegantly in sound production. Keep this in mind when we look at the form and function of each piece.

The nasal cavity runs as far back into the head as the oral cavity, and is quite complicated. It contains a series of bony plates with soft tissue on them, the olfactory organ, responsible for the sense of smell. It also connects to a number of air cavities further up into the head, called sinuses, which regulate the pressure in the brain cavity. Its role in speech is minor but not irrelevant. (You can see how big by pinching your nose closed and then talking. Most of the sounds are unchanged; a few, particularly "n," "m," and "ng," are ruined.) The cavity has no flexibility so there is no way to control its shape, which means that its role in speech is passive.

The oral cavity has a number of features everyone is familiar with, and some people are less familiar with. At the opening are the lips, then the teeth, which are held in by bones: the mandible below and the maxillary bones above. The maxillary bones also form the *front half* of the roof of the mouth, the **hard palate**, which is inflexible. The underside of the oral cavity is mostly taken up by the tongue, which (as you know) is a very flexible organ composed mostly of muscles. Motion of the tongue and lower jaw allows tremendous variability in the size and shape of the air cavity in the mouth; motion of the jaw and lips allow great control of the size and shape of its opening. The muscles of the jaw go up the

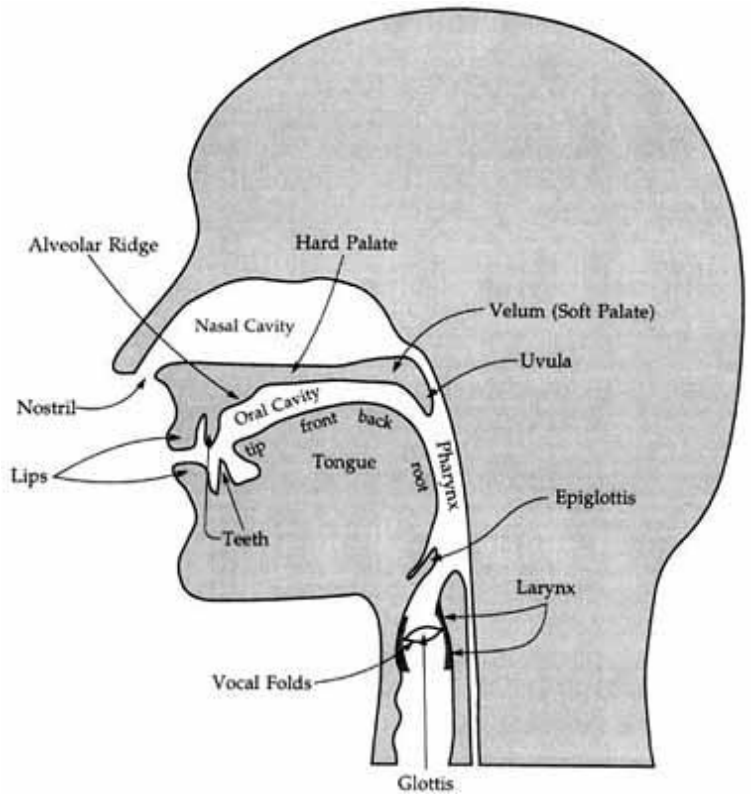

Figure 1: Main parts of the vocal system.

sides of the head and attach to the temporal bones of the skull. Most of the muscles of the tongue are anchored on a bone in the throat, called the **hyoid bone**. This horseshoe shaped bone is at the front top of the throat just above the voicebox. It is held in place by tendons attaching it to muscles and by cartilages; it is the only bone in the body which does not attach directly to any other bone.

The back of the roof of the mouth is the **soft palate** or velum. If you touch it with your finger or tongue, you will feel that it is soft and flexible (and will notice that it is the site of your gag reflex). It can move up and down a little. When it moves up, its tip, the uvula, closes off the nasal cavity from the pharynx (important in swallowing, breathing through the mouth, swimming, and much of speech). The oral cavity can be opened and closed by the base of the tongue, the tip of the tongue, or the lips (or any combination).

The pharynx is a slightly flexible tube, connecting at the top to the nasal and oral cavities and at the bottom to the trachea and esophagus. Everything you breathe and swallow goes into this one tube, which then splits again. Just as the soft palate and base of the tongue control whether the two cavities are open or closed into the pharynx, there *must* be a way to control which of the two tubes is open for things to go down into. The **epiglottis** is a flap at

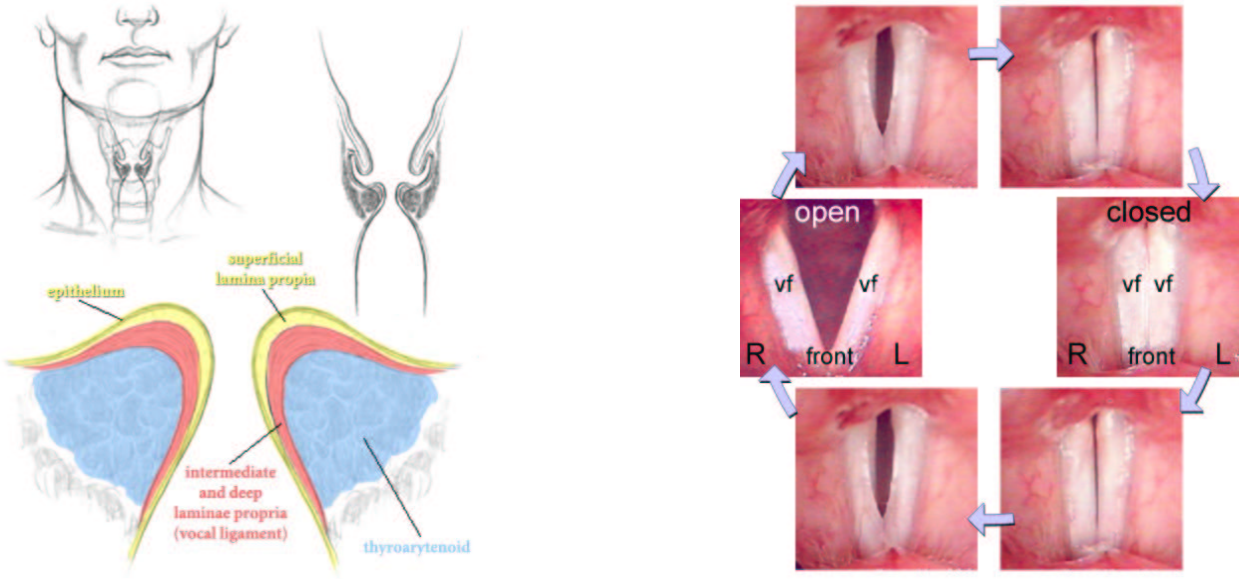

Figure 2: Vocal folds. Left: front view. Right: top view (actual photographs) of how they open and close.

the top of the larynx. Depending on whether it is raised or lowered, it either opens or closes the opening of the larynx. When you swallow you reflexively lower the epiglottis to close the larynx and trachea. This occasionally fails, usually when you are suddenly distracted. Your choking reflex is to use a burst of air to clear out anything which accidentally gets past the epiglottis.

What about the larynx, or "voice box"? It is a cartilage-cased "box" made up of cartilages, ligaments and tendons, muscles, and membranes. The airway goes through the larynx as a roughly tube shaped passage through its middle, but the walls of the tube have a number of pieces and processes which make it far from an even, straight tube. Only two will be of real interest to us. The first is the epiglottis, which we already described. The second are the vocal folds. These are two pieces of tissue, roughly semi-circular, one on either side of the trachea. They extend from the walls towards each other, mostly closing off the airway, see figure 2. They are made of membrane, tendon, and muscle. At the back, they are attached to two cartilages called the **arytenoids**; at the front they attach to a cartilage called the **thyroid cartilage**. In men the larynx is larger and the thyroid cartilage sticks out in the front of the throat, forming the "Adam's apple." A number of muscles attach to their edges and to the arytenoids, and a few muscles actually pass through the vocal folds themselves.

The vocal folds can be opened or closed by using the muscles attached to the arytenoids. One muscular action brings the arytenoids together; the vocal folds then meet in the middle and close off the airway entirely. Relaxing these muscles or using other muscles draws the

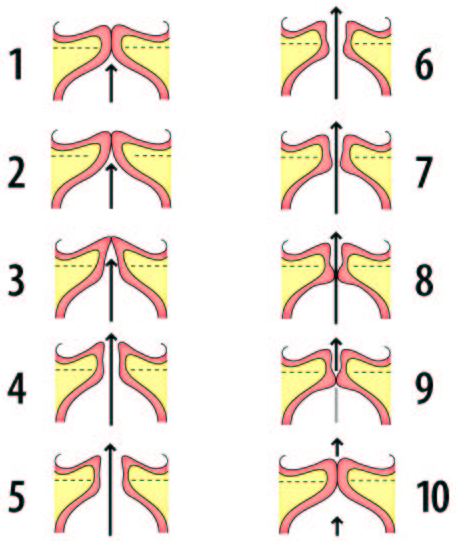

Figure 3: Time series of how the vocal folds move as air is forced through them during sound production.

arytenoids apart, so a space opens between the vocal folds. This gap, through which you breathe, is called the **glottis** (epiglottis means "above the glottis").

The vocal folds are *not* cords. They extend from the walls of the air passage almost or completely to the middle (depending on how you pull the muscles).

2 Sound production and Vowels

When you close the vocal folds and then force air up through them, they make sound. Let us see why. The muscles of the chest compress the lungs and raise the pressure of the air in the lungs and larynx up to the focal folds. This causes a situation rather like blowing out through your pursed lips. The pressure of the air underneath the vocal folds forces them up and open. A burst of air escapes. The vocal folds are not soft and flabby, though; they are elastic membranes, which spring back towards their natural position. They have a resonant frequency, set by their size, thickness, and stiffness (tension), and they will resonate at about this frequency, emitting a burst of air each period of the resonance. See figure 3. Each burst of air temporarily lowers the pressure just below the folds, so there is less upward force as they fall shut than as they open. Therefore the act of blowing through them increases the energy in the resonance, which is what keeps it going. Turning the air stream into a series of puffs generates sound, because it means the pressure and air speed above the vocal folds will vary with time. The sound will be periodic, with a frequency set by the resonant frequency

of the vocal folds.

Several properties of the sound can be controlled at the vocal folds. First, since there are muscles attached to the folds and within the folds, their shape, size, and tension can be adjusted. By relaxing all the muscles, the vocal folds are made loose, which gives a low frequency of vibration. By tensing the muscles within the vocal folds, they become thicker, but also much stiffer (more elastic). A thicker fold is heavier and slower to move, but a stiff fold springs back faster; the latter effect is larger, so the frequency goes up. By tensing the muscles attached to the edges of the fold, they can be made thinner and tighter, raising the frequency. An untrained voice (most of us) can change the frequency of vibration of the vocal folds by about a factor of 4 (two octaves). Vocal training not only increases the proficiency in controlling these muscles, but increases their strength and conditioning, and can increase the range of the voice. Since most of the muscular actions increase the frequency, it is easier to train someone to increase the upper end of their singing range than to enlarge the lower end of the singing range.

In typical speech, men use frequencies around 100 to 150 Hertz, women use frequencies around 200 Hertz, and children (whose whole apparatus is smaller and therefore produces higher frequencies) are more typically around 300 Hertz. The top range of male singing depends on the voice, and might be in the 350 to 450 Hertz range. Womens' singing voices can extend much higher, towards 1000 Hertz, though this also depends on the woman. Mens' voices are classed as **bass, baritone,** and **tenor**, in order from low to high. Women are **alto, mezzo-soprano,** and **soprano** in order from low to high. Alto is higher than tenor; womens' voices are almost invariably higher than mens'. A few men sing counter-tenor (same range as alto) and a few women sing contralto (same range as tenor).

The exact way the muscles are used to stiffen the vocal folds also has an influence on the timbre of the resulting sound. If the folds are not pulled tight all the way, then the folds do not fall all the way shut. This gives a "breathy" tone. If the arytenoids are pulled together tighter, the folds are shut for more of the sound producing cycle, giving more harmonics and a gruffer tone. The two strategies for raising the frequency of the voice also give different sounding voice. When the muscles within *and* around the vocal folds are tightened, the folds become thick and tense. They shut more tightly, producing more harmonics. This is called the "chest voice." When the muscles within the folds are left loose but the muscles around the folds are tightened, so the folds are thin but tight, they do not close completely and fewer harmonics are produced. This singing style is called the "head voice" or "falsetto." This sound making strategy can usually achieve higher pitches than the chest voice, often much higher pitches, but some people find that it sounds silly.

This explains how the frequency of your voice can be adjusted, and a little about the tone color. What makes the difference between different vowel sounds, like "ah," "eh," "oo," and so forth? Making these sounds will convince you that it is the mouth. The key is that the

sound is being produced at the bottom of a tube, running from the vocal folds to the lips, the nose, or both (depending on the position of the soft palate). The tube has resonances and the frequencies produced by the vocal folds which lie close to a resonance are enhanced.

The mouth is about 9 cm from front to back and the pharynx is about 8 cm from the back of the mouth to the top of the vocal folds. These numbers are approximate; they vary between individuals and they change slightly because the shape and length of the pharynx can be slightly adjusted by muscular action. *If* the mouth and pharynx formed a tube of uniform width, then since it is (almost) closed at the vocal folds and open at the mouth, the resonant frequencies would be at approximately,

$$f_{\text{res}} = \frac{v_{\text{sound}}}{4L} \times 1, 3, 5 = 500\,,\ 1500\,,\ 2500\,,\ \ldots \text{ Hz}$$

Note that all of these frequencies are substantially higher than the normal speaking range of frequencies. Therefore, unlike most musical instruments, *the human voice is not played at a resonant frequency of the air cavity*[1] . The frequency is determined by the tension and thickness of the vocal folds, which are heavy enough that their resonation is almost unaffected by the resonances in the air cavity above them. The role of the resonances of the vocal tract is entirely in strengthening certain harmonics of the voice and not others; namely, the harmonics which come close to coinciding with a resonant frequency of the vocal tract are enhanced. Because of this role in "forming" the timbre of the voice, the resonant frequencies of the vocal tract are called **formants**. They are numbered, from the lowest to higher frequencies, that is, they are called the **first formant** F_1, the **second formant** F_2, the **third formant** F_3, and so forth.

The resonant frequencies of the vocal system are not particularly high Q. This is mostly because the vocal folds are not completely closed for the full cycle. When they are partly open, they reflect most of the sound wave but let a part through, down the trachea to the lungs. The lungs are soft spongy tissues which almost perfectly absorb sound. This energy loss mechanism limits the Q. Recall that a resonance enhances any frequency which is close to the resonant frequency. Therefore the formants modify the tone color whether or not a harmonic of the sung frequency happens to coincide exactly with a formant frequency. The timbre of the voice is also affected by the efficiency of radiation from the mouth opening; like the wind instruments we have already discussed, a larger fraction of high frequencies escape an opening than for low frequencies, enhancing the harmonics by about 6 dB per octave. The influence of the formants and radiation efficiency on the spectrum of the voice is illustrated in figure 4.

[1]There is one notable exception: sopranos singing high in their register are singing as high as the lowest resonant frequency. Well trained operatic sopranos actually adjust their throat and mouth with the note they sing to shift the resonance to match the note they are singing, greatly enhancing loudness. This is why operatic sopranos are so famous (infamous) for their loudness.

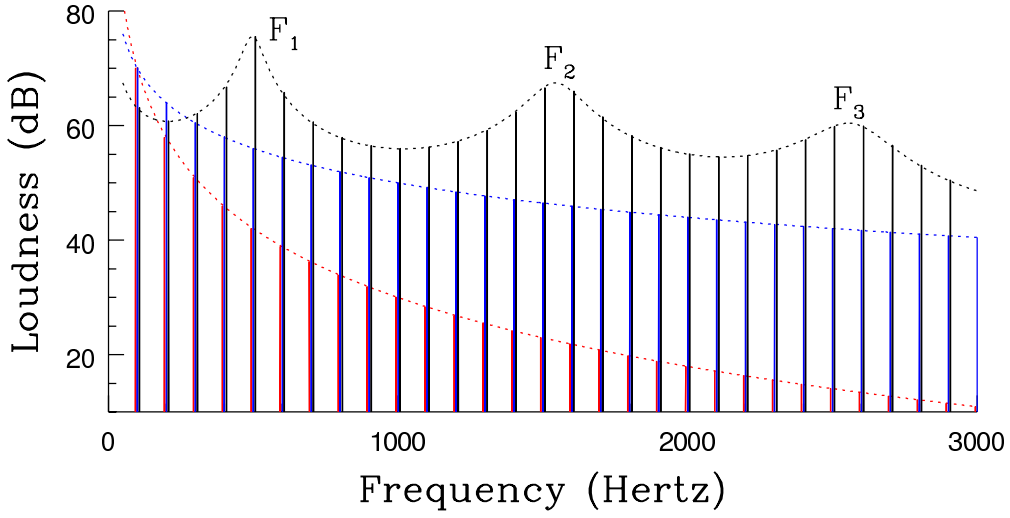

Figure 4: The effect of radiation and formants on sound. Since the voice is periodic, the frequency spectrum is a series of lines at the harmonic frequencies of the vocal frequency. The curve which shows how high these lines go is called the envelope. If the vocal folds were open directly to the air, the sound production would be like the red lines and envelope. Adding in the 6 dB per octave of enhancement for high frequencies to radiate out from the mouth opening gives the blue. The resonant enhancements near the formants gives the black.

The key to producing all the different vowel sounds is, that the shape of the vocal tract can be modified very widely and with great precision by adjusting the tongue, lips, and jaw (and to a lesser extent, the throat and soft palate). What this does is to change the resonant frequencies, modifying which harmonics of the voice are enhanced and which are not. For instance, in the letter "long e," (and even more in the German "ü"), the mouth and lips are narrowed. The pharynx is a wider tube and the mouth a narrower tube. This is a little like the wine bottle (Helmholtz resonator) we encountered in an earlier lecture. The bottom frequency F_1 is lowered, and the two next frequencies F_2 and F_3 move towards each other at around 2000 Hertz. This leads to an enhancement of low frequencies and of a range of frequencies around 2000 Hertz, but nothing in between. (Remember that the appearance of "ee" was close to a sine wave or sine plus sine with twice the frequency, with prominent, much higher frequency ripple. The low frequencies were from F_1, the prominent wiggles were from F_2 and F_3 being at close to the same frequency.) On the other hand, if you open the mouth and lips wide, "aah," then the tube is a narrower tube (pharynx) opening to a wider tube (mouth) opening to the world. The resonances are roughly the resonance of each tube, $f = v_{\text{sound}}/4(L/2)$; so F_1 and F_2 move together around 1000 Hertz, F_3 moves up in frequency around 3000 Hertz. All of the other vowels are other ways of holding the mouth which cause

the formants to move other places or change in Q.

Actually, in English many of the vowel sounds are actually **diphthongs**, that is, two sounds run together so tightly and so regularly that we think of them as a single sound. For instance, long i, as in "like", is really two sounds: "aah" and "yy". Long o, as in "own," is really "ooo" and "www". English uses diphthongs more than most languages.

3 Consonants

This does not begin to exhaust the list of tricks the vocal tract has for making sounds. The other tricks are grouped together as **consonants**. Most consonants involve the narrowing or closing of the airway to produce or modify a sound. However, the definition of what is a consonant and what is a vowel is made partly on usage (linguistics) rather than on the method of sound production. Some consonants involve the vocal folds; others do not. Some can be sustained; others have to be short.

The consonants in English are grouped into 5 types according to how the sound is produced:

- **Fricatives**: The fricatives are hissing sounds made then air is forced through a very narrow passage. This sound production is similar to how sound is produced by a flute, only without the flute. It relies on the physics of turbulence, which we will not try to understand in this course. The thing to know is, that when air passes rapidly through a short, narrow opening, it oscillates erratically as it emerges. If you are blowing through a horizontal slit, the air will alternately go somewhat upwards and downwards as it emerges. The time scale for this alternation is about twice the time it takes a bit of air to rush through the narrow spot (just as the ideal rate to blow across a flute is so the air takes about half a period to cross). For instance, if the air is going 50 meters/second and the slit is 5 mm across, the air needs .005/50 = .0001 seconds to get through, and the airstream will oscillate at around 5000 Hertz. It is *not* a periodic oscillation. It is noise made of frequencies in a range around this central value.

  The fricatives are the consonants made by forming a narrow spot in the airstream and forcing air through it. If the vocal folds are also vibrated, the fricative is **voiced**; if they are not, it is **unvoiced**. The location of the narrowing and resulting consonants are,

  - lip and teeth: "f" (unvoiced) and "v" (voiced)
  - tongue tip and upper teeth: "s" (unvoiced) and "z" (voiced)
  - tongue and both sets of teeth: "th" as in "thistle" (unvoiced) and "th" as in "this" (voiced)

- tongue and forward hard palate: "sh" (unvoiced) and "zh" (voiced)
- vocal folds themselves: "h"

Other languages have additional fricatives; for instance, in German the tongue and the back of the hard palate are used to produce "ch" as in "chanukah."

The fricatives sound different partly because the range of frequencies produced are different, and partly because the fricatives occurring further back in the mouth can be modified by the resonant structure still in front of them.

- **Plosives**: Plosives are consonants made by closing the airpath completely and then letting it open suddenly. There are three ways to do this: with airflow (aspirated); voiced; and both voiced and aspirated, a combination not used in English. (Neither voiced nor aspirated would not produce any sound.)

  The plosive sound depends on where the vocal tract is closed:

  - lips: "p" (aspirated), "b" (voiced), "bh" (both)
  - tongue and teeth: "t" (aspirated), "d" (voiced), "dh" (both)
  - tongue and hard palate: "k" (aspirated), "g" (voiced), "gh" (both)

  There are a number of other possibilities which are not used in English; the tongue and soft palate, the glottis (glottal stop), and so forth.

  Plosives cannot be "sustained," unlike vowels and the other consonants; they are always short.

- **Nasals**: Nasals are voiced sounds where the mouth is completely closed but the soft palate is moved to open the nose, so the sound emerges from there. The resonances of the pharynx-nasal cavity-oral cavity system vary depending on where the mouth is closed off, so different sounds are produced by closing the mouth at different locations:

  - mouth closed at the lips: "m"
  - mouth closed by tongue and hard palate: "n"
  - mouth closed by tongue and soft palate: "ng"

- **Liquids**: Liquids are voiced sounds where the airway is constricted almost as much as in a fricative, but not enough airflow is produced for a fricative sound. There are two in English, "r" and "l".

- **Semi-vowels**: Semi-vowels are vowel sounds which involve a narrowing of the airway which is more than in most vowel sounds. The two semi-vowels in English are "w" and

"y". The distinction between a semivowel and a vowel is made mostly on its linguistic use rather than on the sound production mechanism; both semi-vowels are used as parts of diphthong vowels in English, as well as their use as consonants.

To clarify, the above discussion does not exhaust the list of sounds the vocal system can make, it only lists the consonants commonly used in English (and a few which are not). The full set of sounds the vocal system can make is vast, and no language uses all of them. Different languages use different ones, and make finer distinctions between some sounds which other languages consider equivalent. This is all learned and cultural. Your vocal system is capable of making any sound used in any language. The difficulty is that your brain has not trained to use the sounds you do not need in the language(s) you speak, and your ear has not trained to recognize the important differences between sounds needed for some other languages. The vocal training needed is really nontrivial, because in normal speech you make 10 to 20 different vowel and consonant sounds per second, which means that the muscular motions have to be memorized. Training the vocal system becomes more difficult with age, which is why people who learn languages after the age of about 12 typically retain an accent.

The other aspect of speech is **prosodic** features. This is the rising and falling of the voice frequency, changes in loudness (accents and emphasis), and changes in speed or silent gaps. These convey additional meaning, such as urgency, sarcasm, emphasis, the distinction between a question and an answer, and so forth. In Indo-European languages their use is analogous to punctuation in written language. In *tonal languages,* they are varied syllable by syllable, and depending on how the pitch rises or falls, the syllable has a different meaning (is a different word). Mandarin, Cantonese, and Thai are examples of tonal languages.

4 Singing versus speech

The main thing which makes singing singing rather than speech is the way the pitch and timing of the voice are modified. Typically, each syllable is made the length of a note and is pitched to be steady on a note of the scale, rather than hopping up and down to convey prosodic information. This is the first reason that it is harder to understand song than speech. The other reason is that there are typically some modifications of the way the vocal tract is used, to increase the power of the voice at the expense of intelligibility.

Singers, especially trained choral or operatic singers, typically widen the jaw opening and change the mouth and lip configuration to increase sound radiation and power (think of your choral director yelling over and over to open your mouths wider). There is also a lowering of the larynx and a widening of the pharynx (increasing the size of the pharynx). This often leads to a resonance of the pharynx cavity, called the **singer's formant**, at approximately 3000 Hertz. This happens to be the most sensitive range of human hearing, and is also above

the dominant frequency region of most musical instruments. This is what allows a tenor to compete with an orchestra and still be heard.

What allows a soprano to compete with an orchestra and still be heard, especially in their upper register, is **formant tuning**. They are singing at a high enough frequency that a vocal formant *can* lie within the range of the fundamental of their voice. By adjusting the shape of the vocal tract, they can match a formant frequency of the vocal tract to the note they are singing. This causes a dramatic resonant enhancement of the loudness of the fundamental frequency. It is hard to do, because the vocal tract must be held differently for each note. This is particularly challenging on very fast passages. Therefore it is only highly trained operatic sopranos who can apply this technique. It also means that the note cannot be formed into any specific vowel sound. However, in this range of frequency (500 to 1000 Hertz), you could not tell apart different vowels anyway, because the harmonics of the voice are so far apart that they are not "sampling" the locations of the formants enough for you to tell where the formants lie. Therefore operatic writers only put words intended to be intelligible in the lower part of a soprano's range. It is the combination of sound power and the very large fraction of the sound power in the fundamental which allows opera singers to shatter wine glasses (if they are made of high quality crystal, so the vibration of the glass has a very high Q, and the singer tunes their sung frequency to coincide exactly with the vibration frequency of the glass, then the vibration of the glass grows so large that it shatters).