Chapter 22
Speech

Speech is our most basic and most important means of human communication. Speech conveys more than mere words. It conveys shades of meaning, emotion, attitudes, and opinions, sometimes even more completely than we as talkers would like. As a topic for study, human speech has endless fascination for acousticians, phoneticians, linguists, and philosophers. This chapter can only scratch the surface of this vast topic. It is mainly devoted to the *production* of speech by human talkers. It introduces the anatomy of the vocal tract and then describes how the components of the vocal tract function to create speech sounds, particularly vowels—the basis of singing.

To an unusual extent, the goal of this chapter is merely to cause the reader to become aware of things that he or she already knows from years of experience in talking. Properly read, this chapter requires some vocal involvement from the reader, if only to check that certain sounds are really made the way that the text says that they are made. If it's done right, it should be a noisy experience.

22.1 Vocal Anatomy

Speech sounds are made by using air from the lungs. The lungs provide all of the energy that ultimately ends up in a speech waveform. The air from the lungs passes through the trachea, to the larynx—including the vocal folds, and then to a part of the vocal tract that serves as a resonator. We consider those processes in turn (Fig. 22.1).

Diaphragm and lungs: The diaphragm is an arched sheet of muscle separating the lung cavity from the abdominal cavity. To understand what happens as you inhale, take a few deep breaths and use your hands to feel how your upper body moves.

W.M. Hartmann, *Principles of Musical Acoustics*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-6786-1_22,

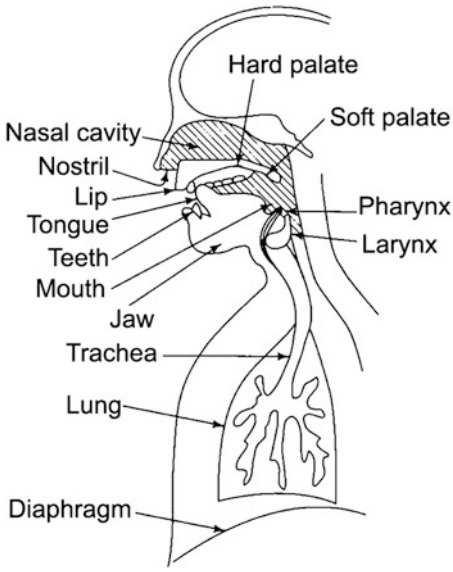

Fig. 22.1 Vocal anatomy, cross section after Farnsworth, 1940

Notice that when you inhale, your belly seems to protrude as the diaphragm is lowered to expand the lungs, and in the process moves other organs around lower down. Notice too that your upper chest walls rise. Both actions expand the chest cavity, allowing air to rush in because of normal atmospheric pressure. This process of inhalation, or inspiration, can expand the volume of air in the lungs from 2 to 7 L for the average adult human. However, in normal quiet breathing, the change in lung volume is only 10% of that.

Expiration (or exhalation) occurs when the normal elasticity of the lungs and chest walls forces air out of the lungs. Also, the lungs have a reserve capacity that can be made to expel additional air using expiratory muscles. When we stress certain syllables while speaking, the momentary added air volume is expelled by chest-wall muscles called "intercostal muscles."

A liter
The *liter* is a metric unit of volume. It is equivalent to 0.001 cubic meters or 1,000 cubic centimeters (cc). English equivalents to a liter are 61 $in.^3$ or slightly more than a USA quart.

Trachea: The trachea, commonly known as the "windpipe," allows air to flow in and out of the lungs. As the trachea descends toward the lungs, it divides into two bronchi (bronchial tubes), one for each lung. The trachea has a mucous lining that helps to catch dust particles, preventing them from entering the lungs. In that sense, its function is the same as the air intake and air filter on your car's engine.

Larynx: At the top of the trachea, and contiguous with it, is the larynx or voice box. The larynx serves many functions, not the least of which is to channel air to the trachea and food to the esophagus. Within the larynx is the *glottis*, consisting of the vocal folds (vocal cords)—two mucous membranes with a separation that is under exquisite, active, conscious control by the brain.

As you exhale normally through the trachea, the vocal folds are apart and air passes by them. If you hold your breath, the vocal folds are tight shut. When you make a *voiced* sound, the vocal folds are loosely together. The vocal folds vibrate against one another to make a buzz driven by a steady stream of air from the trachea. For unvoiced sounds the vocal folds are apart, as in normal breathing. Because there is no vocalization, unvoiced sounds have to be made by creating constrictions using pharynx, soft palate (vellum), tongue, teeth, and lips and then blowing air past these constrictions. The constrictions may be fixed in time, as for the sound "ssss," or the constrictions may suddenly open as for the sound of "p." In whispering, the vocal folds are somewhat apart, but they vibrate in a noisy way to create a sound that can be filtered later, just as for voiced sounds.

Resonators: The pharynx (pharyngeal cavity), mouth, and nasal cavities have resonances that act as filters, reinforcing particular frequency regions of the voiced or whispered sounds from the larynx. The frequencies of the resonances can be precisely controlled because of all the flexibility that you have with the shape of your mouth cavity and your tongue and lips. The different resonances are responsible for creating the different vowel sounds.

22.2 Voiced Sounds

The buzzing sound made by vibrating vocal folds is normally a periodic wave, and it has a complex spectrum with many harmonics. The vocal buzz, waveform and spectrum, is shown in Fig. 22.2. To some degree, the vocal buzz resembles the sounds of wind musical instruments—the brass instruments and woodwinds. Both the vocalized sounds and the wind instrument sounds begin with air from the lungs. Both sounds are generated by some physical element that vibrates with periodic motion in the air stream, as controlled by the intelligence of the person producing the sound. The analogy is particularly valid in the case of singing where sustained sounds have stable frequencies or frequencies that move only slowly or regularly. Physically, however, there is an important difference between the production of voiced sounds and the production of wind instrument sounds.

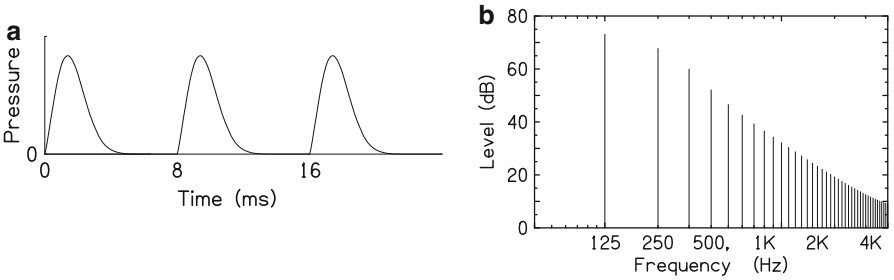

Fig. 22.2 (**a**) The waveform of a vocal sound measured in the throat shows peaks of air pressure when the vocal folds are open and intervals of zero pressure when the vocal folds are closed. The waveform is periodic with a period of about 8 ms, typical of an adult male voice. (**b**) The spectrum shows a fundamental frequency of 125 Hz (=1,000/8) and strong harmonics. Different shapes of oral cavities further upstream will lead to different vowel sounds. Note that in this figure, both axes are logarithmic. The harmonics are really equally spaced by 125 Hz, as expected for a periodic tone, but the logarithmic horizontal axis (a scale of octaves) does not give that impression

The vibrating element in the wind instrument, for example a trumpet player's lips or a clarinet player's reed, requires feedback from the rest of the instrument in order to produce the tone with the intended pitch. It is not so with voiced sounds. Although the parts of the vocal tract that follow the larynx (pharynx, mouth cavity, and nasal cavity) act as resonators or filters, they do not provide important feedback to the larynx. The larynx is capable of generating the voiced tone, with the correct frequency, all by itself without help from the rest of the vocal tract. Consequently, the human voice is easier to understand physically than are wind musical instruments. That is why the chapter on the human voice comes first in this book, before the chapters on musical instruments.

The most prominent voiced sounds are the vowels. Vowel sounds carry most of the energy in speech, though not necessarily most of the information. Because of their power, vowels are the main element in singing. It is easy to sing without consonant sounds. It is impossible to sing without vowels.

To make vowel sounds, one shapes the vocal tract—mouth and nasal cavities—to give these cavities particular acoustical resonances as shown in Fig. 22.3. These resonances cause particular harmonics of the vocal fold buzz to be emphasized. Thus, the vocal folds create a complex spectrum and the vocal tract filters it. The frequency regions containing strong harmonics due to vocal tract resonances are called "formants." It is usual to characterize vowel sounds by three formants, though it is possible to find as many as five.

The formant concept

The idea of a formant is nothing new. A formant is a band of frequencies that is prominent in the output of an acoustical source. This band of frequencies is prominent because the source includes a filter, or resonator, that emphasizes those frequencies.

The concept of a formant can be illustrated in a speech or music context in which the band of strong frequencies remains constant while the fundamental frequency of the sound changes. For instance, if a singer sings three different notes all with the vowel "AH" there are three different fundamentals, but the range of emphasized frequencies stays the same. A formant near 1,000 Hz strongly emphasizes the eighth harmonic of a 125-Hz tone, or the fifth harmonic of a 200-Hz tone, or the third harmonic of a 333-Hz tone.

A mute inserted into the bell of a brass instruments (e.g., trumpet or trombone) introduces a formant into the sounds of the instrument. This changes the tone color of every note played. For instance, a trumpet mute called a "cup mute" introduces a resonance peak extending from 800 to 1,200 Hz. Although the trumpet may play dozens of different notes, each with its own fundamental frequency and harmonic series, the formant, 800–1,200 Hz, applies equally to every note and leads to an unmistakable coloration of the sound.

Formant frequencies for different American vowel sounds are given in Table 22.1. The table shows the average formant frequencies for 76 talkers—33 men, 28 women, and 15 children. Average formants for adult women are higher than those for adult men because, on the average, vocal tracts are shorter for women. If you think of a vocal tract as a cylindrical pipe open at one end and closed at the other (Chap. 8), you are not surprised to learn that a shorter pipe has higher resonance frequencies. Children have much shorter vocal tracts and correspondingly much higher formant frequencies. As listeners, we associate high formant frequencies with small creatures. We also associate high fundamental frequencies with small creatures. The high fundamental frequencies of children occur because their vocal folds are lighter than for adults. Thus, high formant frequencies and high fundamental frequencies occur for very different reasons.

22.3 Speech Sounds

Years ago, you learned that speech is composed of vowels and consonants. This distinction is not necessarily wrong, but it greatly oversimplifies the phonological and linguistic nature of speech and language. Although it is not hard to define a

Table 22.1 Average formant frequencies and levels for vowels spoken by mainly American talkers

Vowel	EE	I	E	A	AH	AW	U	OO	UH
(as in)	heed	hid	head	had	hod	hawed	hood	who'd	hud
Frequencies (Hz)									
F1									
Men	270	390	530	660	730	570	440	300	640
Women	310	430	610	860	850	590	470	370	760
Children	370	530	690	1,010	1,030	680	560	430	850
F2									
Men	2,290	1,990	1,840	1,720	1,090	840	1,020	870	1,190
Women	2,790	2,480	2,330	2,050	1,220	920	1,160	950	1,400
Children	3,200	2,730	2,610	2,320	1,370	1,060	1,410	1,170	1,590
F3									
Men	3,010	2,550	2,480	2,410	2,440	2,410	2,240	2,240	2,390
Women	3,310	3,070	2,990	2,850	2,810	2,710	2,680	2,670	2,780
Children	3,730	3,600	3,570	3,320	3,170	3,180	3,310	3,260	3,360
Relative levels (dB)									
L1	−4	−3	−2	−1	−1	0	−1	−3	−1
L2	−24	−23	−17	−12	−5	−7	−12	−19	−10
L3	−28	−27	−24	−22	−28	−34	−34	−43	−27

steady vowel, as we have done with the formant concept, the variety of human speech behavior makes even the study of vowel sounds a complicated one. A thorough approach to a more advanced treatment is beyond the scope of this book. This section on speech sounds will deal with the some classes of speech sounds with emphasis on the means of production.

diphthongs: Within the vowel context there are diphthongs, combinations of two vowels that are interpreted as a single speech sound. The word "toil" is an example. As for the word "out" from the speech spectrogram, it is impossible to pronounce the vowel sound in "toil" without changing the shape of the vocal tract. Even the pronunciation of the first letter of our English alphabet seems to end in an EEE sound if you listen to it closely.

glides: A glide is a vowel-like sound that requires the vocal tract to move. The WH sound in "when" is an example.

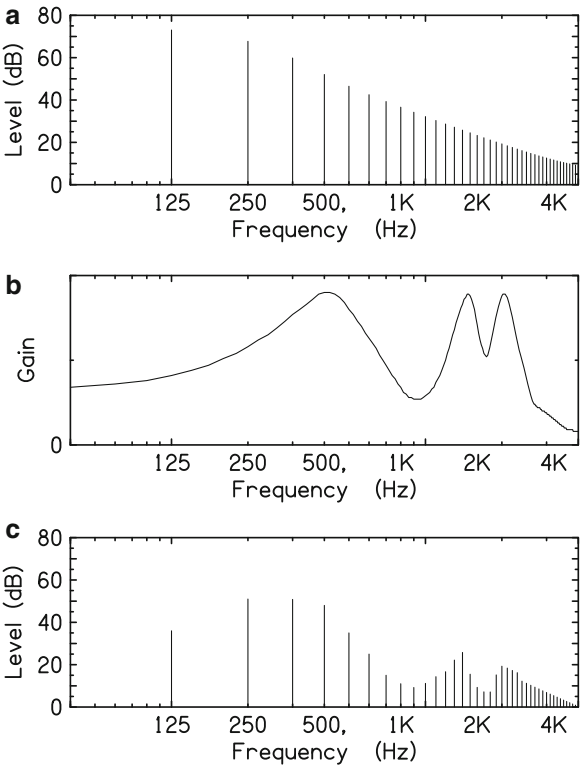

Fig. 22.3 Creating a voiced vowel. (**a**) The spectrum of a vocal sound measured in the throat from Fig. 22.2. (**b**) The gain of the vocal tract filter making the vowel E as in "head." (**c**) The spectrum of the vowel finally produced by the talker. The peak near 500 Hz is the first formant. It appears to be broader than the other peaks, but that is a distortion of the logarithmic plot. In units of linear frequency, the first formant peak is actually the narrowest

fricatives: Fricatives are consonant sounds that are essentially noise. They are made by creating a constriction somewhere in the vocal tract and blowing air past the constriction to make a turbulent flow with a noisy character. The FF sound is made with a constriction between the upper teeth and lower lip. To make the TH sound you make a constriction with your tongue in back of your teeth. The HH, as in "hah" comes from a constriction in the glottis itself. The sounds known as sibilants are the hissy fricatives like SS and SH. Spectrally, fricative sounds are broadband with no harmonic structure. However, different fricatives have different frequency regions of spectral strength. For instance, the SS sound has much greater high-frequency power than the SH sound.

plosives: As the name implies, a plosive is a little explosion. Plosives (sometimes called "stops") are made by building up air pressure behind a closed constriction and then releasing it. The plosives P,T, and K are not voiced. The plosives B,D, and G are normally accompanied by voicing. Plosives, as well as fricatives, are said to be manners of articulating consonant sounds.

22.4 Spectrograms

A spectrogram, for example Fig. 22.4, is a plot of power in a vocalization as a function of frequency and time. The frequency is plotted on the vertical axis and runs from 0 to 4,000 or 5,000 Hz because this is the vital frequency range for speech. The time is along the horizontal axis, as for an oscilloscope, but the time scale is much longer than the time for a trace on the oscilloscope. From left to right the time is several seconds—enough time to speak an entire sentence. The power that occurs at a particular frequency (vertical axis) and at a particular time (horizontal axis) is shown by darkness.

Figure 22.4 shows a spectrogram for the utterance, "Joe took father's shoe bench out." The first sound that occurs is the "J" sound in "Joe." Notice that it is a noise with most of its power between 2 and 4 kHz. Next comes the "OH" sound in "Joe," and it shows motion in time. The formants are moving. There is some weak noise energy for the "t" sound in "took," and then comes the short "oo" vowel sound in "took." The short "oo" vowel is too short for much motion. It appears to show dark regions near 400, 1,300, 2,200, and 3,200 Hz. We can compare that with the "U" sound in Table 22.1, as in "hood." If this is a male talker (we don't really know from this spectrogram), then formants are expected at 440, 1,020, and 2,240 Hz. These first three formants agree pretty well with the dark regions, though 1,020 Hz seems low compared to 1,300 Hz. This discrepancy indicates that individual talkers aren't necessarily the same as the average talker.

Figure 22.4 shows that the "s" sound that ends the word "father's" and the "sh" sound that begins the word "shoe" are totally merged in time. The breaks that experienced listeners perceive between different words may have no real acoustical

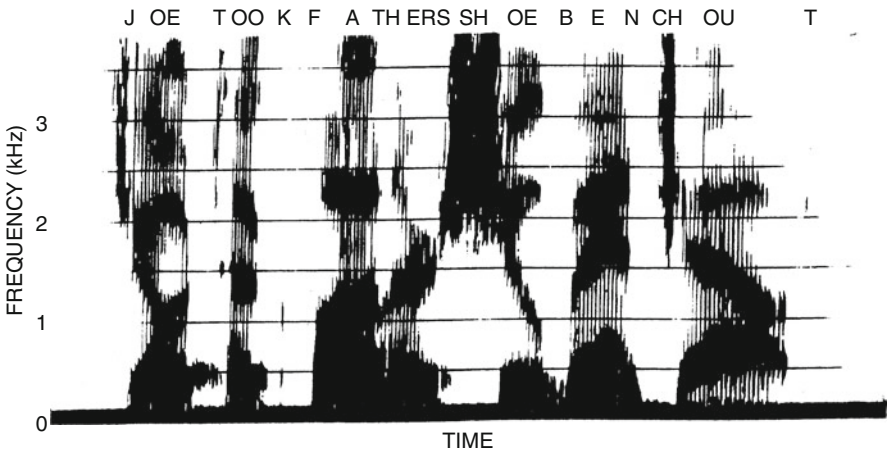

Fig. 22.4 Speech spectrogram. Power is indicated by *dark regions*. From W.J. Strong, J. Acoust. Soc. Am. **41**, 1434 (1967)

existence in the speech waveform. The segmentation of different words from the continuum can be a problem for listeners who are less familiar with the language.

Something is moving a lot in the "oe" sound that ends the word "shoe." Perhaps it is a second formant that is moving. In any case, there is no necessary motion in pronouncing the word "shoe." You will find it possible, though perhaps a little unnatural, to say the word "shoe" with all your articulators in a fixed position. Individual modes of speaking show greater or lesser motion. By contrast, it is impossible to pronounce the word "out" without a lot of motion of the mouth. That can be seen in the huge time variation in the first and second formants at the end of the sentence in Fig. 22.4. It looks as though the second formant drops from about 1,500 to 1,000 Hz.

The speech spectrogram is a very useful method of studying speech. It is detailed, but not too detailed. Spectrograms of different individuals saying the same sentence can show what parts of speech are essential to convey the intended words and what parts are individual differences. The spectrogram gives clues as to the articulatory gestures used in speech, but they are only cues because there is more than one way to make any given speech sound.

Exercises

Exercise 1, Breathing-in speech.

Prove that it is not necessary to breathe out when you speak. It is possible to speak while breathing in. Can you learn to control this kind of speech?

Exercise 2, We are all windbags!

The text says that the lung capacity of the average human is about 5 L. A table of weights and measures says that 1 L is 0.2642 gallons (that's 0.22 British imperial gallons). (a) Calculate the average lung capacity in gallons; (b) in quarts.

Exercise 3, Feeling the buzz

(a) Feel the buzz of your vocal folds when you make a voiced sound. (b) Which of the following sounds are voiced: FF, SS, B, V, P, T?

Exercise 4, The nose knows.

(a) Which speech sounds can you make with your mouth closed? (b) Practice making the transition between "M" and "N" with your mouth closed and in one position. Can you get a friend to distinguish between these two sounds?

Exercise 5, Diphthongs or not?

Say the vowels as you learned them in school: A, E, I, O, and U. Which vowels would you expect to sound most nearly the same if played backwards?

Exercise 6, Plosive sounds

Where are the "explosions" taking place in your mouth in the sequence, "Pah, Tah, Kah, Gah?" Does this sequence go from front to back or from back to front?

Exercise 7, Fricative sounds

Fricative sounds are made by blowing air past a constriction. Where are the constrictions in the sequence, "Sah, Fah, Thah, Hah?"

Exercise 8, Whispered vowels

Say a vowel out loud. Now whisper the same vowel. Did you have to change the shape of your vocal tract when you started to whisper? Why?

Exercise 9, Chimeras

The discussion of Table 22.1 noted that children have high fundamental frequencies because their vocal folds are light, and they have high formant frequencies because their vocal tracts, from the pharynx on up, are small. Can you imagine a creature with low fundamental frequencies and high formants or vice versa?

Exercise 10, Read a spectrogram

In the spectrogram, identify the spectral character of vowel formants, formant transitions, fricatives (like sibilants), and plosives.

Exercise 11, The remarkable "AH"

Refer to Table 22.1 and notice the separation in frequency between the first and second formants of the AH and AW vowel sounds as in "hod" and "hawed." Notice that the separation is the smallest separation for any of the formants of any of the vowels in the table. Figure 22.4 has an AH sound in "father." Can you now account for the big blob of power below 1,500 Hz in the spectrogram of that word?

Exercise 12, Benched!

Look at the blobs of power in the "E" sound in the word "bench" in Fig. 22.4. What are the frequencies for these regions? Do they agree with the formant frequencies in Table 22.1 for the "E" sound as in "head?"