CHAPTER

15 Speech Production

Throughout human history, the principal mode of communication has been the spoken word. The systems in the human body that send and receive oral messages are sophisticated in design and complex in function. Our understanding of both hearing and speech has progressed dramatically in recent years, largely because of new techniques for making acoustical as well as physiological measurements. The auditory system and its function were discussed in Chapter 5, and now it is appropriate to devote the same attention to the vocal organs.

In this chapter you should learn:

- How the human vocal organ makes speech sounds;
- How speech sounds are the product of the source, the filter function, and the radiation efficiency;
- About speech articulation by the different parts of the vocal tract;
- About formants as resonances of the vocal tract;
- How the glottis and the vocal tract are studied by speech acousticians.

15.1 ■ THE VOCAL ORGANS

The human vocal organs, as well as a representation of the main acoustical features, are shown in Fig. 15.1. The lungs serve as both a reservoir of air and an energy source. In

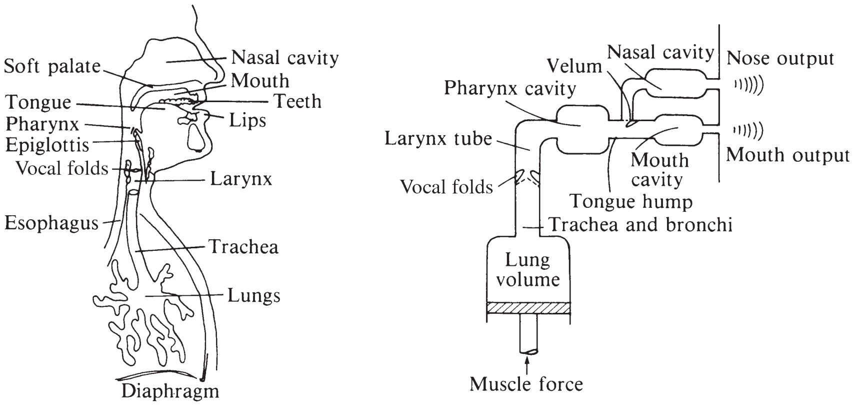


FIGURE 15.1 Human vocal organs and a representation of their main acoustical features. (After Flanagan 1965)

speaking, as in exhaling, air is forced from the lungs through the larynx into the three main cavities of the vocal tract: they pharynx and the nasal and oral cavities. From the nasal and oral cavities, the air exits through the nose and mouth, respectively.

Air can be inhaled or exhaled with little generation of sound if desired. In order to produce speech sounds, the flow of air is interrupted by the vocal cords or by constrictions in the vocal tract (made with the tongue or lips, for example). The sounds from the interrupted flow are appropriately modified by various cavities in the vocal tract and are eventually radiated as speech from the mouth and, in some cases, the nose.

15.2 ■ THE LARYNX AND THE VOCAL FOLDS

The most important sound source in the vocal system is the *larynx*, which contains the *vocal folds* or *vocal cords*. The larynx is constructed mainly of cartilages, several of which are shown in Fig. 15.2. One of the cartilages, the thyroid, forms the projection on the front of the neck known as the Adam's apple.

The vocal folds are not at all like cords or strings, but consist rather of folds of ligament extending from the thyroid cartilage in the front to the arytenoid cartilages at the back. The arytenoid cartilages are movable and control the size of the V-shaped opening between the vocal cords, which is called the *glottis*. Figure 15.3 shows how the arytenoids control the size of the glottis. Normally, the arytenoids are positioned well apart from each other to permit breathing; however, they come together when sound is produced by the vocal folds.

The vocal folds may act on the air stream in several different ways during speech. From a completely closed position in which they cut off the flow of air, they may open suddenly as in a light cough or a glottal stop (such as the glottal "h" that occurs in Cockney English). A glottal stop may also give a hard beginning to a vowel sound, such as the "Idiot!" expressed vehemently. On the other hand, the vocal folds may be completely open for un-

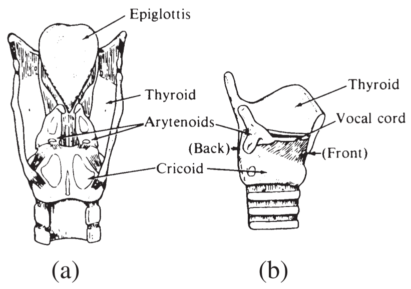


FIGURE 15.2
Various views of the larynx: (a) back; (b) side.

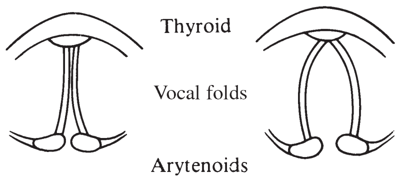


FIGURE 15.3
Control of the glottal opening by the arytenoids.

voiced consonants such as “s,” “sh,” “f,” etc. An intermediate position occurs in the “h” sound, where the air stream interacts lightly as it passes between the vocal folds.

The most useful function of the vocal folds, however, is to modulate the air flow by rapidly opening and closing. This rapid vibration produces a buzzing sound from which vowels and voiced consonants are created. These functions of the vocal folds are somewhat analogous to the functions of the lips. The glottal stop corresponds to the action of the lips in a plosive consonant such as p; the light friction of the folds in producing the “h” sound corresponds to the action of the lips in pronouncing “f.” The rapid vibration of the vocal folds is similar to the rolling noise made by a child’s lips to imitate a motor, or the sound used to indicate coldness (“brrr”), or a trumpet player buzzing his or her lips in a practice exercise. You can feel the vibrations set up by the vocal folds by placing a finger lightly against your Adam’s apple. Make sounds “zzzzzz” and “ssssss” alternately to turn the vibrations on and off (these are examples of voiced and unvoiced consonants, respectively).

The rate of vibration of the vocal folds is determined primarily by their mass and tension, although the pressure and velocity of the air do contribute in a smaller way. The vocal folds are typically longer and heavier in the adult male than in the female and, therefore, vibrate at a lower frequency (pitch). During normal speech, the vibration rate may vary over a 2 : 1 ratio (one octave), although the range of a singer’s voice is more than two octaves. Typical frequencies used in speech are 110 Hz in the male, 220 Hz in the female, and 300 Hz in the child, with wide variations from one individual to another.

Speech scientists describe three different modes in which the vocal folds can vibrate. In the normal mode, they open and close completely during the cycle and generate puffs of air roughly triangular in shape when air flow is plotted against time. In the open phase mode, the folds do not close completely over their entire length, so the air flow does not go to zero. This produces a breathy voice, sometimes used to express shock (“No!”) or passion (“I love you”). A third mode, in which a minimum of air passes in short puffs, gives rise to a creaky voice, such as might result if you attempt to talk while lifting a heavy weight. A fourth mode, called *head voice*, or *falsetto*, is normally not used in speech and is discussed in Section 17.6.

The vocal folds are caused to open by air pressure in the trachea, which tends to blow them upward and outward. As the air velocity increases, the pressure decreases between them, and they are pulled back together by the Bernoulli force (see Section 11.4). Ordinarily, however, the restoring force supplied by the muscles exceeds the Bernoulli force. Feedback from the vocal tract has relatively little influence on the vocal fold vibrations (as compared with the close cooperation between the air column in a brass instrument and the player’s lips, for example; see Section 11.3).

Another use of the vocal folds is in the production of a whisper. For a quiet whisper, the vocal folds are in much the same position as they are for an “h” sound. For louder whispers, the folds are brought closer so as to interfere more strongly with the flow of air. Say the word “hat” in a loud whisper and note the different rate of air flow during the “h” and the “a” by holding your hand in front of your mouth.

The vocal folds can be observed by placing a small dental mirror far back in the mouth. Using this technique, several investigators have taken high-speed motion pictures (4000 frames/s) of the vocal folds in vibration. Figure 15.4 illustrates this technique, and Fig. 15.5

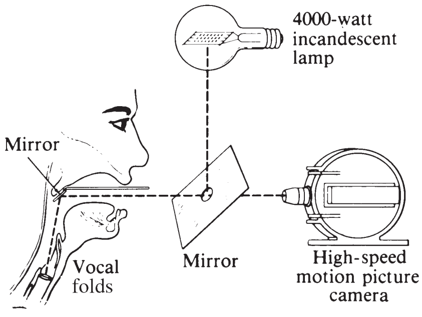


FIGURE 15.4 Technique for high-speed motion picture photography of the vocal folds. (From Flanagan 1965)

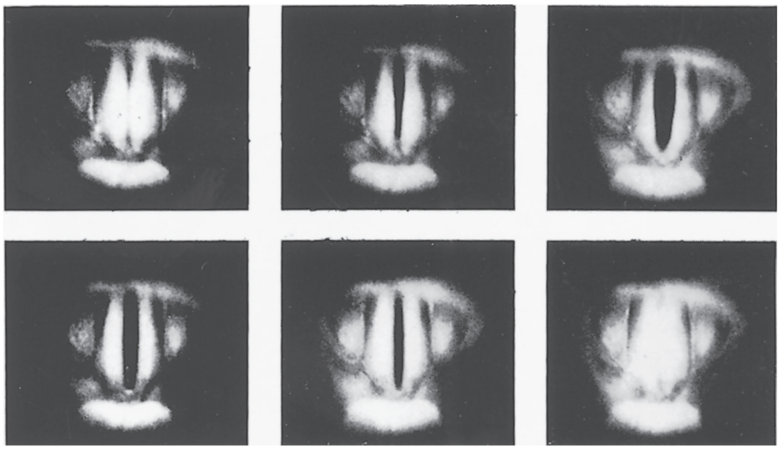


FIGURE 15.5 Successive phases in one cycle of vocal fold vibration. The total elapsed time is approximately 8 ms. (From Flanagan 1965)

shows one cycle of vocal fold vibration at a frequency of about 125 Hz. Less obtrusive fiber optic probes inserted into the throat allow continuous observation of the vocal folds.

The flow of air through the glottis is roughly (though not exactly) proportional to the area of the glottal opening. For normal vocal effort, the waveform of the air flow is roughly triangular in shape with a duty factor (that is, the ratio of time open to the total period of a vibration) of 30 to 70%, as shown in Fig. 15.6. The sound resulting from this interrupted air flow is characterized as a "buzz" and is rich in overtones. A triangular waveform is composed of harmonics that diminish in amplitude as $1/n^2$ (at a rate of 12 dB/octave), and the sound spectrum of the output of the larynx shows approximately this character for the higher harmonics, as seen in Fig. 15.7.

One might guess that the loud speaking would require greater air pressure from the lungs and a greater amplitude of vocal fold vibration. This is only partly true. Even if we assume the pressure in the trachea to be constant, the pressure in the larynx will fluctuate due to standing waves in the vocal tract. It may be a bit surprising to learn that one of the most important parameters affecting loudness of phonation is the *rate of glottal closure*. Rapid closure introduces higher harmonics in the glottal airflow spectrum, and these harmonics excite resonances of the vocal tract, leading to considerable buildup in sound level.

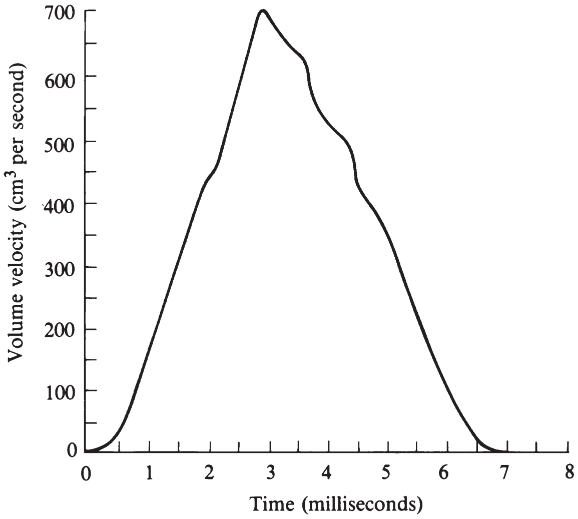


FIGURE 15.6 The variation of air flow in a glottal puff. The curve repeats once every 8 ms (a frequency of 125 Hz).

Although the vocal folds serve as the principal source of sound in speech, other sources are used, especially in the production of unvoiced consonant sounds. Sounds such as "f," "th," "s," "sh," (fricative consonants) and "l" are produced by a turbulent flow of air through a constriction somewhere in the vocal tract. The spectrum of such turbulence is quite broad, with many overtones that are not harmonic. Another source of sound is generated by a sudden release of pressure, such as that used in the plosive consonants p, t, and k. The sounds of consonants will be discussed in Section 15.4.

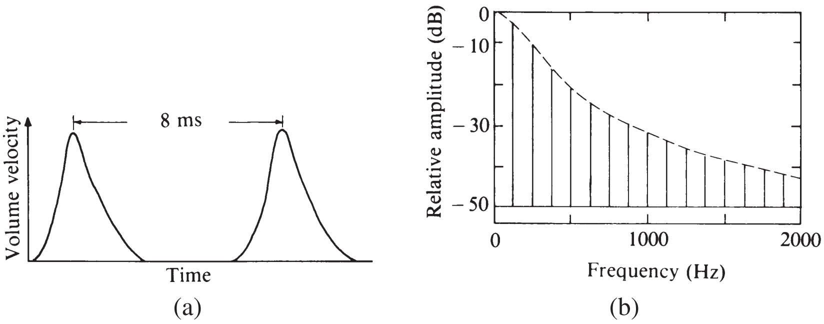


FIGURE 15.7 A typical waveform of the volume velocity of the glottal output for a fundamental frequency of 125 Hz, and a Fourier spectrum corresponding to this type of waveform. (From Stevens and House 1961.)

15.3 ■ THE VOCAL TRACT

The function of the vocal tract is a most remarkable one; it transforms the "buzzes" and "whooshes" from the vocal folds and other sources into the intricate, subtle sounds of speech. This demanding function is accomplished by changes in shape to produce various acoustic resonances. Intensive studies in recent years have produced a great deal of information concerning the details of how this is accomplished. This information is the heart of a branch of science called *acoustical phonetics*.

The vocal tract, as shown in Fig. 15.1, can be considered a single tube extending from the vocal folds to the lips, with a side branch leading to the nasal cavity. The length of the tube is typically about 17 cm, which can be varied slightly by raising or lowering the larynx and by shaping the lips. For the most part, however, the resonances in the vocal tract are tuned by changing its cross-sectional area at one or more points.

The *pharynx* connects the larynx with the oral cavity. It is not easily varied in shape, although its length can be changed slightly by raising or lowering the larynx at one end and the soft palate at the other end. The soft palate also acts as a valve to isolate or connect the nasal cavity to the pharynx. Since food also passes through the pharynx on its way to the esophagus, valves are necessary at the lower end to prevent food from reaching the larynx and to isolate the esophagus acoustically from the vocal tract. The *epiglottis* serves as such a valve, with the “false vocal cords” at the top of the larynx serving as a backup in case some food gets past the epiglottis. The epiglottis, false vocal cords, and vocal folds (cords) are open during normal breathing but closed during swallowing, thus forming a triple barrier to protect the windpipe. The epiglottis and false vocal cords do not appear to play any significant role in the production of speech.

The *nasal cavity* has fixed dimensions and shape, so that it is virtually untunable. In the adult male, the cavity has a length of about 12 cm and a volume on the order of 60 cm^3. The soft palate serves as a valve to control the flow of air from the pharynx into the nasal cavity. If the soft palate is lowered, air and sound waves flow into the nasal cavity and a nasal effect results from resonance within the nasal cavity. If, at the same time, flow through the mouth is blocked, air and sound exit through the nose, and humming results. Nasalized vowel sounds, which are common in French, are made by allowing sound to exit through both the mouth and the nose.

You can observe the soft palate in a mirror as it moves up and down (closing the nasal cavity in the up position). Say “ah” and the soft palate will rise; relax, and it will lower to normal breathing position. A cleft palate is a defect that allows air into the nasal cavity for all sounds, even those that should be entirely oral.

The *oral cavity*, or mouth, is probably the most important single part of the vocal tract because its size and shape can be varied by adjusting the relative positions of the palate, the tongue, the lips, and the teeth. The *tongue* is very flexible; its tip and edges can be moved independently or the entire tongue can move forward, backward, up and down. Movement of the lips, cheeks, and teeth also changes the size, shape, and acoustics of the oral cavity.

The lips control the size and shape of the mouth opening through which sound is radiated. Since the mouth opening is small compared to the wavelength of most components of the radiated sound, the size and shape of the opening are not of particular significance, except as they affect the all-important resonance frequencies of the oral cavity (this will be discussed further in Chapter 17). The mouth radiates more efficiently at higher frequencies where the wavelength approaches the size of the opening. In fact, a rise of 6 dB per octave in radiation efficiency is a good approximation to this effect.

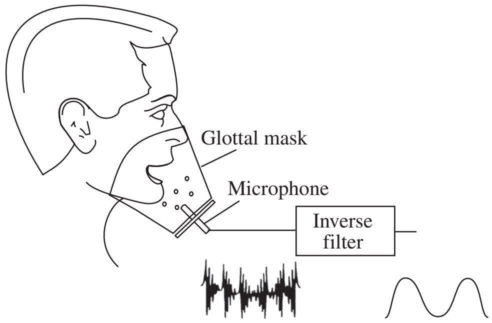
The spectrum envelope of speech sound can be thought of as the product of three components:

$$\text{Speech sound} = \text{source} \times \text{filter function} \times \text{radiation efficiency}.$$

If each of these quantities is expressed in decibels, then the contributions are added rather than multiplied. When the source consists of the vocal folds vibrating in their usual manner, the source function decreases in strength approximately 12 dB per octave (see Fig. 15.7). To this should be added the radiation efficiency of the mouth (which rises approximately 6 dB per octave) giving a net decrease of 6 dB per octave due to the first and last terms in the equation above. It remains to consider the more complicated way in which the filter function of the vocal tract varies with frequency, and that is the subject of Section 15.5.

Inverse Filtering and the Glottogram

Speech sound, we learned, is the product of the glottal flow (source), the vocal tract (filter), and the mouth opening (radiator). It is difficult but not impossible to study these independently. One way to study the glottal flow, for example, is to cancel the filtering effect of the vocal tract by *inverse filtering*. The subject speaks into a mask that measures (by means of a flow resistance and a pressure microphone) the waveform of the air flow. The signal then passes through a set of filters that are adjusted to remove the formants as well as possible. This works quite well, especially for low-pitched phonation and for open vowels. The waveform after filtering is called a *glottogram*, and it gives a fairly accurate representation of the glottal air flow in various modes of phonation.



15.4 ■ ARTICULATION OF SPEECH

Before discussing the resonances of the vocal tract, it is appropriate to briefly describe the articulation of English speech sounds, or *phonemes*. In speech structure, one or more phonemes combine to form a syllable, and one or more syllables to form a word. Phonemes can be divided into two groups: vowels and consonants. Vowel sounds are always voiced; that is, they are produced with the vocal folds in vibration. Consonant sounds may be either voiced or unvoiced.

Various speech scientists list from 12 to 21 different vowel sounds used in the English language. This discrepancy in number comes about partly because of a difference of opin-

TABLE 15.1 The vowels of Great American English

Pure vowels						Diphthongs		
ee	heat	/i/	aw	call	/ɔ/	ou	tone	/oʊ/
i	hit	/ɪ/	u̇	put	/ʊ/	ei	take	/eɪ/
e	head	/ɛ/	oo	cool	/u/	ai	might	/aɪ/
ae	had	/æ/	ǔ	ton	/ʌ/	au	shout	/aʊ/
uh	the	/ə/	er	bird	/ɜ/	oi	toil	/ɔɪ/
ah	father	/ɑ/				ju	fuse	/ju/

ion as to what constitutes a pure vowel sound rather than a *diphthong* (a combination of two or more vowels into one phoneme). Table 15.1 lists the vowel sounds of Great American, the dialect of English spoken throughout most of western and midwestern United States. Also given are the corresponding symbols from the International Phonetic Alphabet (Denes and Pinson 1973). Figure 15.8 shows the approximate tongue positions for articulating these vowels.

Whereas the vowel sounds are more or less steady for the duration of the phoneme, consonants involve very rapid, sometimes subtle, changes in sound. Thus consonants tend to be more difficult to analyze and to describe acoustically.

Consonants may be classified according to their *manner of articulation* as plosive, fricative, nasal, liquid, and semivowel. The *plosive* or stop consonants (p, b, t, etc.) are produced by blocking the flow of air somewhere in the vocal tract (usually in the mouth) and releasing the pressure rather suddenly. The *fricatives* (f, s, sh, etc.) are made by constricting the air flow to produce turbulence. The *nasals* (m, n, ng) are made by lowering the soft palate to connect the nasal cavity to the pharynx and then blocking the mouth cavity at some point along its length. The *semivowels* or glide consonants (w, y) are produced by keeping the vocal tract briefly in a vowel position and then changing it rapidly to the vowel sound that follows; thus, semivowels are always followed by a vowel. In sounding the *liquids* (r, l), the tip of the tongue is raised and the oral cavity is somewhat constricted.

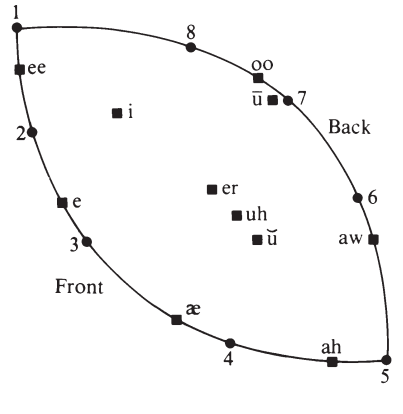


FIGURE 15.8 Approximate tongue positions for articulating vowels listed in Table 15.1. Numbers 1–8 are the eight cardinal vowels, which serve as a standard of comparison between languages.

TABLE 15.2 The classification of English consonants

	Manner of articulation						
	Plosive		Fricative		Nasal	Semivowel	Liquids
Place of articulation	Unvoiced	Voiced	Unvoiced	Voiced			
Labial (lips)	p	b			m	w	
Labiodental (lips and teeth)			f	v			
Dental (teeth)			th /θ/ (thin)	th /ð/ (then)			
Alveolar (gums)	t	d	s	z	n	y /j/	l, r
Palatal (palate)			sh /ʃ/	zh /ʒ/			
Velar (soft palate)	k	g			ng /ŋ/		
Glottal (glottis)			h				

Phonetic symbols are given where they differ from the English letter.

Consonants are further classified according to their *place of articulation*, primarily the lips, the teeth, the gums, the palate, and the glottis. Terms used by speech scientists to denote place of articulation include *labial* (lips), *dental* (teeth), *alveolar* (gums), *palatal* (palate), *velar* (soft palate), *glottal* (glottis), and *labiodental* (lips and teeth). Finally, consonants are classified as to whether they are *voiced* or *unvoiced*.

Twenty-four consonants of English are thus classified in Table 15.2. Note the seven pairs of voiced/unvoiced consonants. In addition, the pair /tʃ, dʒ/, which refer to the "ch" (church) and "j" (judge) sounds, are sometimes included as separate consonants, although each of them consists of a plosive followed by a fricative (ch $\simeq$ t + sh; j $\simeq$ d + zh). Consonants are more independent of language and dialect than vowels are.

15.5 ■ RESONANCES OF THE VOCAL TRACT: FORMANTS

The *vocal tract* consists of three main sections: the pharynx, the mouth, and the nasal cavity. These can be shaped by movements of other vocal organs, such as the tongue, the lips, and the soft palate (see Fig. 15.1).

Although the pitch and intensity of speech sounds are determined mainly by the vibrations of the vocal folds, the spectrum of these sounds is strongly shaped by the resonances of the vocal tract. It is the character of these resonances that distinguishes one phoneme from another.

The peaks that occur in the sound spectra of the vowels, independent of the pitch, are called *formants*. They appear as envelopes that modify the amplitudes of the various harmonics of the source sound. Each formant corresponds to one or more resonances in the vocal tract. Formant frequencies are virtually independent of the source spectrum.

Figure 15.9 illustrates the effect of formants on the source sound from the larynx. Both the waveform and the spectrum of the source sound are shown along with the waveform and the spectrum of the transmitted speech sound. (Note that in the waveform graphs, the horizontal axis is time; in the spectra, the horizontal axis is frequency.)

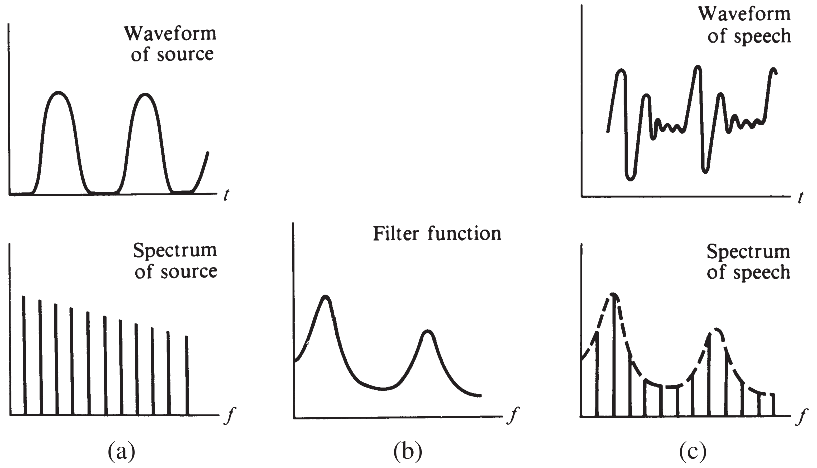


FIGURE 15.9 The effect of formants on sound: (a) waveform and spectrum of source sound; (b) filter function showing two formants (resonances); (c) waveform and spectrum of transmitted sound. t = time; f = frequency.

Table 15.3 gives the average formant frequencies for the vowel sounds of men, women, and children. The first nine rows give the average frequencies of the first three formants. The last three rows indicate the relative strengths of the three formants for each vowel. For example, /ɑ/ (ah) has the strongest second formant, only 4 dB weaker than the first formant. For /i/ (ee), on the other hand, the second formant is 20 dB below the first.

TABLE 15.3 Formant frequencies and amplitude of vowels averaged for 76 speakers

Formant frequencies (Hz)		/i/ (ee)	/ɪ/ (i)	/ɛ/ (e)	/æ/ (ae)	/ɑ/ (ah)	/ɔ/ (aw)	/ʊ/ (u̇)	/u/ (oo)	/ʌ/ (u)	/ɜ/ (er)
F_1	M	270	390	530	660	730	570	440	300	640	490
	W	310	430	610	860	850	590	470	370	760	500
	Ch	370	530	690	1010	1030	680	560	430	850	560
F_2	M	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	W	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ch	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
F_3	M	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	W	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ch	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160
		−4	−3	−2	−1	−1	0	−1	−3	−1	−5
Formant amplitudes (dB)		−24	−23	−17	−12	−5	−7	−12	−19	−10	−15
		−28	−27	−24	−22	−28	−34	−34	−43	−27	−20

Source: Peterson and Barney (1952).

15.6 ■ MODELS OF THE VOCAL TRACT

Although the vocal tract, with its many curves and bends, is a rather complex acoustical system, simple models help us to understand the origin of the various formants or resonances.

The simplest acoustic model of the vocal tract is a pipe closed at one end (by the glottis) and open at the other end (lips). Such a pipe has resonances (see Fig. 4.8) given by $f_1 = v/4L$, $f_3 = 3v/4/L, \ldots, f_n = nv/4L$ $(n = 1, 3, 5, \ldots)$. For a pipe 17 cm long (the typical length of a vocal tract), the resonances occur at approximately 500, 1500, and 2500 Hz, which are surprisingly close to the peaks in the spectrum of the vowel sound /ϵ/ (typically at 500, 1800, and 2500 Hz).

Suppose we fasten a small loudspeaker to one end of a 17-cm pipe and place a microphone near the open end, as shown in Fig. 15.10. If the loudspeaker were driven by a pure tone (sine wave) of varying frequency, we would note strong resonances at about 500, 1500, and 2500 Hz, as we just discussed. If the loudspeaker were then driven with a sawtooth waveform (or some other waveform with many harmonics) having a frequency of 100 Hz, and the output of the microphone were displayed on a spectrum analyzer, we would see something similar to the spectrum shown in Fig. 15.10(b). The heights of the various harmonics in the source spectrum have now been shaped by the resonances of the pipe. Formants are present at 500, 1500, and 2500 Hz.

In addition to frequency, the parameters *amplitude* and *bandwidth* are used to describe a formant. The amplitude describes the height of a resonance (see the last three rows in Table 15.3), and the bandwidth describes its breadth.

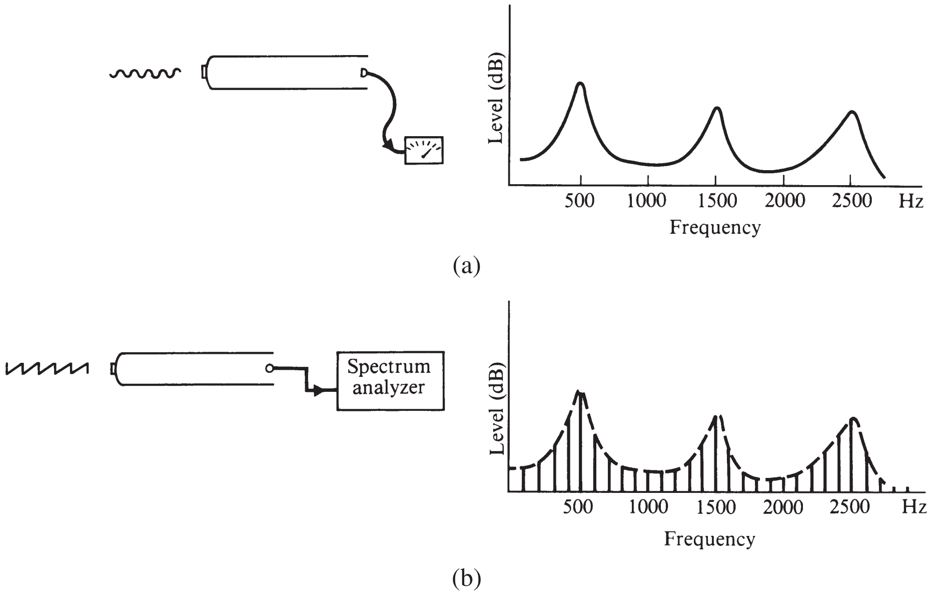


FIGURE 15.10 Response of a closed-pipe model of the vocal tract: (a) resonances of a 17-cm pipe excited with a pure tone of varying frequency; (b) spectrum of a 100-Hz sawtooth wave shaped by the resonances (formants) of the pipe.

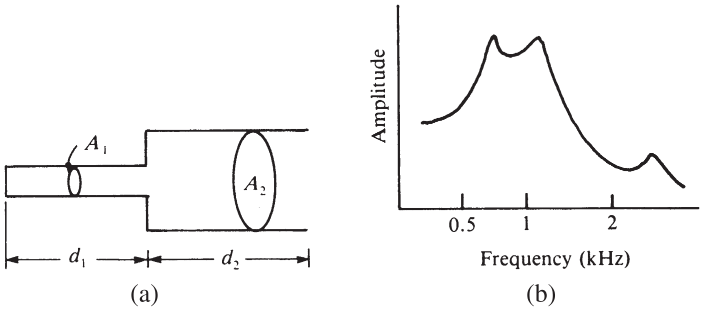


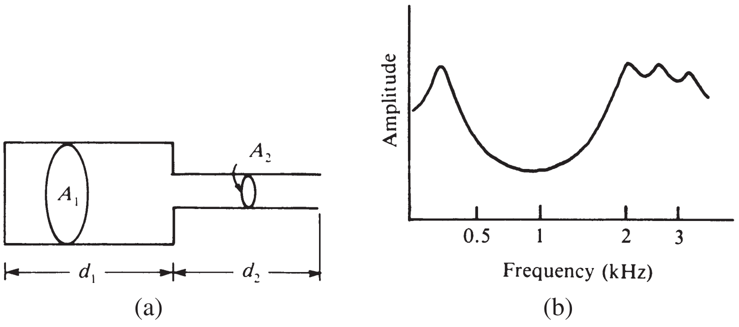
FIGURE 15.11 (a) Two-tube approximation of vocal-tract configuration for the vowel /*a*/ (ah). (b) Approximate form of spectrum envelope of vowel generated by the configuration in (a). (From Stevens 1972.)

Simple models for the vocal tract add to our understanding of how various phonemes might be articulated. Models for the vowel sounds /ɑ/, /i/, and /u/ (“ah,” “ee,” and “oo”) using tubes of two diameters are shown in Figs. 15.11, 15.12, and 15.13 (Stevens 1972).

It is more difficult to construct simple models adequate to describe the articulation of consonant sounds. Stevens (1972) shows, however, that the simple constriction illustrated in Fig. 15.14, if moved to different positions in the vocal tract, can approximate the formants associated with several consonant sounds. The corresponding resonances are shown in Fig. 15.15. For the back portion of the tube, which is essentially closed at both ends, the resonance frequencies are $f_b = nv/2l_b$. For the front portion, the frequencies are $f_f = mv/4l_f$, with $m = 1, 3, 5, \ldots$.

For a given value of the distance l_b from the glottis to the constriction, a vertical line can be drawn and the formant frequencies determined from the intersections with the various curves. Note that for $l_b < 8$ cm, the front cavity (dashed curve in Fig. 15.15) provides the lowest resonance, whereas for $l_b > 10$ cm, the back cavity (solid curve) does so. Near the crossover points, the resonances interact, as represented by the dotted curves. The first formant is not indicated on the graph, but the lowest resonance of the model would be a Helmholtz resonance whose frequency depends on the volume of the back cavity along with the length and cross section of the constriction.

FIGURE 15.12 (a) Two-tube approximation of vocal-tract configuration for the vowel /i/ (ee). (b) Approximate form of spectrum envelope of vowel generated by the configuration in (a). (From Stevens 1972.)



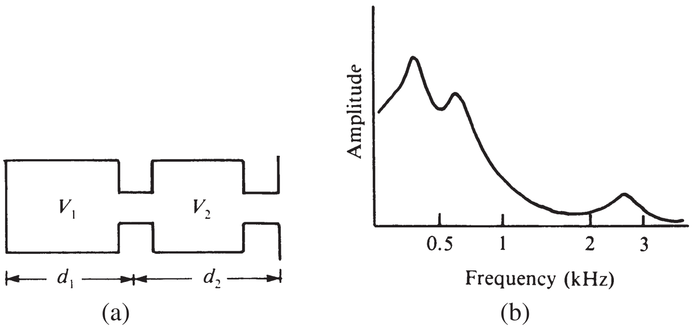


FIGURE 15.13 (a) Approximation of vocal-tract configuration for the vowel /u/ (oo). V_1 and V_2 represent the volumes of the two cavities. (b) Approximate form of spectrum envelope of vowel generated by the configuration in (a). (From Stevens 1972.)

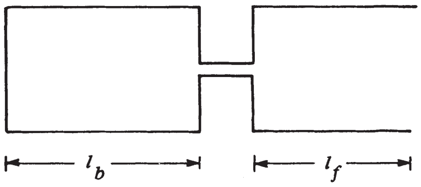


FIGURE 15.14 Idealized model of constricted vocal-tract configuration corresponding to a consonant. The constriction is adjusted to different positions to represent different places of articulation. (From Stevens 1972.)

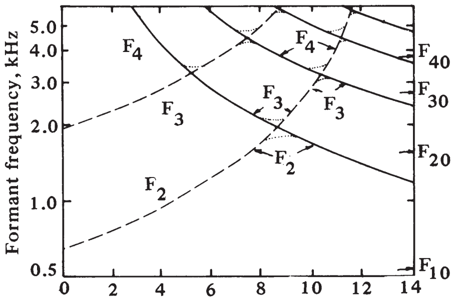


FIGURE 15.15 Relations between natural frequencies and the position of the constriction for the configuration shown in Fig. 15.14. The overall length of the tube is 16 cm and the length of the constriction is 3 cm. The dashed lines represent the lowest two resonances of the front cavity (anterior to the constriction); the solid lines represent the lowest four resonances of the back cavity. The dotted lines near the points of coincidence of two resonances represent the resonant frequencies for the case in which there is a small amount of coupling between front and back cavities. The resonances of a 16-cm tube with no constriction are shown by the arrows at the right. The curves are labeled with the appropriate formant numbers. (From Stevens 1972.)

Constriction positions corresponding to $l_b < 8$ cm represent configurations for uvular and pharyngeal consonants that do not ordinarily occur in English. Rather abruptly moving the constriction from left to right through the F_2–F_3 crossover point ($l_b \approx 8.5$ cm) and at the same time enlarging the constriction toward the vowel configuration resembles the articulation of the velar consonants /g/ or /k/. Similar abrupt shifts near the F_3–F_4 crossover point correspond to the fricatives /s/ and /ʃ/ ("sh"). The labial and labiodental consonants would have $l_b > 13$ cm.

15.7 ■ STUDIES OF THE VOCAL TRACT

A number of speech scientists have made X-ray photographs of the vocal tract during the production of speech sounds. From these photographs, profiles of the vocal tract can be constructed. It is interesting to compare the profiles for the vowel sounds, ɑ, i, and u ("ah," "ee," "oo"), shown in Fig. 15.16, with the simple models shown in Figs. 15.11–15.13.

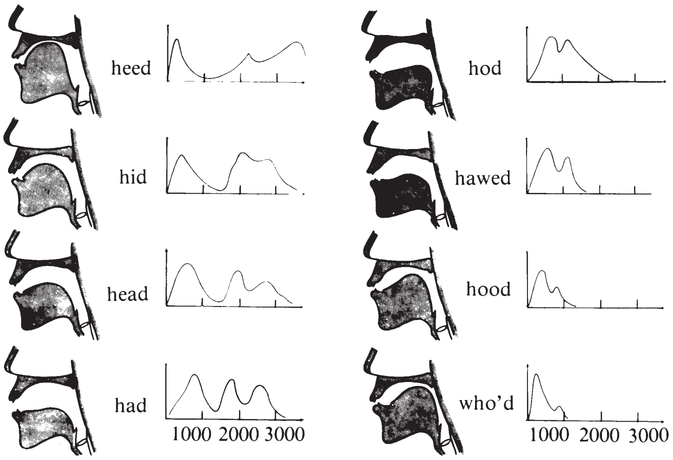


FIGURE 15.16 The positions of the vocal organs (based on data from X-ray photographs of the author) and the spectra of the vowel sounds in the middle of the words, *heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who'd*. Compare the sounds *hod*, *heed*, and *who'd* with the corresponding two-tube models of the vocal tract in Figs. 15.11, 15.12, and 15.13 (Ladefoged 1962).

In the case of consonants, the profile of the vocal tract depends to some extent on the vowel sounds that precede and follow the consonant. The location of the constriction (the place of articulation) changes very little, however. Profiles for the stop or plosive consonants are shown in Fig. 15.17. Note that the voiced consonants are nearly identical to the corresponding unvoiced consonants. The only difference between the words "to" and "do," for example, is that the vocal folds vibrate during the /d/ sound, but do not begin vibrating until after the /t/ sound has been articulated.

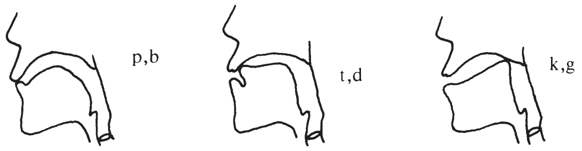


FIGURE 15.17 Profiles of the vocal tract showing place of articulation of the stop or plosive consonants.

15.8 ■ PROSODIC FEATURES OF SPEECH

Prosodic features are characteristics of speech that convey meaning, emphasis, and emotion without actually changing the phonemes. They include pitch, rhythm, and accent. In English, prosodic features play a secondary role to that of phonemes in the communication of information.

In certain languages, such as Chinese, however, a phoneme can take on several different meanings depending on its "tone." The manner in which the frequency changes with time for the four tones of Mandarin Chinese is shown in Fig. 15.18.

One of the common uses of prosodic features is to change a declarative sentence ("You are going home.") into a question ("You are going home?"). This is done mainly by raising the pitch of the final word. The same sentence could be made imperative by adding stress (increase in both loudness and pitch) to the second word ("You *are* going home!").

Prosodic features tend to indicate the emotional state of the speaker. "Raising one's voice" in anger, for example, increases both loudness and pitch. A state of excitement frequently causes an increase in the rate of speaking. Several attempts have been made to accomplish acoustic "lie detection" by analyzing the prosodic features of recorded speech for evidence of stress.

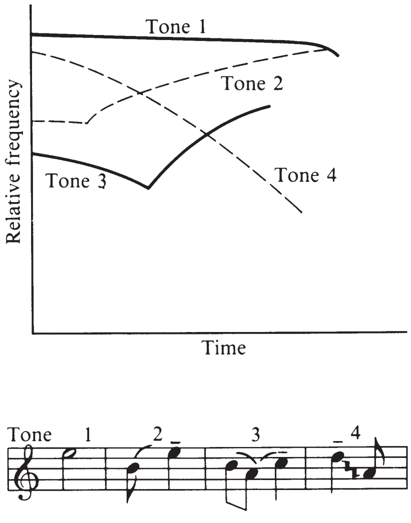


FIGURE 15.18 Frequency changes with time for tones in Mandarin Chinese. (After Luchsinger and Arnold 1965.)

15.9 ■ SUMMARY

The principal parts of the vocal tract are the larynx and vocal cords, pharynx, nasal cavity, oral cavity, tongue, lips, and teeth. Speech sounds originate with the vibrations of the vocal folds or with a constriction of the air flow, are filtered in the vocal tract, and finally are radiated through the lips or nose. Resonances of the vocal tract, called *formants*, determine the vowel sounds, the first and second formant being the most important. Consonants involve rapid changes in sound generated by changing a constriction somewhere in the vocal tract. Consonants can be classified according to their place and manner of articulation.

Simple models of the vocal tract, constructed from tubes of different lengths and diameters, help us understand the acoustical behavior of the vocal tract. Profiles of the vocal tract can be drawn from X-ray photographs. Prosodic features, such as pitch, rhythm, and accent, convey meaning, emphasis, and emotion.

REFERENCES AND SUGGESTED READINGS

David, E. E., Jr., and P. B. Denes, eds. (1972). *Human Communication: A Unified View*. New York: McGraw-Hill.

Denes, P. B., and E. N. Pinson (1973). *The Speech Chain*. New York: Anchor-Doubleday.

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

Flanagan, J. L. (1965). *Speech Analysis, Synthesis and Perception*. New York: Academic.

Ladefoged, P. (1962). *Elements of Acoustic Phonetics*. Chicago: University of Chicago Press.

Lehiste, I., ed. (1967). *Readings in Acoustic Phonetics*. Cambridge, Mass.: MIT Press.

Luchsinger, R., and G. E. Arnold (1965). *Voice-Speech-Language*, Belmont, Calif.: Wadsworth.

Peterson, G. E., and H. L. Barney (1952). "Control Methods Used in a Study of Vowels," *J. Acoust. Soc. Am.* **24**: 175.

Pickett, J. M. (1999). *The Acoustics of Speech Communication*. Needham Heights, Mass.: Allyn and Bacon.

Schouten, J. F. (1962). "On the Perception of Sound and Speech," Congress Report, 4th ICA, Copenhagen, 196.

Stevens, K. N. (1972). "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in *Human Communication: A Unified View*, p. 51. Eds., E. E. David, Jr., and P. B. Denes. New York: McGraw-Hill.

Stevens, K. N., and A. S. House (1971). "An Acoustical Theory of Vowel Production and Some of Its Implications," *J. Speech & Hearing Research* **4**(4): 75.
Strong, W. J., and G. R. Plitnik (1983). *Music, Speech and High Fidelity*, 2nd ed. Provo, Utah: Soundprint. Chaps. 23–27.

Sundberg, J. (1977). "The Acoustics of the Singing Voice," *Sci. Am.* **236**(3): 82.

Thomas, I. B. (1969). "Perceived Pitch of Whispered Vowels, " *J. Acoust. Soc. Am.* **46**: 468.

GLOSSARY

cardinal vowels Eight vowel sounds that serve as a standard of comparison for the vowels of various languages.

diphthong A combination of two or more vowels into one phoneme.

epiglottis A thin piece of cartilage that protects the glottis during swallowing.

filter A device that allows signals in a certain frequency range to pass and attenuates others.

formants Vocal tract resonances that determine speech sounds.

fricatives Consonants that are formed by constricting air flow in the vocal tract (such as f, v, s, z, th, sh, etc.).

glottis The V-shaped opening between the vocal folds.

larynx The section of the vocal system, composed mainly of cartilage, that contains the vocal folds.

nasals Consonants that make use of resonance of the nasal cavity (m, n, ng).

palate The roof of the mouth.

pharynx Lower part of the vocal tract which connects the mouth to the trachea.

phonemes Individual units of sound that make up speech.

phonetics The study of speech sounds.

plosives Consonants that are produced by suddenly removing a constriction in the vocal tract (p, b, t, d, k, g).

prosodic feature A characteristic of speech, such as pitch, rhythm, and accent, that is used to convey meaning, emphasis, and emotion.

semivowels Consonants for which the vocal tract is formed in a configuration generally used for vowels (w, y).

vocal folds or vocal cords Folds of ligament extending across the larynx that interrupt the flow of air to produce sound.

REVIEW QUESTIONS

1. What are the three main cavities of the vocal tract? Which of these play a role in the production of speech?
2. Describe the way in which the vocal folds vibrate.
3. Describe the role of the vocal folds in producing an "h" sound.
4. The spectrum envelope of speech sound can be thought of as the product of what three components?
5. Describe the spectrum of the glottal source function.
6. Give examples of the following types of consonants: fricative, nasal, liquid, semivowel.
7. Give examples of voiced and unvoiced consonants.
8. Sketch simple two-tube models of the vocal tract configuration for the vowels /ɑ/ and /ɪ/.
9. What voiced and unvoiced consonants are formed with the lips?
10. What voiced and unvoiced consonants are formed with the soft palate?
11. Describe the role of the vocal tract in whistling.
12. What are prosodic features of speech? Give two examples.
13. What are typical vocal fold vibration frequencies in male and female speakers?
14. How is it possible to observe the motion of the vocal folds?
15. How is it possible to observe the shape of the vocal tract for different vowel sounds?
16. How does the glottal waveform change when one speaks louder?

QUESTIONS FOR THOUGHT AND DISCUSSION

1. Discuss the function of each of the principal parts of the vocal tract.
2. If a person partially fills his or her lungs with helium and then speaks, the speech sounds distorted (it is sometimes described as sounding like Donald Duck). Explain this on the basis of formants (the information in Table 3.1 may be helpful). (This type of distortion will be discussed in Chapter 16.)
3. Discuss the acoustics of
   (a) a "hoarse" throat;
   (b) a stuffed nose;
   (c) swollen tonsils.
4. Although the vibrations of the vocal folds are similar to the vibrations of a trumpeter's lips, the control of frequency by the air column through feedback is all but missing in the case of the vocal folds. Can you explain why? (Consider the mass of the vibrating members, the sharpness of the air resonances, and damping in each case.)

EXERCISES

1. Calculate the first three resonances of a tube 11 cm long (the approximate length of a child's vocal tract) open at one end and closed at the other. Compare these to the formant frequencies for /ϵ/ given in Table 15.3.
2. Make a graph of the second formant frequency (vertical axis) versus the first formant frequency (horizontal axis) for the ten vowel sounds given in Table 15.3. Do this for either the average male or female voice. Select a scale for each axis that is appropriate for the data you intend to graph.
3. Take a simple sentence (e.g., "You always give the right answers") and attempt to give it several meanings by changing prosodic features. For each different way of speaking the sentence, indicate the pattern of pitch and loudness used.
4. Express in newtons/meter^2 the maximum and minimum lung pressures used in speech (4 cm and 20 cm of water). Atmospheric pressure (10^5 N/m^2) corresponds to a manometer pressure of about 34 ft of water.
5. Calculate the frequencies of resonance for a tube 16 cm long closed at one end and open at the other, and show that they correspond to F_{10}, F_{20}, F_{30}, in Fig. 15.15.
6. Determine whether there is a "scaling factor" relating male and female vowel formants by the following calculations. Determine the ratios of the female-to-male for-

mant frequencies for the vowels given in Table 15.3. Find the average ratio. Could recording male speech and playing it back 16% faster make it resemble female speech?

7. Suppose a vocal tract 17 cm long were filled with helium ($v = 970$ m/s). What formant frequencies would occur in a neutral tract?

8. Estimate relative male and female vocal tract lengths by:

(a) Averaging the ratios of a male and female formant frequencies for several vowels;

(b) Assuming they have about the same ratio as male and female heights.

9. The resonance frequency of a Helmholz resonator (Section 2.3) is

$$f = \frac{v}{2\pi}\sqrt{\frac{A}{lV}},$$

where v is the velocity of sound, A and l are the cross-sectional area and length of the neck, and V is the volume of the resonator. For the model in Fig. 15.14, assume a neck area of 0.6 cm^2, a length of 3 cm, a volume V of 20 cm^3, and a sound velocity of 344 m/s, and calculate the resonance frequency. Compare this to F_{10}, calculated in Problem 5.

EXPERIMENTS FOR HOME, LABORATORY, AND CLASSROOM DEMONSTRATION

Home and Classroom Demonstration

1. *Formant tube* The experiment illustrated in Fig. 15.10 can be expanded to simulate other vocal tract resonances, as shown in Fig. 15.19. A pure tone is varied over the frequency range of 200–4000 Hz, and the resonances of the tube are observed. Listening to a sawtooth waveform with a fundamental frequency of 100–150 Hz should suggest vowellike sounds as the constriction is moved (b) or tubes of different diameters and lengths are joined together (c).

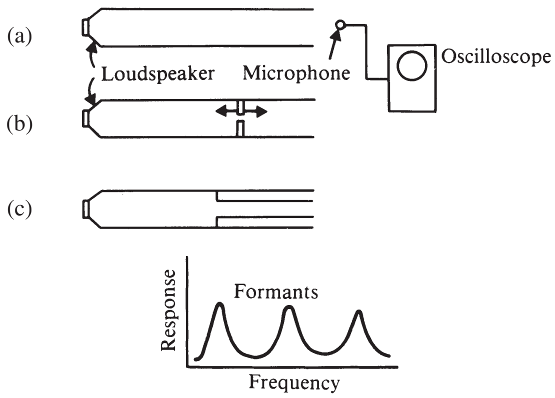


FIGURE 15.19 Formant tube, which can be used to demonstrate resonances of the vocal tract.

2. *Movable constriction* The effect of constricting the air flow at different places in the vocal tract can be simulated by inserting a small nozzle into a short length of pipe, as shown in Fig. 15.20. As the constriction moves up and down the pipe, the sound changes in character, growing louder when the source reaches the position of a pressure maximum for one of the pipe resonances.

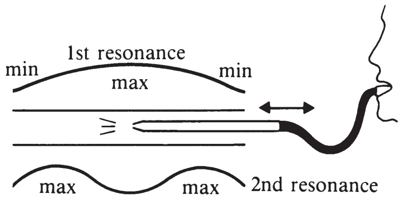
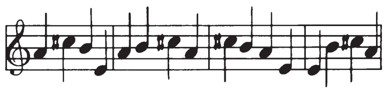


FIGURE 15.20 Demonstration to show the effect of moving a constriction up and down the vocal tract.

3. *Whispered vowels* During a whisper, the vocal folds produce broadband ("white") noise, which contains a wide range of frequencies and virtually no sensation of pitch. Whispering vowel sounds, however, shapes the vocal tract so that bands of noise near the formant frequencies are emphasized. A rather faint sense of pitch develops, which usually corresponds to the second formant frequency (Thomas 1969).

Professor J. F. Schouten of the Netherlands was well known for his demonstrations of acoustic phenomena as well as his research in hearing in speech. He demonstrated the whispered-vowel phenomenon by whispering the following four lines of vowels to produce the well-known Westminster chime:



$\phi - \mathrm{I} - \epsilon - a$
$\phi - \epsilon - \mathrm{I} - \phi$
$\mathrm{I} - \phi - \epsilon - a$
$a — \epsilon — \mathrm{I} — \phi$

The second formant of ϕ (a vowel sound common in Scandinavian and Germanic languages) is around 1760 Hz, which is two octaves above the musical standard $A_4 = 440$ Hz (Schouten 1962).

4. *Single formants* The character of vowel sounds is mainly determined by the first and second formants. The waveform and spectrum /i/ are shown in Fig. 15.21(a). Filtering out one of the formants changes the vowel sound appreciably. The first formant, produced by itself, sounds much like /u/. The second formant by itself has a sharp timbre that produces no phonetic association, since the human voice is unable to produce it singly (Schouten 1962).

5. *Whistling* When one whistles, the vocal tract is excited from the end opposite the vocal folds. The normal whistling range is from about F_5 to F_7 (700 to 2800 Hz), although with practice many people can whistle down to 500 Hz and below. This suggests that the vocal tract acts somewhat like a closed cylindrical pipe (see Fig. 15.10). Whistled sound is very nearly a pure tone, with exceptionally weak overtones. Incidentally, the lowest whistled note for most people is near the pitch of their highest sung note (in falsetto for a male voice); thus, the human vocal system can emit sounds over a total range of about five octaves.

6. *Air flow in various phonemes* Hold a hand directly in front of your mouth and say "what"; note the air flow. Say "ah-h-h," and compare the air flow. Compare pairs of words beginning with voiced and unvoiced plosive consonants (e.g., to/do, pet/bet, kit/bit).

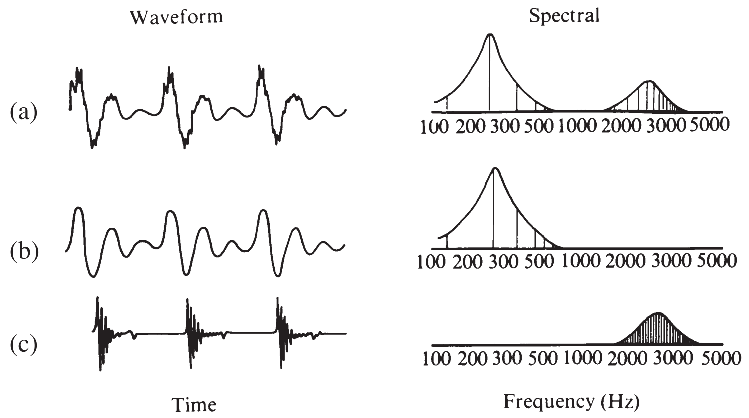


FIGURE 15.21 (a) Waveform of vowel /i/ and its spectral pattern showing two formants. (b) First formant only gives an /u/-like sound. (c) Second formant only. (After Schouten 1962.)

Laboratory Experiment

Vowel formants (Experiment 22 in *Acoustics Laboratory Experiments*)

CHAPTER 16 Speech Recognition, Analysis, and Synthesis

Speech is not just the future of Windows but the future of computing itself.
Bill Gates (quoted by *Business Week*, Feb. 23, 1998)

Our ability to recognize the sounds of language is truly phenomenal. Speech can be followed at rates as high as 400 words per minute. If we assume an average of five phonemes or individual sounds per word, this means recognizing over 30 phonemes per second; even normal conversation requires the recognition of 10 to 15 phonemes per second. In this chapter, we will consider the way in which speech recognition takes place, particularly through certain types of *cues* in the complex speech sounds we hear. Before we consider speech recognition, however, it is appropriate to discuss the acoustical analysis of speech sounds.

In this chapter you should learn:

- About speech spectrograms;
- About recognition of vowels and consonants;
- About filtered and compressed speech;
- About speech recognition and synthesis by computers;
- About speaker identification and voiceprints.

16.1 ■ THE ANALYSIS OF SPEECH

Some speech sounds change rapidly and, therefore, require special techniques for analysis. Graphs of sound level versus time and graphs of sound level versus frequency (sound spectra) are useful but inadequate. It is more useful to display the sound level as a function of both frequency and time. Various techniques have been used for creating such displays.

One way to display three variables is on a three-dimensional graph. When comparing sound level, frequency, and time, this can be approximated by making multiple graphs of sound level versus frequency, each one displaced slightly in time to create perspective. Such a three-dimensional display is illustrated in Fig. 16.1.

Sound spectra have appeared throughout this book. In Section 2.7 we introduced spectra as recipes for describing a complex vibration or sound. In Section 7.10 we discussed how spectrum analysis (or Fourier analysis) is done by spectrum analyzers. An instrument that rapidly analyzes the spectrum of sound is known as a *real-time spectrum analyzer*. Such

From Chapter 16 of *The Science of Sound*, Third Edition. Thomas D. Rossing, Richard F. Moore, Paul A. Wheeler.

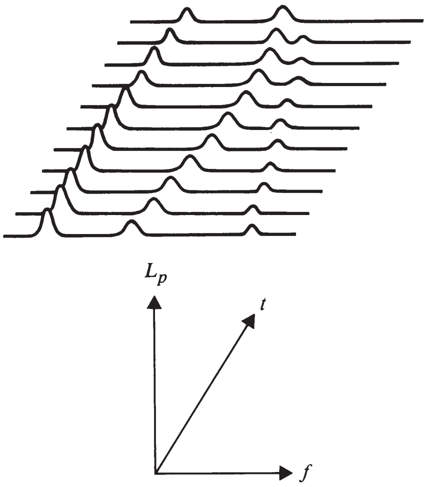


FIGURE 16.1 Three-dimensional display of sound level versus frequency and time.

instruments nowadays are nearly always digital; they generally use a technique called a fast-Fourier transform (FFT) and are referred to as *FFT analyzers*. FFT analyzers are useful in producing three-dimensional ("waterfall") displays such as that shown in Fig. 16.1.

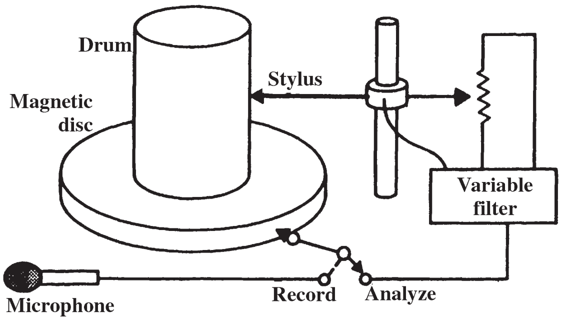
An instrument that is particularly useful for speech analysis is the *sound spectrograph*, originally developed at the Bell Laboratories around 1945. This instrument records a sound-level–frequency–time plot for a brief sample of speech on which the third dimension, sound level, is represented by the degree of blackness in a two-dimensional time-frequency graph.

A modern digital version of the sound spectrograph is shown in Fig. 16.2(b). Filters divide the incoming speech signal into many different frequency bands (from about 50 to 250, depending on the type of analysis being done). The amount of sound power that comes through each filter is measured as a function of time, and the speech spectrograph is printed on the grayscale printer shown at the right. The format printed by the digital sound spectrograph in Fig. 16.2(b) is similar to that of the older analog version (Fig. 16.2(a)), which speech scientists have found so useful over the years.

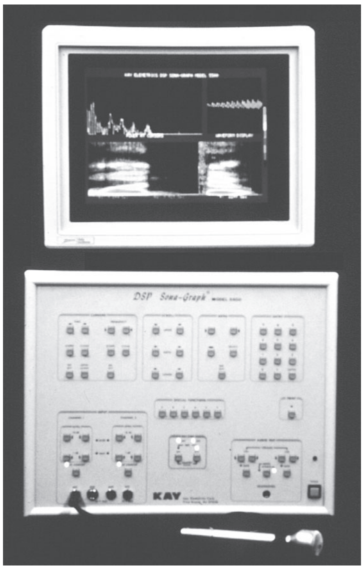
A speech spectrogram is shown in Fig. 16.3. The horizontal axis is time, and the vertical axis frequency. The vertical striations show the fundamental period of the vocal cord vibrations. Two filter bandwidths are customarily used with the instrument, 45 and 300 Hz. The broader band gives better time resolution at the expense of frequency resolution. In the analog version speech sample must be played many times to record a spectrogram, so playback is at a much higher speed than that used for recording on the magnetic disc.

16.2 ■ THE RECOGNITION OF VOWELS

Although spectrograms of vowel sounds may show four or five formants, the first two or three formants are generally sufficient to identify vowel sounds. On the other hand, experiments have shown that, under some conditions, vowels can be recognized from only the



(a)



(b)

FIGURE 16.2 (a) A schematic diagram of a sound spectrograph; (b) a digital sound spectrograph.

higher formants when the lowest two formants are missing. Thus, in normal speech, there are multiple acoustic cues to aid in the recognition of vowel sounds. Some of these extra cues make it possible to determine vowel sounds even when distortion and interference are present, as the following examples illustrate.

One familiar type of distortion occurs when speech is recorded at one speed and played back at a faster speed, producing what has been called "duck talk." Even when the pitch and all formants are raised by an octave or more (by doubling the speed of a tape recorder, for example), it is possible to understand most of what is being said. Apparently our speech-

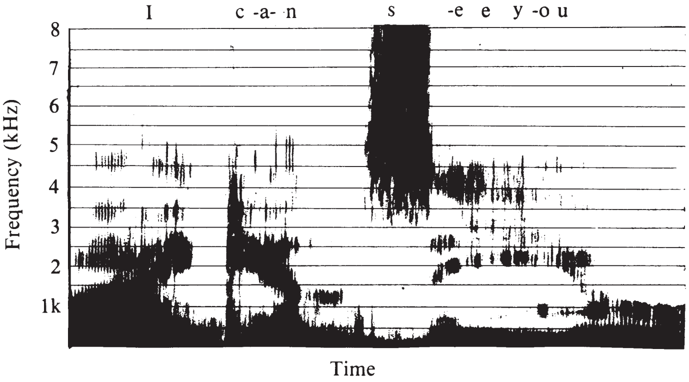


FIGURE 16.3 A speech spectrogram. The vertical axis is frequency, and the horizontal axis is time. Sound level is indicated by darkness.

processing system is able to “scale” the entire structure of speech. That is, when we hear speech at a higher pitch, we also look for formants in a higher frequency range. How this is accomplished is not very well understood at the present time.

Another example of distortion caused by formant transposition is *helium speech*. The velocity of sound in pure helium is nearly three times greater than in air (see Table 3.1). If one takes a deep breath of helium, the resonances of the vocal tract (formants) will increase in frequency by some factor, which is typically 1.5 rather than 3, because our exhalation will contain a mixture of helium with nitrogen, carbon dioxide, etc. Speech produced under these conditions sounds quite similar to duck talk, although the nature of the distortion is quite different. Analysis of helium speech indicates that the fundamental pitch is virtually unchanged, because the mixture of gas in the vocal tract has little effect on the vibration frequency of the vocal cords. To understand helium speech, then, we must recognize the raised formants even though the pitch of the vowel sounds corresponds to the normal formants we are accustomed to hearing.

In order to prevent nitrogen narcosis, deep-sea divers breathe mixtures of helium, nitrogen, and oxygen at high pressure, and speech becomes unnatural or even unintelligible. (In Sealab II, for example, the inside pressure was maintained at 6.8 times atmospheric and the gas mixture at 80% helium, 15% nitrogen, and 5% oxygen.) Thus the problem of helium speech has attracted considerable attention, and several experimental speech processors (*formant restorers*) have been developed (Stover 1966).

The vocal tracts of young children are considerably smaller than those of adults; hence, the formant structure is considerably different. The pharynx tends to be proportionately shorter than the oral cavity, so the formant configuration is not scaled to that of an adult. Yet our speech decoder enables us to recognize a vowel spoken by a young child as being the same vowel spoken by an adult. Some of the advantages of multiple speech cues become more apparent in light of these considerations.

16.3 ■ THE RECOGNITION OF CONSONANTS

Unlike vowel sounds, which change slowly, consonant sounds change very rapidly. As we described in Section 15.4, consonants are articulated by constricting or blocking the flow of air somewhere in the vocal tract. The sound cues by which the consonant is recognized often occur in the first few milliseconds after the block is released and air is allowed to flow through the vocal tract.

Figure 16.4 shows a sound spectrogram of a simple phrase, “this is a sound spectrogram.” The vowel formants appear as dark horizontal bars, whereas some of the up-and-down movements of these formants signal the consonants. The “s” sound is a burst of noise extending up to 8000 Hz. The fine vertical lines represent the vibrations of the vocal folds.

Experiments with the sound spectrograph have contributed a great deal to our understanding of the recognition of speech sounds, especially of consonants. In some experi-

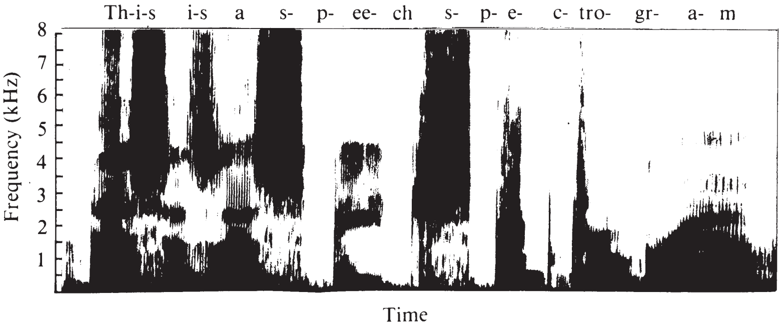


FIGURE 16.4 A sound spectrogram of a spoken phrase. The vertical axis is frequency, and the horizontal axis is time.

ments, certain features of speech are altered or eliminated to determine the intelligibility changes. Another type of experiment, however, generates speechlike sounds artificially. In such artificially synthesized speech, it is relatively easy to adjust separately the acoustic features to determine their effect on speech recognition.

Much early knowledge about the recognition of consonants was obtained with the pattern-playback machine built some years ago at the Haskins Laboratories. This machine works like a speech spectrograph in reverse: when a pattern similar to a spectrogram is fed in, it generates a sound with the designated intensity-frequency-time pattern. Arbitrary patterns may be painted on plastic belts in order to study the effects of varying the features of speech one by one.

A dot presented to the pattern-playback machine produces a "pop" that is like a plosive consonant, but difficult to recognize as any particular consonant unless it is followed by a vowel sound. In experiments by Cooper et al. (1952), listeners were presented with 15-ms noise bursts of varying frequencies, followed by two-formant vowel sounds, as shown in Fig. 16.5. High-frequency bursts were heard as /t/ for all vowels, but bursts at lower frequencies could be heard as either /p/ or /k/, depending on the vowel sound that followed. Bursts were heard as /k/ when they were on a level with or slightly above the second formant of the vowel; otherwise, they were heard as /p/.

Another way to generate plosive consonants on the pattern-playback machine is by a frequency transition in the second formant, which may be upward or downward. Transitions of the second formant of the type shown in Fig. 16.6 will produce the unvoiced plosives /t/, /p/, or /k/, depending on the vowel formants that follow. A remarkable result emerged

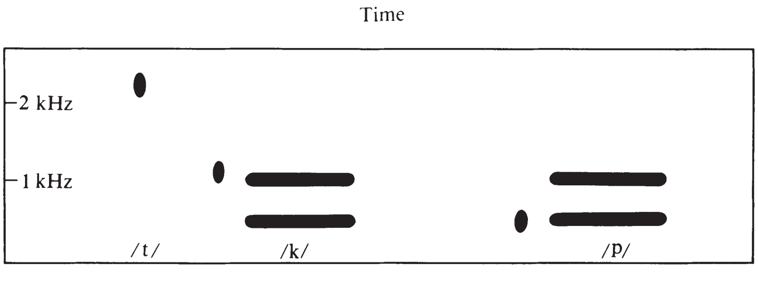


FIGURE 16.5 Stimulus patterns for producing /t/, /k/, and /p/ sounds on the pattern playback machine. A single burst of high-frequency noise is heard as /t/. A noise burst at the frequency of the second formant of a following vowel is heard as /k/; a noise burst below the second formant is heard as /p/. (From data in Cooper et al. 1952.)

from experiments with the pattern-playback machine: All the second-formant transitions perceived as one particular plosive pointed back toward one particular frequency. The transitions in Fig. 16.7, which appear to originate from about 1800 Hz, are all heard as the sound /t/. Similarly, transitions that produce /p/ appear to originate from about 700 Hz, and /k/-producing transitions originate from about 3000 Hz.

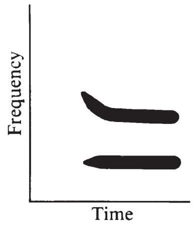


FIGURE 16.6 A formant transition, which may produce a /t/, /p/, or /k/ depending on the vowel that follows.

The voiced plosives /b/, /d/, and /g/ have associated with their second-formant transition an upward transition in the first formant as well. The first formant is raised from a very low frequency to a level appropriate for the vowel. Figure 16.8 shows patterns that synthesize /b/, /d/, and /g/ sounds before various vowels. Note that although the first formant always moves upward, the second formant can move either upward or downward, depending on the vowel.

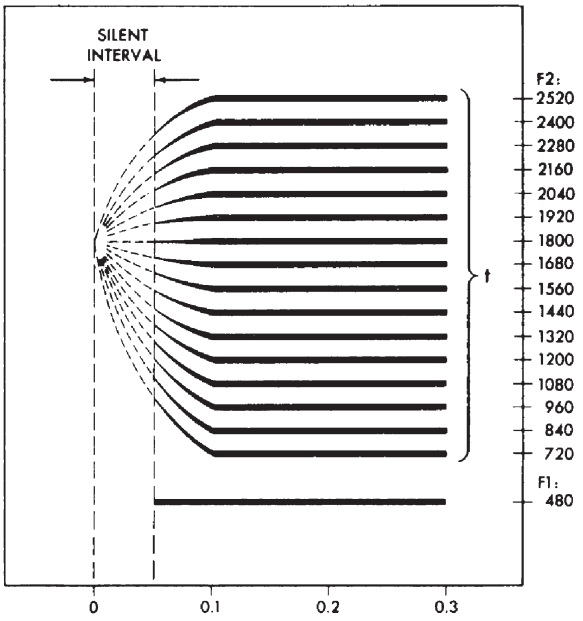


FIGURE 16.7 Second-formant transitions perceived as the same plosive consonant "t." (After Delattre, Liberman, and Cooper 1955.)

/d/ Sounds

In the case of the consonant /d/, the second formants appear to originate from a "d-locus" at about 1800 Hz; the key to distinguishing the voiced /d/ from the unvoiced /t/, therefore, lies in the cue provided by the first-formant transition. It is interesting to note, however, that patterns extending all the way back to the d-focus, as in Fig. 16.9(a), do not always produce a clear /d/. In order to have a /d/ sound in every case, it is necessary to erase the first part of the transition so that it "points" at the locus but does not actually begin there, as in Fig. 16.9(b). Presumably this is the way we are accustomed to receiving these cues, and major change confuses our speech decoder.

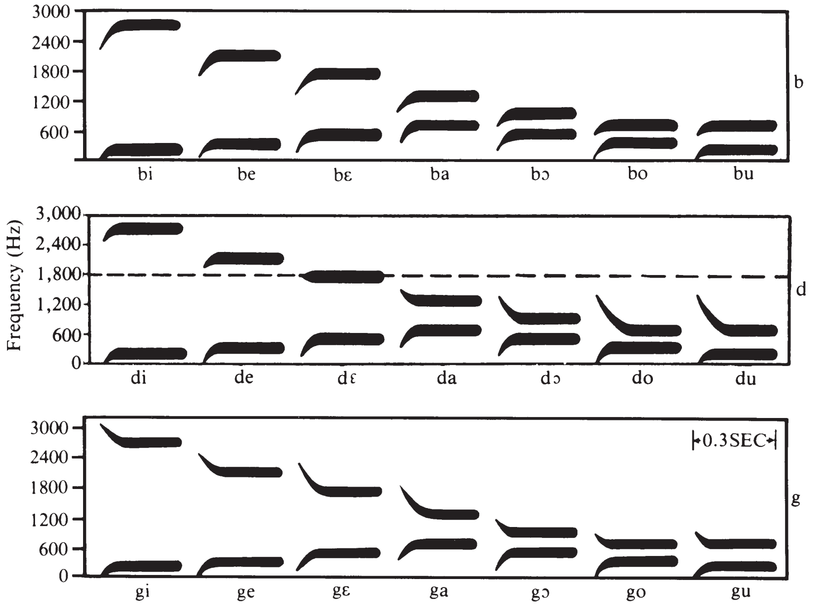


FIGURE 16.8 Spectrographic patterns sufficient for the synthesis of /b/, /d/, and /g/ before vowels. The dashed line at 1800 Hz shows the locus for /d/. (From Delattre, Liberman, and Cooper 1955.)

For the liquids and semivowels /r/, /ℓ/, /w/, and /j/, the second-formant transition begins at the locus, although the exact character of the transition can vary with context.

Fricative consonants can be distinguished from all other sounds by the hissing noise of the turbulent air stream, which appears as a fuzzy area on speech spectra. We may ask, however, "What are the cues for distinguishing one fricative from another?" Experiments on both natural and synthesized speech have indicated that /s/ and /ʃ/ ("sh") are distinguished

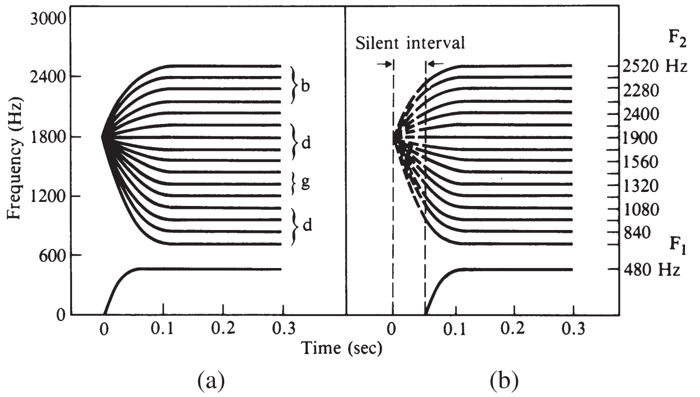


FIGURE 16.9 (a) Second-formant transitions that start at the /d/ locus. (b) Comparable transitions that merely "point" at it, as indicated by the dotted lines. Those of (a) produce syllables beginning with /b/, /d/, or /g/, depending on the frequency level of the formant; those of (b) produce only syllables beginning with /d/. (From Delattre, Liberman, and Cooper 1955.)

from other fricatives by their greater intensities and from each other by their spectra. In the case of /ʃ/, sound energy is concentrated in the 2000- to 3000-Hz range, whereas for /s/, it is above 4000 Hz. The weaker fricatives /f/ and /θ/ ("th") depend on second-formant transitions in the vowel sound that follows to provide clues about the place of articulation.

The *duration* of a sound may provide an important clue for phoneme recognition. In Chapter 15 we pointed out that a fricative appeared to change into a plosive when its duration was shortened. This can be demonstrated by tape recording a normally pronounced word such as see; if the tape is cut to reduce the duration of the initial /s/ from its normal 0.1-s duration to about 0.01 s, the word is heard as "tee."

The effects of third-formant transitions on the perception of consonants are more complicated and less understood than those of second- and first-formant transitions. For example, a third-formant transition provides a clue for the perception of /d/ in "di" but not in "du" (Liberman et al. 1967). Apparently, no simple explanation has been made of this phenomenon, although it may be noted that the second-formant transitions are in opposite directions in the two cases.

16.4 ■ FILTERED SPEECH AND NOISY ENVIRONMENTS

Filters are devices that respond selectively to certain frequencies. The vocal tract acts like a series of filters, each tuned to one of the resonances that we associate with formants; however, electrical filters can be constructed to have a much sharper frequency response than the vocal tract does. Some experiments with electrically filtered speech will be described in this section.

Electrical filters may have high-pass, low-pass, band-pass, or band-reject characteristics (see Section 18.6). A high-pass filter transmits only those frequencies above its cutoff frequency, and a low-pass filter only those frequencies below its cutoff frequency. A band-pass filter has both high and low cutoff frequencies and transmits only frequencies that lie in the band between; a band-reject, or notch, filter rejects only signals between the two cutoff frequencies.

Speech intelligibility is usually measured by *articulation tests* in which a set of words is spoken and a listener or group of listeners is asked to identify them. Articulation tests customarily use lists of specially selected words of one or two syllables. The articulation score, which is the percentage of words correctly identified, will be lower for these isolated test words, of course, than for words used in the normal context of speech.

Articulation scores for filtered speech are shown in Fig. 16.10 for both high-pass and low-pass filtering. The curves are seen to cross at 1800 Hz, where the articulation score for both is about 67%. Normal conversation, therefore, would be completely intelligible by listening only to components above 1800 Hz, or, equally so, by listening only to components below 1800 Hz. It is also possible to achieve an acceptable level of intelligibility for speech after passage through a band-pass filter with a surprisingly narrow passband. The minimum acceptable passband is found to vary with frequency; in the range around 1500 Hz, for example, a 1000-Hz band width is sufficient to give a sentence articulation score of about 90% (Denes and Pinson 1973). Using a narrowband filter ($\frac{1}{3}$-octave bandwidth), intelligibility reached 50% around 2000 Hz but was very low at most frequencies (Chari, Herman, and Danhauer 1977). Needless to say, speech quality deteriorates more

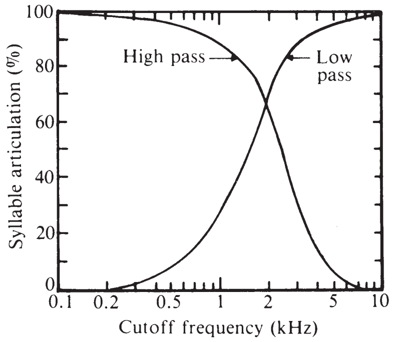


FIGURE 16.10 Intelligibility of filtered speech for different cutoff frequencies of both high-pass and low-pass filters. Note that the two curves cross at about 1800 Hz, where the articulation score is 67% for both types of filter. (After French and Steinberg 1947.)

than does intelligibility with filtering. Filtered speech sounds thin and unpleasant even when it is understandable.

The effects of waveform distortion have also been investigated. *Peak clipping* is a type of distortion that often results from overdriving an audio amplifier, but it is sometimes introduced deliberately into speech communication systems in order to reduce the bandwidth required to carry the speech. Figure 16.11 illustrates moderate and severe peak clipping of a speech waveform. Intelligibility of speech is impaired surprisingly little by peak clipping, although the quality of the speech suffers. Even after severe peak clipping, similar to that shown in Fig. 16.11(c), intelligibility remains at 50 to 90%, depending on the skill of the listener (Licklider and Pollack 1948).

The intelligibility of speech in noisy environments is a timely subject that has been studied at a number of laboratories. The degree of "masking" of speech depends on the intensity and the spectrum of the interfering noise. Using broadband noise or white noise (noise with equal intensity at every audible frequency), the intelligibility of words drops to about 50% when the average intensities of the speech and the noise are about equal. The intelligibility of sentences remains higher, however, because of linguistic and semantic cues. Figure 16.12 indicates how the thresholds of intelligibility and detectability of speech depend on the level of broadband noise.

A tone of lower frequency can mask a tone of higher frequency much more effectively than the converse (see Section 6.10). Thus narrowband noise is most effective in masking speech if its frequency is below the speech band. Potential for speech interference in a

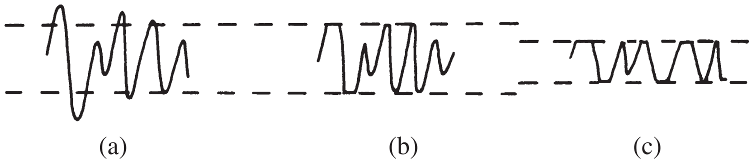


FIGURE 16.11 Peak clipping: (a) waveform of original speech; (b) waveform after peak clipping; (c) waveform after severe peak clipping.

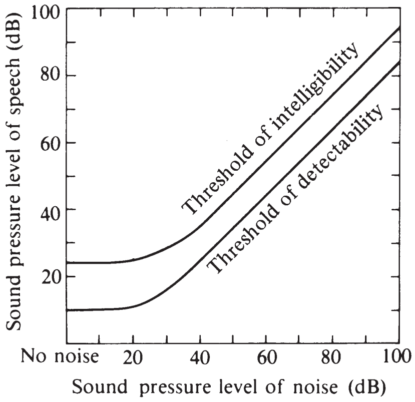


FIGURE 16.12 The thresholds of intelligibility and detectability as functions of the intensity of the masking noise. (From Hawkins and Stevens 1950.)

noisy environment is sometimes expressed by the speech-interference level, which is the average of the noise level in three appropriate frequency bands (see Section 31.5).

An interesting demonstration illustrating one property of speech interference may be performed using *elliptical speech*. As an interfering noise source rises in intensity, one of the first features of speech that is lost is the place of articulation, so "cat" becomes indistinguishable from "tat" and "bed" from "dead," etc. Elliptical speech, in which such substitutions have been made, is difficult to understand under normal conditions, but as the noise level rises, the confusion gradually fades away and linguistic and semantic cues eventually make elliptical speech more understandable.

16.5 ■ THE SYNTHESIS OF SPEECH

"If computers could speak, they could be given many useful tasks. The telephone on one's desk might then serve as a computer terminal, providing automatic access to such things as airline and hotel reservations, selective stock market quotations, inventory reports, medical data, and the balance in one's checking account" (Flanagan 1972a). Providing computers with the ability to speak has been one target of a substantial amount of research on speech synthesis.

Early efforts to imitate speech sounds resulted in various mechanical "talking machines." One such machine, invented in 1791 by Wolfgang von Kempelen of Vienna and later improved by Sir Charles Wheatstone, is shown in Fig. 16.13. A bellows supplies air to a reed, which serves as the main voice source. A leather "vocal tract" is shaped by the fingers of one hand. Consonants, including nasals, are simulated by four constricted passages controlled by the fingers of the other hand.

During his boyhood in Scotland, Alexander Graham Bell had an opportunity to see the Wheatstone reconstruction of von Kempelen's talking machine. Assisted by his father and his brother, he constructed a talking machine of his own, molding the lips, tongue, palate, pharynx, and velum in guttapercha, wood, and rubber. A larynx box of tin had vocal folds made of rubber sheet (Flanagan 1972b).

Modern talking machines use electronic rather than mechanical techniques. Computers have brought about rapid advances in speech analysis. State-of-the-art speech synthesis

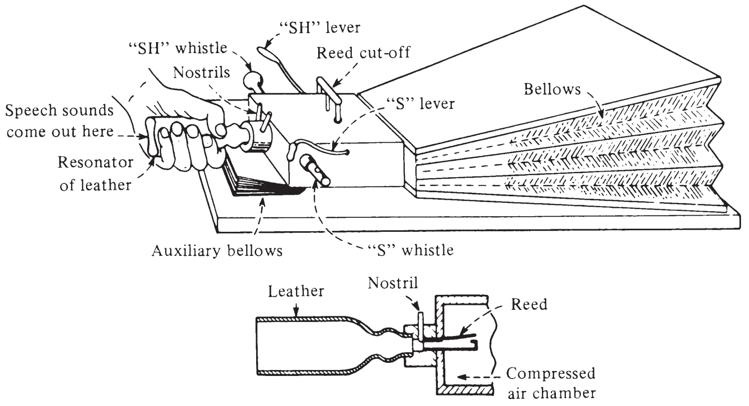


FIGURE 16.13 Wheatstone's reconstruction of von Kempelen's talking machine (Flanagan 1972b).

generally uses one of two methods: *formant synthesis* by rule or *concatenative synthesis* by computer assembly of speech from pieces of natural speech.

Formant synthesizers employ the source-filter theory of speech production (see Chapter 15, especially Section 15.5). They utilize formant patterns for each phoneme and tracks for F_0 (fundamental frequency) that specify source variations and other aspects of articulation. For example, for the word *saw*, the /*s*/-phoneme pattern might specify excitation of a filter with a formant at about 5000 Hz, followed by the /aw/ pattern, consisting of three filters with peak frequencies of 600, 1300, and 2500 Hz representing formants F_1, F_2, and F_3 (Bickley, Syrdal, and Schroeter 1999).

Of course there need to be rules for modifying the basic phoneme patterns to create natural connections between them. This is one of the critical problems for good synthesis. Formant synthesizers are efficient because they do not require a large storage capacity and the computational demands are relatively simple. The goal of natural-sounding speech for text-to-speech (TTS) applications continues to drive research on the refinement of formant synthesizers.

Concatenative synthesis methods result in highly intelligible and potentially very natural-sounding speech. Imagine that one recorded all desired words in all desired voices and intonations. This would lead to synthesized speech that sounds very natural, but it would also require a prohibitive amount of storage. Therefore, a compromise has to be made. The voice recordings need to be coded and compressed. There are various ways to do this. A device for coding speech is often called a *vocoder* (short for voice coder), although the term is also applied to a combined speech analyzer and synthesizer.

16.6 ■ SPEECH CODING AND COMPRESSION

Devices such as *channel vocoders* (which transmit information about the output from 16 filters plus another channels for information about unvoiced consonants and fundamental frequency) and*formant vocoders* (which transmit information about the formants themselves) have been successful, but the most popular technique for speech coding has probably been *linear predictive coding* (LPC), which describes a speech waveform in terms of a set of 10 or 12 time-varying parameters derived from analysis of speech samples (Atal and Hanauer 1971). Most "talking chips" in microcomputers make use of LPC, as did the remarkable "Speak and Spell" toy introduced by Texas Instruments in 1978 (Franz and Wiggins 1982). A detailed description of LPC is beyond the scope of this chapter. Other coding techniques include *pitch-synchronous overlap add* (Moulines and Charpentier 1990) and sinusoidal coding techniques (Dutoit 1997).

Bandwidth compression is desirable when speech is to be transmitted over long distances, such as in transoceanic cables, coaxial or optical. The efficiency of a speech-coding method can be expressed in terms of the perceptual quality of the decoded speech versus the required amount of information storage or transmission capacity to move it (the bandwidth). Rate of information is measured in kilobits per second (Kb/s). A speech waveform coded into digital form may require 64 Kb/s to preserve the naturalness and intelligibility of the original speech, although waveform compression techniques can reduce this to 16 Kb/s without noticeable loss of quality. Vocoding techniques, which attempt to preserve perceptual quality rather than waveform accuracy, can often reduce this to 8 Kb/s or even less. Good intelligibility can be obtained as low as 2 Kb/s, although naturalness is lost.

16.7 ■ SPEECH RECOGNITION BY COMPUTERS

Designing a machine that understands language is more difficult than building one that talks. Human listeners have learned to accept a wide range of speech input, including different dialects, accents, voice inflections, and even speech of rather low quality from talking computers. Machines to recognize speech have not yet reached this degree of flexibility, however. Machines that can recognize a limited vocabulary from one speaker will have difficulty recognizing the same words from a different speaker.

Speech recognition may focus either on recognizing individual words or on recognizing connected words in a phrase or sentence. A common strategy for recognizing isolated words is template matching. Templates of appropriate time-varying parameters are created for the words in the desired vocabulary as spoken by selected speakers. These same parameters in a spoken word are then compared to the stored templates, and the closest match is assumed to be the word spoken. Isolated word recognition is practical for such tasks as digit recognition, recognizing simple computer commands, and machine control, but not for general communication.

Continuous speech recognition is much more difficult than isolated word recognition, because it is difficult to recognize the beginning and end of words, syllables, and phonemes. In natural speech, articulatory gestures are made quickly, so that each is modified, to a certain extent, by its neighbors in the spoken sequence. This modification, which can be quite considerable, is known as *coarticulation*. The degree of coarticulation will depend on the rapidity of speech and the mode of speech. Its effect, in some ways, is analogous to the difference between hand-printed letters and handwriting in which the letters are modified as they are connected together.

Much research effort has been devoted to machine recognition of speech, because the potential applications are many. Voice-controlled typewriters and word processors may soon be possible. Voice programming of computers, control of machines, telephone dialing, data entry for materials handling and sorting, financial transactions, etc., while leaving the hands free for other tasks, are of obvious benefit.

16.8 ■ SPEAKER IDENTIFICATION BY SPEECH SPECTROGRAMS: VOICEPRINTS

Can one reliably identify a person by examining the spectrographic patterns of his or her speech? This is a question of considerable legal as well as scientific importance. The Technical Committee on Speech Communication of the Acoustical Society of America asked six distinguished speech scientists to review the matter from a scientific point of view a number of years ago (Bolt et al. 1970). They concluded that "the available results are inadequate to establish the reliability of voice identification by spectrograms." Speech spectrograms, or *voiceprints*, are not analogous to fingerprints, because they do not represent anatomical traits in a direct way. The article by Bolt et al. (1970) does, however, summarize methods used for speaker identification and their validity.

Speech spectrograms portray short-term variations in intensity and frequency in graphical form. Thus they give much useful information about speech articulation. When two persons speak the same word, their articulation is similar but not identical. Thus spectrograms of their speech will show similarities but also differences. However, there are also differences when the same speaker repeats a word, as can be seen in Fig. 16.14.

Our auditory system exhibits an amazing ability to identify speakers, especially if the voices are well known to us, even in the presence of substantial interference. However, wrong identifications are within the experience of all of us. Careful studies, in fact, have shown that listening provides more dependable identification of the speaker than the examination of spectrograms of the same utterances does (Stevens et al. 1968). Work is being done on the design and evaluation of methods for objective voice identification using com-

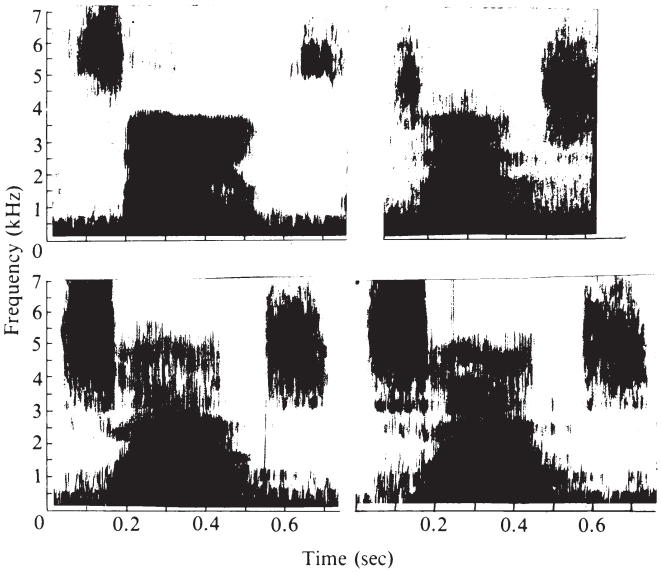


FIGURE 16.14 Four spectrograms of the spoken word *science*. The vertical scale represents frequency, the horizontal dimension is time, and darkness represents sound level. The two spectrograms at the left are by the same speaker. (Compare similar spectrograms in Bolt et al. 1970.)

pletely automatic procedures but, at this time at least, they do not inspire great confidence in the use of voiceprints for error-free speaker identification.

16.9 ■ SUMMARY

To analyze speech, it is desirable to display sound level as a function of both frequency and time. This can be done on a three-dimensional graph or by a sound spectrograph.

The first two or three formants are usually sufficient for recognition of vowel sounds even in the presence of distortion or interference. The cues for consonant recognition often depend on the vowel sound that follows. The pattern-playback machine, which generates synthesized speech with specified features, has added much to our knowledge about consonant recognition. Filtering speech and masking speech with noise reduce intelligibility.

It is now possible to build machines that synthesize speech of acceptable quality, and machines that can recognize small vocabularies of words. Other machines can identify a speaker by his or her voiceprint, but not with a high degree of reliability. Future research and development will most likely lead to machines that can speak, understand speech, and even identify a speaker.

REFERENCES AND SUGGESTED READINGS

Atal, B. S., and S. L. Hanauer (1971). "Speech Analysis and Synthesis by Linear Predication of the Speech Wave," *J. Acoust. Soc. Am.* **50**: 637.

Bickley, C., A. Syrdal, and J. Schroeter (1999). "Speech Synthesis," in *The Acoustics of Speech Communication* by J. M. Pickett. Needham Heights, Mass.: Allyn & Bacon.

Bolt, R. H., F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens (1970). "Speaker Identification by Speech Spectrograms: A Scientists' View of Its Reliability for Legal Purposes," *J. Acoust. Soc. Am.* **47**: 369.

Chari, N. C., G. Herman, and J. L. Danhauer (1977). "Perception of One-Third Octave-Band Filtered Speech," *J. Acoust. Soc. Am.* **61**: 576.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman (1952). "Some Experiments on the Perception of Synthetic Speech Sounds," *J. Acoust. Soc. Am.* **24**: 597.

David, E. E., Jr., and P. B. Denes (1972). *Human Communication: A Unified View*. New York: McGraw-Hill.

Delattre, P. C., A. M. Liberman, and F. S. Cooper (1955). "Acoustic Loci and Transitional Cues for Consonants," *J. Acoust. Soc. Am.* **27**: 769.

Denes, P. B., and E. N. Pinson (1973). *The Speech Chain*. Garden City, N.Y.: Anchor/Doubleday.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer.

Flanagan, J. L. (1972a). "The Synthesis of Speech," *Sci. Am.* **226**(2): 48.

Flanagan, J. L. (1972b). "Voices of Men and Machines," *J. Acoust. Soc. Am.* **51**: 1375.

Flanagan, J. L., and L. R. Rabiner (1973). *Speech Synthesis*. Stroudsburg, PA: Dowden, Hutchinson & Ross.

Franz, G. A., and R. H. Wiggins (1982). "Design Case History: Speak and Spell Learns to Talk," *IEEE Spectrum* **19**(4): 45.

French, N. R., and J. C. Steinberg (1947). "Factors Governing the Intelligibility of Speech Sounds," *J. Acoust. Soc. Am.* **19**: 90.

Hawkins, J. E., Jr., and S. S. Stevens (1950). "The Masking of Pure Tones and Speech by White Noise," *J. Acoust. Soc. Am.* **22**: 6.

Levinson, S. E., and M. Y. Liberman (1981). "Speech Recognition by Computer," *Sci. Am.* **244**(4): 64.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967). "Perception of the Speech Code," *Psych. Rev.* **74**: 431.

Licklider, J. C. R., and I. Pollack (1948). "Effects of Differentiation, Integration, and Infinite Peak Clipping Upon the Intelligibility of Speech," *J. Acoust. Soc. Am.* **20**: 42.

Moulines, E., and F. Charpentier (1990). "Pitch-synchronous Waveform Process Technique for Text-to-Speech Synthesis Using Diphones," *Speech Commun.* **9**, 453–467.

Pickett, J. M. (1999). *The Acoustics of Speech Communication*. Needham Heights, Mass.: Allyn & Bacon.

Rossing, T. D. (1982). *Acoustics Laboratory Experiments*. Copies from the author.

Stevens, K. N., C. E. Williams, J. P. Carbonell, and B. Woods (1968). "Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentation of Speech Material," *J. Acoust. Soc. Am.* **44**: 1596.

Stover, W. R. (1966). "Technique for Correcting Helium Speech Distortion," *J. Acoust. Soc. Am.* **41**: 70.

Strong, W. J., and G. R. Plitnik (1983). *Music, Speech and High-Fidelity*, 2nd ed. Provo, Utah: Soundprint.

GLOSSARY

coarticulation Modification of speech sounds when they are connected to other sounds in a spoken sequence.

concatenative synthesis Uses an inventory of natural speech pieces as building blocks from which an arbitrary utterance can be constructed.

cues Characteristics of speech sounds that help us to recognize them.

formant synthesis Employs the source-filter theory of speech production to synthesize speech.

fricative consonant Consonant that is formed by constructing air flow in the vocal tract (e.g., f, v, s, z, th, sh).

linear predictive coding (LPC) Describing a speech waveform in terms of a set of time-varying parameters derived from speech samples.

masking Obscuring of one sound by another (see Section 6.10).

peak clipping Limiting the amplitude of a waveform so that peaks in the waveform are eliminated; this distorts the waveform.

phonemes Individual units of sound that make up speech.

real-time spectrum analyzer An instrument that rapidly creates a spectrum of a sound.

sound spectrograph An instrument that displays sound level as a function of frequency and time for a brief sample of speech.

spectrogram A graph of sound level versus frequency and time as recorded on a sound spectrograph or similar instrument.

speech synthesis Creating speechlike sounds artificially.

vocoder A combined speech analyzer and synthesizer ("voice coder").

voiceprints Speech spectrograms from which a speaker's identity may be determined.

REVIEW QUESTIONS

1. What is a phoneme?
2. What three variables are plotted by a sound spectrograph?
3. What is a real-time spectrum analyzer?
4. Describe what happens when a speaker inhales helium before speaking.
5. What is a pattern-playback machine?
6. Describe the first and second formants associated with the consonant /t/.
7. Give an example of a consonant whose formants can move in different directions depending upon the vowel that follows.
8. What is a fricative consonant? Give an example.
9. Give an example of a semivowel.
10. What is "elliptical" speech?
11. What is concatenative synthesis?
12. What is linear predictive coding (LPC)?
13. What is a voiceprint?

QUESTIONS FOR THOUGHT AND DISCUSSION

1. When you fail to understand an indistinctly spoken word, is it more apt to be the initial consonant, the final consonant, or the vowel that is not recognized? Try to give reasons for your answer.
2. Discuss why a baritone voice played back at a higher speed than that at which it was recorded does not sound like a soprano.
3. Discuss the acoustics of the frequently heard phrase "he projects his voice." Think of other expressions used to describe good speaking techniques and their possible acoustical basis.
4. Would synthesized speech of high intelligibility but unnatural quality be useful in telecommunication? Would it be acceptable to most users of the telephone?

EXERCISES

1. From the spacing of the small striations in the speech spectrogram shown in Fig. 16.3, estimate the fundamental frequency of the speaker. Can you tell whether the speaker was male or female? (The duration of the spectrogram is 1.9 s.)
2. Recognition of vowels requires frequencies from about 200 to 3000 Hz, whereas recognition of certain consonants requires frequencies up to 8000 Hz. While listening to a radio newscast, quickly turn down the treble tone control, and note which consonants are the most difficult to identify.
3. From Fig. 16.8 estimate the frequency change in the first and second formants during articulation of "di," "da," and "du." Estimate also the time over which these formant shifts take place.
4. Time your own speech rates when speaking normally and when speaking as fast as you can. Then count the number of words in a particular paragraph, and time yourself as you read it at each of these rates.
5. Listen to speech recorded at one rate and then played back at both a faster and a slower rate. Describe the speech you hear (quality, pitch, intelligibility, etc.).

EXPERIMENTS FOR HOME, LABORATORY, AND CLASSROOM DEMONSTRATION

Home and Classroom Demonstration

1. *Vowel formants* Formants of the cardinal vowel sounds spoken by volunteers can be determined by means of a sound spectrograph or a real-time (FFT) spectrum analyzer. If the latter is used, it is best to sum up a number of words or syllables that use the vowel of interest by means of the signal averager.

2. *Changing the fundamental and formant frequencies* Tape record speech at one speed and play it back both faster and slower than recorded. The sounds are still recognizable, because the pitch and the formant frequencies have been changed by the same ratio. The quality has suffered, however.

3. *Changing the fundamental or formant frequencies* If possible, obtain tapes of synthesized speech in which the pitch and formant frequencies are scaled abnormally. Compare to the results obtained in Experiment 2.

4. *Helium speech* Take a deep breath of helium and speak. The resulting speech distortion results from a change in formant frequencies, although the pitch (determined by the vocal-fold vibration frequency) remains the same. *Be careful to replace the oxygen in your lungs as soon as possible.*

5. *Filtered speech* Record a short phrase and play it back through a variety of filters (high-pass, low-pass, octave band-pass, one-third octave band-pass). Compare the relative intelligibilities.

6. *Speech synthesis with a personal computer* A number of speech synthesis programs are available. Compare synthesized vowels with spoken vowels by listening and by real-time spectrum analysis.

7. *Speak and Spell* Texas Instruments' Speak and Spell is an expensive device for demonstrating speech synthesis.

8. *Speaker Identification* Record spectrograms (or real-time spectra) of the same two sentences or phrases spoken by several volunteers. Can you recognize which ones are spoken by the same person?

9. *Out of Context* Observe how difficult it is to understand when someone abruptly changes the topic of conversation.

10. *Visual Cues* The video "Speech Perception" (Acoustical Society of America, Melville, NY, 1997) includes a demonstration showing how one hears a different consonant when listening with the eyes closed (no visual cue) and open.

Laboratory Experiments

Speech sounds: The sound spectrograph (Experiment 23 in *Acoustics Laboratory Experiments*)

Synthesis of vowel sounds (Experiment 24 in *Acoustics Laboratory Experiments*)

Spectra of front vowels (Lab 5 in Pickett 1999)

Spectra of back vowels and diphthongs (Lab 6 in Pickett 1999)

Nasals and glides (Lab 9 in Pickett 1999)

Fricative/stop distinction (Lab 10 in Pickett 1999)

Voiced/unvoiced distinction (Lab 11 in Pickett 1999)

Coarticulation effects (Lab 12 in Pickett 1999)

CHAPTER 17

Singing

It is somewhat ironic that the oldest musical instrument of all, the human voice, is less well understood than the various instruments we discussed in Part III. This is certainly due, in part, to the inaccessibility of its various components within the human body. Studying the human voice might be likened to studying the violin without being allowed to open the case or, at best, to hearing it played from behind an opaque screen with a small hole through which to peek.

In this chapter you should learn:

- About formants in the singing voice;
- About factors influencing the spectra of sung notes;
- About breathing and air flow;
- About registers in singing;
- About the acoustics of choir singing.

The vocal organ, as shown schematically in Fig 15.1, consists of the *lungs*, the *larynx*, the *pharynx*, the *nose*, and the *mouth*. Air from the lungs is forces through the *glottis*, a V-shaped opening between the vocal cords or folds, causing them to vibrate and thus modulate the flow of air through the larynx. The output from the vocal folds is characterized as a buzz (a nearly triangular waveform), rich in harmonics that diminish at a rate of about 12 dB per octave (see Fig. 15.7).

The vocal tract, which consists of the larynx, the pharynx, the oral cavity, and the nasal cavity, acts as a filter-resonator to transform this buzz into musical sound, somewhat in the manner of the tubing of a trumpet or oboe (but without a large amount of feedback to the source). Unlike the horns of the orchestra, however, the vocal tract creates its formants (resonances) mainly by changing its cross-sectional area at various points of articulation along its length. (The vocal tract was discussed in Section 15.3.)

17.1 ■ FORMANTS AND PITCH

In both speech and singing, there is a division of labor, so to speak, between the vocal folds and the vocal tract. The vocal folds or cords control the pitch of the sound, whereas the vocal tract determines the vowel sounds through its formants and also articulates the consonants. The pitch and the formant frequencies are virtually independent of each other in speech, but trained singers (especially sopranos) sometimes tune their vowel formants

From Chapter 17 of *The Science of Sound*, Third Edition. Thomas D. Rossing, Richard F. Moore, Paul A. Wheeler.

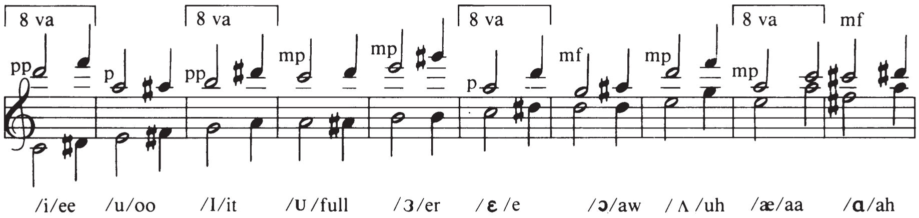


FIGURE 17.1 Typical formants of male (𝅗𝅥) and female (♩) speakers represented on a musical staff. (Compare with Table 15.3.)

to match one or more harmonics of the sung pitch. The loudness and timbre of the sung sound depend on both the vocal folds and the vocal tract.

Figure 15.16 shows the vocal tract profiles for 12 English vowels, and Table 15.3 tabulates typical formant frequencies for ten of them as spoken by both male and female voices. In Fig. 17.1, these same data are presented on a musical staff for the reader who is more familiar with this notation. It may be surprising to learn that although female voices are pitched about an octave higher than male voices, the formants usually differ by less than a musical third (less than 25% in frequency).

In Table 15.3 the relative formant amplitudes are given. For most spoken vowels, the second formant is considerably weaker than the first; the "ah" and "aw" sounds have the strongest second formants. We have added dynamic markings (pp, p, mp, mf) to Fig. 17.1 to indicate the relative strengths of the second formant. Although the first and second formants contribute almost equally to vowel sounds (the third formant contributes slightly less), the first formant will usually contribute more to timbre because of its greater amplitude and lower frequency, closer to the fundamental.

Note the position of the formants in Fig. 17.1 in relation to the singing range. In the case of the bass or baritone singer, the fundamental rarely is enhanced by a formant resonance (exceptions are "ee" and "oo" sounds near the top of the singing range). In most cases, the formants enhance higher harmonics of the fundamental; for example, if a bass sings "ah" with a pitch G_2 ($f = 98$ Hz), the first formant gives its greatest boost to the seventh harmonic, the second formant boosts harmonics around the eleventh, and the third formant gives a smaller (but important because of the frequency range in which it lies) boost to the 24th and 25th harmonics and their neighbors. The pitch, of course, remains at G_2, because the overtones are harmonics of this "almost-missing" fundamental. A few people have learned to shape their mouths in such a way that harmonics of a sung pitch can be made audible (see demonstration experiment on Single Formants in Chapter 15).

Another way to present formant frequencies of vowel sounds is shown in Fig. 17.2. Frequencies of the first formant are plotted on the horizontal scale, and those of the second formant on the vertical scale. The egg-shaped regions represent the approximate limits of formant frequencies that the ear will recognize as a given vowel. Note that there is overlap; that is, certain sounds can be interpreted as more than one vowel, depending on the context.

17.2 ■ DIFFERENCES BETWEEN SPOKEN AND SUNG VOWELS

Sung vowels are fundamentally the same as spoken vowels, although singers do make vowel modifications in order to improve the musical tone, especially in their high range.

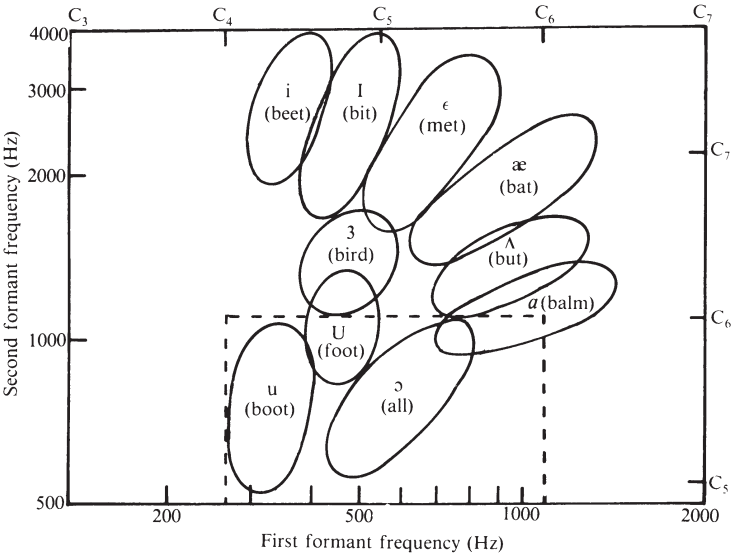


FIGURE 17.2 Frequency of first and second formants for 10 vowels. The dotted line shows the approximate range of a soprano voice. (After Peterson and Barney 1952.)

For example, "ee" is often sung like the umlauted "ü" of the German "für," and the short "e" of bed sounds more like the vowel sound in herd.

Analysis of the individual vowel formants reveals changes that may be substantial. Figure 17.3 shows spectra of the vowel /æ/ (as in bat) spoken and sung by a professional bass-baritone singer. Note that the first formant is virtually unchanged, but the second formant is lower in frequency in the sung vowel. The third and fourth formants remain at about the same pitch but are markedly stronger in the sung vowel.

Four articulatory differences between spoken and sung vowels were noted by Sundberg (1974) as a result of studying X-ray pictures of the vocal tract and photographs of the lip openings. In singing,

1. The larynx is lowered;
2. The jaw opening is larger;

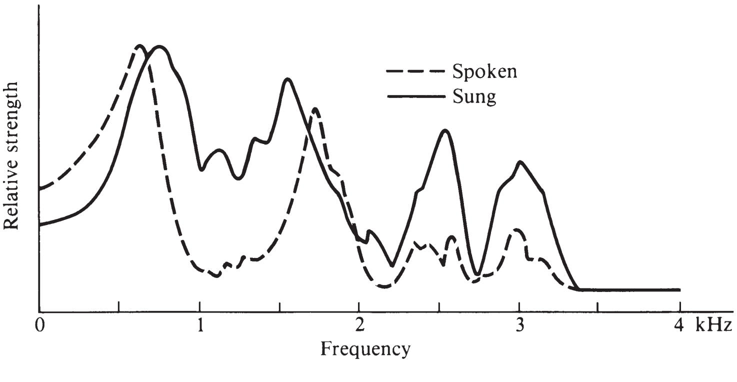


FIGURE 17.3 Spectra of vowel sound /æ/ as spoken and sung by a professional singer.

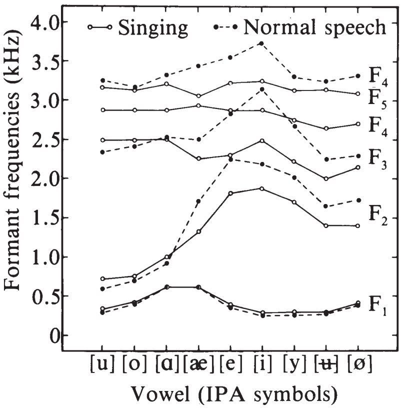


FIGURE 17.4 Formant frequencies of long Swedish vowels in normal male speech (dashed lines) and in professional male singing (solid lines). (From Sundberg 1974.)

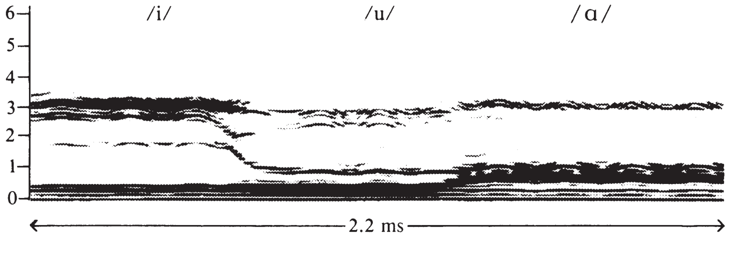
3. The tongue tip is advanced in the back vowels /u/, /o/, and /ɑ/; and
4. The lips are protruded in the front vowels.

Formant frequencies of nine sung and spoken vowels are shown in Fig. 17.4. Frequencies of the lower two formants do not change very much, although the second formant is slightly lower in singing the front vowels. The third and fourth formants are substantially lower when sung, and a fifth formant is now apparent.

Trained singers, especially male opera singers, show a strong formant somewhere around 2500–3000 Hz. This "singer's formant," which seems to be more or less independent of the particular vowel and the pitch, usually lies between the third and fourth formants and adds brilliance and carrying power to the male singing voice. It is interesting to note that the frequency of this formant is near the resonance frequency of the ear canal, which gives it an additional auditory boost. A formant of 3000 Hz is evident in the spectrograms shown in Fig. 17.5.

Sundberg (1974) attributes the singer's formant to a lowered larynx, which, along with a widened pharynx, forms an additional resonance cavity (about 2 cm long) with a frequency in the range of from 2500 to 3000 Hz. Lowering the larynx also produces the darker vowel sounds favored by most singers. The larynx, which is lowered as much as 30 mm during

FIGURE 17.5 Spectrogram of vowels /i/, /u/, and /ɑ/ (ee, oo, ah). The pitch is E3 (f = 165 Hz). Note the strong formant at 3 kHz for all three vowels. (From van den Berg and Vennard 1959.)



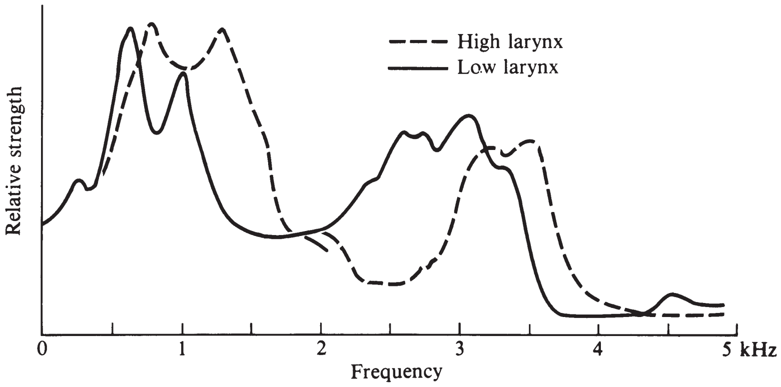


FIGURE 17.6 Spectrum of vowel sound /ɑ/ sung with a high and a low larynx.

swallowing, may be lowered up to 15 mm during singing (Shipp 1977). Untrained singers tend to raise their larynxes as they raise the pitch.

Figure 17.6 shows spectra of the vowel sound /ɑ/ ("ah") sung with both a high and a low larynx by a professional bass-baritone singer in our laboratory. The broad resonance extending from 2500 to 3000 Hz in the spectrum of the low larynx is a blend of the third vowel formant and the singer's formant.

Because the singer's formant requires a widened pharynx ("open throat"), it is characteristic of good singing in the chest register (see Section 17.4). Professional contraltos usually have such a formant, but sopranos, who sing mainly in the head register, may not. It is not usually present in the falsetto voice of the male singer, either. Figure 17.7 shows how

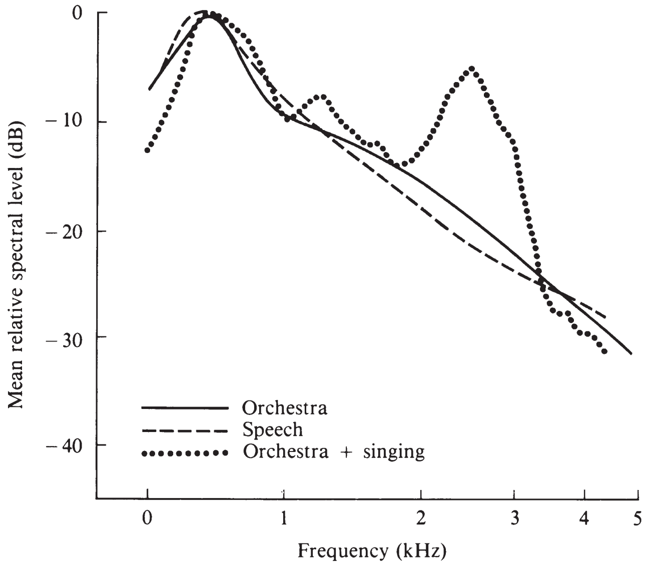


FIGURE 17.7 Idealized average spectra of normal speech and orchestra music. The dotted curve shows the average spectrum of Jussi Björling singing with a loud orchestra accompaniment. (From Sundberg 1977a.)

TABLE 17.1 Formant frequencies of basic sung vowels

Formant frequency (Hz)		/i/ (ee)	/I/ (i)	/ϵ/ (e)	/æ/ (aa)	/ɑ/ (ah)	/ɔ/ (aw)	/U/ (u̇)	/u/ (oo)	/ʌ/ (u)	/ɜ/ (er)
F_1	M	300	375	530	620	700	610	400	350	500	400
	W	400	475	550	600	700	625	425	400	550	450
F_2	M	1950	1810	1500	1490	1200	1000	720	640	1200	1150
	W	2250	2100	1750	1650	1300	1240	900	800	1300	1350
F_3	M	2750	2500	2500	2250	2600	2600	2500	2550	2675	2500
	W	3300	3450	3250	3000	3250	3250	3375	3250	3250	3050

Source: Appelman (1967).

the singer's formant in the voice of operatic tenor Jussi Björling helped him "cut through" a large orchestra.

It is obvious from Fig. 17.2 that the formant frequencies of different speakers (and singers) may vary rather widely, yet still result in understandable vowel sounds. Furthermore, in certain ranges of singing, the vowel formants change substantially from their normal frequencies. Nevertheless, it is instructive to compare the formant frequencies of typical *sung* vowels given in Table 17.1 to the corresponding formant frequencies of the *spoken* vowels given in Table 15.3.

Formant changes that occur throughout the singing range may be roughly described as the gradual substitution of one vowel sound for another. As the pitch rises, for example, many singers find it convenient to make the following substitutions (Appelman 1967):

Normal range	/i/	/ϵ/	/æ/	/*a*/	/ɔ/	/u/
High range	/I/	/ɑ/	/ɑ/	/ʌ/	/ʌ/	/U/

Some voice teachers recommend the substitution of a long (closed) vowel for a short (open) vowel as the pitch rises and the converse for a downward crescendo (Haasemann and Jordan 1991). The vowel sounds /I/, /u/, /ʌ/, and /æ/, which do not change substantially with pitch, are termed *stable*.

17.3 ■ FORMANT TUNING BY SOPRANOS

In low voices, the various formants of the local tract emphasize various harmonics of the source sound from the glottis, as we discussed in Section 17.1. Sopranos, however, do much of their singing in a range in which the pitch exceeds the frequency of the first formant. Thus they would not receive the benefit of a boost from formant resonance, and their tones would suffer in quality and loudness. Experienced sopranos have learned how to "tune" their formants over a reasonable range of frequency in order to make a formant coincide with the fundamental or one of the overtones of the note being sung.

For example, a soprano singing /i/ ("ee") at a pitch of F_5 (698 Hz) might find her normal first formant at 310 Hz, more than an octave below the sung pitch. She would receive little support from this formant. However, if she opens her lips somewhat wider than the normal position for speaking /i/, the formant can be pushed up to the vicinity of the sung pitch. Or if she were singing /ɑ/ ("ah") at a pitch of A_4 (440 Hz), she would find her normal

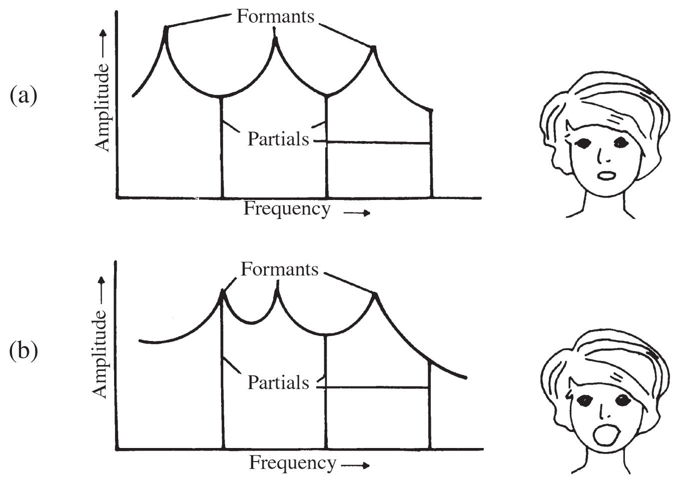


FIGURE 17.8 Formant tuning by wider jaw opening: (a) normal first formant lies below the sung pitch; (b) first formant raised to coincide with sung pitch. (From Sundberg 1977a.)

first formant around 700 Hz, between the fundamental (440 Hz) and the second harmonic (880 Hz) of the sung note. She would probably find it more convenient, in this case, to raise the formant to the vicinity of the second harmonic in order to provide the needed boost. Figure 17.8 shows how formant tuning can be accomplished by increasing the jaw opening to change the shape of the vocal tract.

Figure 17.9, also based on the work of Sundberg, shows the extent to which formant tuning can take place to match one of the harmonics of various sung pitches. The numbered

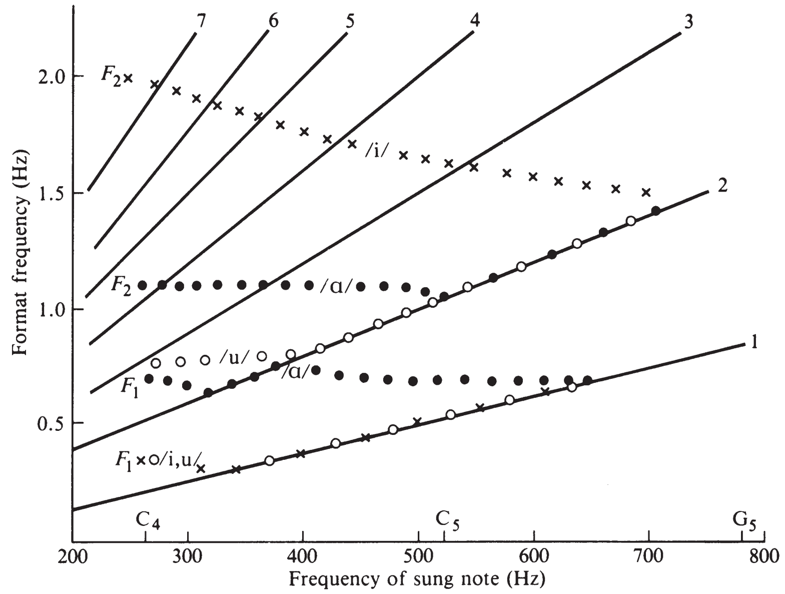


FIGURE 17.9 The tuning of formants to match harmonics of the sung note. F_1 and F_2 are the lowest formants of vowels /i/, /ɑ/, and /u/. The solid lines are the first seven harmonics of the sung note. (After Sundberg 1975.)

lines represent the harmonics of the pitch. Note that formants are usually tuned upward, although downward tuning is also possible.

Formant tuning might be expected to produce objectionable distortion of vowel sounds, but this does not seem to be the case. We are accustomed to recognizing vowels produced at various pitches in the speech of men, women, and children (see Table 15.3) with vocal tracts of different lengths. If the pitch is high, we associate it with relatively high formant frequencies. Recording a vowel sound at one speed and playing it back at another may change it to another vowel sound because the same ratio of f_2/f_1 in the new pitch range is interpreted differently. An "ah" changes to an "oh" when played at half speed (Benade 1976).

Near the top of the soprano range, where formant tuning is particularly marked, it is difficult to distinguish one vowel sound from another. (Try listening to a soprano sing various vowels at a high pitch, and see if you can recognize them.) Composers are quite aware of this difficulty in vowel recognition and generally do not present important text at the top of a soprano's range (if they must, they generally repeat the text at a lower pitch).

17.4 ■ BREATHING AND AIR FLOW

Most singers attach quite a measure of importance to good breathing habits. This may be somewhat of a mystery, however, because the only thing required from the breathing mechanism is that it supply air to the larynx at the desired *subglottal pressure*. Different singers appear to use different muscular strategies to accomplish this. For example, some singers sing with the abdominal wall expanded ("belly out") and some with it contracted ("belly in"). Likewise, different singers use their diaphragms differently during singing. In order to understand the different strategies of breathing, we briefly consider how the human respiratory system operates.

The lungs are spongy, elastic structures; they act somewhat like toy balloons. When inflated, they exert a passive expiratory force that increases with the amount of air inhaled. After maximum inhalation, the pressure is around 2 kPa (equivalent to 10 cm of water gage), which is about one-fiftieth of atmospheric pressure.

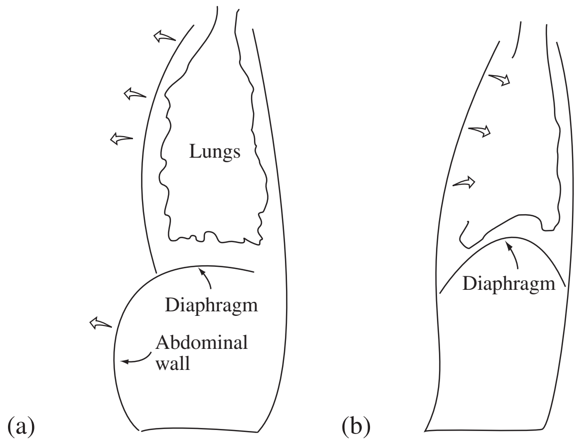


FIGURE 17.10 Schematic of the breathing apparatus. (a) Expansion of the chest cavity by the external intercostals, along with a flattening of the diaphragm, reduces the pressure and causes the lungs to expand and fill with air. (b) Contraction of the chest cavity by the internal intercostals, along with raising the diaphragm, causes the lungs to contract.

Because the lungs have no muscles of their own, breathing is accomplished by changing the size of the chest cavity. There are two basic mechanisms for doing this:

1. Downward movement of the diaphragm to lengthen the chest cavity;
2. Elevation of the ribs to increase the front-to-back thickness of the chest cavity.

Normal quiet breathing is accomplished almost entirely by movement of the diaphragm. During maximum breathing, however, increasing the thickness of the chest cavity may account for up to half of the chest cavity enlargement.

Breathing is handled by two major muscle groups. In one group are the external and internal intercostals that expand and contract the rib cage. The second group, which includes the muscles in the abdominal wall and the diaphragm, changes the abdominal cavity, as shown in Fig. 17.10. The rib cage is an elastic system that is expanded and contracted by the *intercostal muscles* that join the ribs. Muscles can only contract, and they frequently work in opposing pairs (such as the biceps and triceps in the upper arm). The *external* (inspiratory) intercostals function so that a contraction leads to an increase of the rib cage volume, whereas the *internal* (expiratory) intercostals decrease the rib cage volume when they contract. The diaphragm and abdominal muscles also work as an opposing pair, with the diaphragm acting as an inspiratory muscle, and the abdominal muscles are expiratory in nature.

The diaphragm is an important breathing muscle. When relaxed, it assumes a shape like an upside-down bowl; when it contracts, it flattens into a plate, as shown in Fig. 17.10(a). This action also increases the volume of the rib cage, and helps to draw air into the lungs. The diaphragm is a muscle specifically for inhalation. The volume of the abdominal cavity cannot be easily altered, so flattening the diaphragm tends to push the abdominal wall outward. Outward movement of the abdominal wall is, therefore, a visible indication of diaphragm contraction. Conversely, when the abdominal muscle is contracted, the abdominal wall moves inward, and the diaphragm returns to its bowl shape. Thus breathing can be mainly by the use of the intercostal muscles, mainly by the use of the diaphragm, or a combination of both.

The muscular activity required for maintaining the desired subglottic pressure is dependent on the lung volume, because the passive elastic forces of the lungs and the rib cage tend to raise or lower the air pressure inside the lungs, depending on whether the lung volume is greater or less than the *functional residual capacity* (FRC), the volume of air in the lungs at the end of a quiet expiration. When the lungs are filled with a large quantity of air, the passive exhalation force is large, and thus the pressure tends to rise. If this pressure is too high for the intended phonation, it can be reduced by a contraction of the diaphragm and/or the inspiratory intercostals. When the lung volume drops below the FRC due to singing a long phrase, the subglottal pressure will tend to drop, and it can be brought back up to the desired level by contracting the abdominal wall muscle and/or the expiratory intercostals.

Figure 17.11 shows how the total lung capacity is divided into four different volumes:

1. The *tidal volume* is the volume of air moved in and out during normal breathing (about 500 cm^3 in the normal male adult).

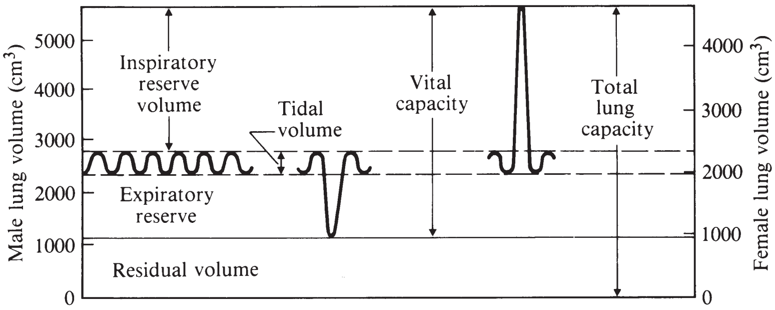


FIGURE 17.11 Lung capacity in the normal young adult and its subdivision into functioning volumes. The volume of the male lung is indicated at the left, female at the right.

2. The *inspiratory reserve volume* is the volume that can be inspired beyond the normal tidal volume (about 3000 cm^3).
3. The *expiratory reserve volume* is the volume that can be expired by forceful effort at the end of normal tidal expiration (about 1100 cm^3).
4. The *residual volume* is the volume of air that remains in the lung after forceful expiration (about 1200 cm^3).

Corresponding volumes in the female lung average about 20 to 25% less than those given for the male lung.

Vital capacity is the amount of air that can be moved in and out of the lung with maximum effort. The average vital capacity in the young adult male is about 4600 cm^3 and in the young adult female is about 3100 cm^3. Pathological conditions such as tuberculosis, emphysema, chronic asthma, lung cancer, bronchitis, and pleurisy can greatly decrease vital capacity. At a normal breathing rate of 10 breaths per minute, about 5000 cm^3 of air will be moved in and out of the lungs per minute. A young male adult can breathe at a rate as high as 2500 cm^3/s for a short period of strenuous exercise.

17.5 ■ SUBGLOTTAL PRESSURES IN SINGING

Few measurements of subglottal pressure have been published due to the difficulty of measuring it. The most reliable results are obtained by inserting a thin needle into the trachea through the tissues below the cricoid cartilage, and a few singers have submitted themselves to this procedure. There is, however, an indirect method of measurement. When the lips are closed and the glottis is open, the subglottal pressure is equal to the pressure in the mouth cavity. Therefore, the subglottal pressure can normally be determined from the oral pressure during the production of the consonant /p/ (Sundberg 1987).

In order to increase the sound pressure level, it is necessary for the singer to increase the subglottal pressure. This is illustrated in Fig. 17.12, which shows subglottal pressure for a tenor who sang a chromatic scale at piano, mezzoforte, and forte levels. Note that the subglottal pressure increases with phonation frequency as well as with loudness.

During normal quiet breathing, the air pressure in the lungs will be about 100 N/m^2 (1 cm H_2O) above and below atmospheric pressure. During maximum expiratory effort

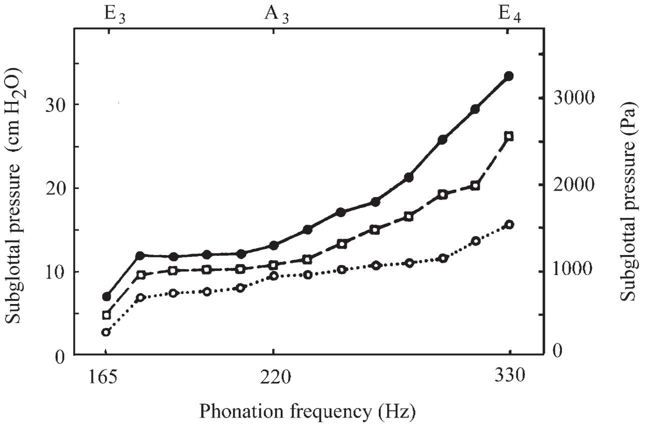


FIGURE 17.12 Subglottal pressure in a tenor who sang a chromatic scale between E_3 and E_4 at *f*, *mf*, and *f* levels. The pressure increases for increasing level and also with increasing frequency. (After Cleveland and Sundberg 1983.)

with the glottis closed, a pressure of 10,000 N/m^2 is possible in the strong healthy male lung. Fortissimo singing requires a pressure of 3000 to 4000 N/m^2 (compare woodwind blowing pressures given in Fig. 12.22). Figure 17.13 shows the subglottal air pressure (essentially equal to lung pressure) required for different sound levels in the case of four singers.

The relationship between sound level and subglottal pressure, shown in Fig. 17.13, tends to be the same for trained and untrained singers. This is not true of the relationship between sound level and air-flow rate, however. In trained singers, the pressure and rate of air flow tend to increase together with sound level, so a trained singer can sustain a soft tone for a long time. Many untrained singers, however, require a fairly large air-flow rate in order to sing softly, the flow rate reaching a minimum for *mf* or *f* dynamic. Flow rates range from about 100 to 400 cm^3/s at different dynamic levels of singing (Bouhuys et al. 1968).

One recent set of experiments (Leanderson, Sundberg, and von Euler 1987) set about determining the role of diaphragm activity during singing (some singers use their diaphragms

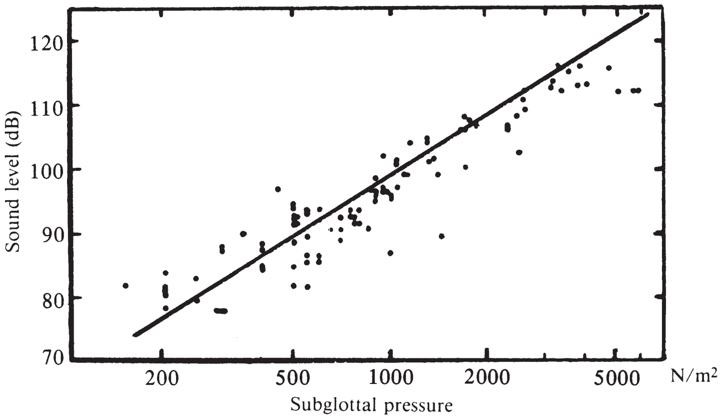


FIGURE 17.13 Subglottal pressures and sound levels for many different tones from four singers. (After Bouhuys et al. 1968.)

only during inspiration, others contract them during the entire phrase). By simultaneously monitoring pressure across the diaphragm, sound pressure level, and air flow, it was found that the flow rate tended to be higher when the diaphragm was activated, although there were substantial differences in diaphragm use between individual singers. It is probably safe to conclude that use of the diaphragm is not the key to good singing.

Although it would appear that all a singer requires from his or her breathing apparatus is to maintain a stable subglottic pressure, the emphasis put on breathing technique by many singing teachers suggests that the manner of breathing is of some importance.

Even though the way in which the vocal folds vibrate at a given subglottal pressure is determined by the laryngeal musculature, there appear to be some ties between the musculature used for breathing and that used for phonation. The way in which the subglottal pressure is controlled by the respiratory muscle system may generate reflexes that affect the laryngeal muscles.

In normal speech the passive expiratory recoil forces tend to be more important in establishing the desired subglottal pressure, whereas in singing, active muscles appear to be more important (Sundberg 1987). Learning to control these muscles is an important part of learning to sing.

17.6 ■ REGISTERS, VOICES, AND MUSCLES

We discussed the larynx and vocal cords and their functions in speech in Section 15.2. Although their functions are essentially the same in singing, there are a few additional features, such as muscular action, which become important when analyzing the singing voice.

The principal muscles internal to the larynx are the thyroarytenoids, the cricothyroids, and the cricoarytenoids (the names indicate which two cartilages they connect). The *cricoarytenoid muscles* operate the arytenoid cartilages to which the posterior ends of the vocal folds or vocal cords are attached, as shown in Fig. 15.3.

The *cricothyroids* connect the two large cartilages of the larynx, the thyroid, and the cricoid (see Fig. 15.2). They can pull the thyroid forward, with respect to the cricoid, and also downward, closer to it. Both of these actions stretch the vocal folds longitudinally, which is one way of increasing their rate of vibration. The action of the cricothyroid muscles can be observed in two ways: One is to press inward on the Adam's apple while singing a note in midrange. Sudden release of the pressure will cause the pitch to go up. A second experiment consists of placing a finger in the small space between the thyroid and cricoid cartilages while singing. Raising the pitch an octave will force the finger outward as the thyroid is pulled down closer to the cricoid.

The *thyroarytenoids*, also called the vocalis or vocal muscles, form the body of the vocal folds themselves. They extend from the notch of the thyroid to the arytenoid cartilages at the rear and are covered with a membrane that is continuous with the lining of the rest of the larynx. The tension on the vocal folds is a complex balance of forces from all three muscles, and coordination between them is necessary for smooth transition from one pitch to another. In order to hold a steady pitch during a crescendo or diminuendo, these muscles must compensate for the tendency of pitch to rise as the velocity of air flow is increased (due to the Bernoulli force; see Section 11.4).

A good description of the singing voice requires the expression of at least three quantities: fundamental frequency, amplitude, and spectrum (which are closely related, as we know from Chapters 5–7, to the perceived qualities of pitch, loudness, and timbre). The fundamental frequency of the vocal folds is controlled mainly by the laryngeal muscles we have just discussed. The amplitude (loudness) is controlled mainly by the subglottic pressure (which, in turn, is controlled by the respiratory muscles, as discussed in Sections 17.4 and 17.5). The timbre of the singing voice also depends upon the nature of the vocal-fold vibrations and thus depends upon both the laryngeal and the respiratory muscles.

Most descriptions of the way in which these muscles are used in singing include the term *register*. Unfortunately, there is no universally accepted definition of this term. The most common description is that a register is a frequency range in which all tones are perceived as being produced in a similar way and possess a similar voice timbre. According to Holien (1974), "a vocal register is a totally laryngeal event; it consists of a series or a range of consecutive voice frequencies which can be produced with nearly identical phonatory quality." Sundberg's (1987) book on *The Science of the Singing Voice* includes an excellent discussion of registers, and Miller (2000) has written an entire book on the subject.

Much has been said about the various registers used in singing. An idealistic approach is *one register*. The voice, if possible, should produce all the pitches of which it is capable without breaks or radical changes in technique. Some teachers feel that the best way to make this ideal come true is to assume that it *is* true, that it *can* be accomplished.

A more realistic approach is *three registers*. These correspond to the differences in tone caused by different adjustments of the larynx. The registers go by various names, but the most common are *chest*, *middle*, and *head* (in male voices, they are sometimes labeled *chest*, *head*, and *falsetto*). The famous teacher Mathilde Marchesi was obviously a proponent of this approach as she wrote, "I most emphatically maintain that the female voice possesses *three* registers, and not *two*, and I strongly impress upon my pupils this undeniable fact, which, moreover, their own experience teaches them after a few lessons." According to Marchesi (1970), the highest note in the chest register is about E_4 to F_4 for sopranos and F_4 to $F_4^\sharp$ for mezzo-sopranos and contraltos. The highest note in the middle register is about F_5. This is in agreement with the register ranges shown in Fig. 17.14.

The third approach is *two registers*, which considers that every voice has a potential of roughly two octaves of "heavy" mechanism, with about one octave of overlap. This middle octave can be sung in either laryngeal adjustment, and it is possible to combine some of the best qualities of both. Basses and contraltos sing almost exclusively with the heavy mechanism, mixing in just a bit of the light mechanism at the top of their range. However, the light mechanism is never used exclusively except for comic effects. Lyric

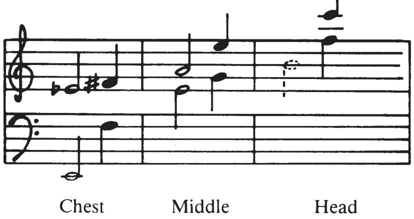


FIGURE 17.14 Ranges of three registers (according to Mackworth-Young, 1953). Half notes represent the male voice, quarter notes represent the female voice. Male voices use head register when singing falsetto.

and coloratura sopranos, on the other hand, sing with the light mechanism, mixing in just a little of the heavy at the bottom of their range, but never singing in a pure chest voice (Vennard 1967).

Because of the confusion associated with the use of the term register, it is preferable to refer to the two modes of vocal fold vibration as two mechanisms, heavy and light. We will call them *chest* (modal) *voice* and *head* (falsetto) *voice*. The distinguishing feature seems to be in the state of the thyroarytenoid muscles. In the heavy or chest voice, these muscles are active; in the light or head voice, they are virtually passive. An analogy can be drawn between these two modes of vocal fold vibration and the vibrations of the lips of a trumpet player (with active muscles) as opposed to the (passive) vibrations of a clarinet reed.

In the chest voice, the thyroarytenoids or vocalis muscles are active and hence shortened. At the lowest tones, the muscles are relaxed and the vocal folds are thick. Because of their thickness, the glottis closes firmly and remains closed an appreciable part of each cycle of vibration, as it does during speech (see Fig. 15.5).

As the pitch rises in chest voice, the cricothyroid muscles contract and apply tension to the vocal folds. The folds do not elongate rapidly, however, because the thyroarytenoid muscles come into action, and indeed thicken the vocal folds as they do so. At the top notes of the chest voice, the thyroarytenoids of the inexperienced singer sometimes give way to excessive force from the cricothyroids, and the voice "cracks" into an involuntary head tone (Vennard 1967).

In the light mechanism or head voice, the thyroarytenoids offer little resistance to the cricothyroids, which can then apply substantial longitudinal tension to the vocal folds, thus elongating them and making them thin. The vocalis muscles fall to the sides, and the vibration takes place almost entirely in the ligaments with much less amplitude of movement than in the chest voice. The glottis closes only briefly, or not at all, and the resulting sound has fewer harmonics than the chest voice does. According to studies by van den Berg (1968) on isolated larynxes, elongations of 30% are typical, as shown in the graph of stress versus strain in Fig. 17.15. Stress is the force applied per unit area and strain is the percentage by which the length increases.

The manner in which the vocal folds vibrate in the chest voice and in the head voice is shown schematically in Fig. 17.16.

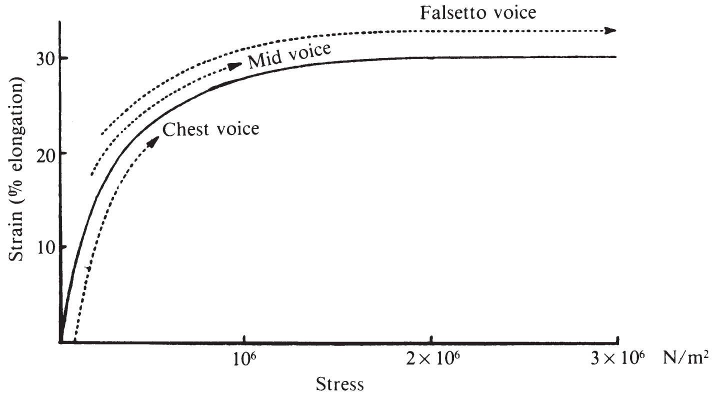


FIGURE 17.15 Stress-strain graph for vocal ligaments under passive tension (similar to head voice). The horizontal axis represents applied stress of force. (From van den Berg 1968.)

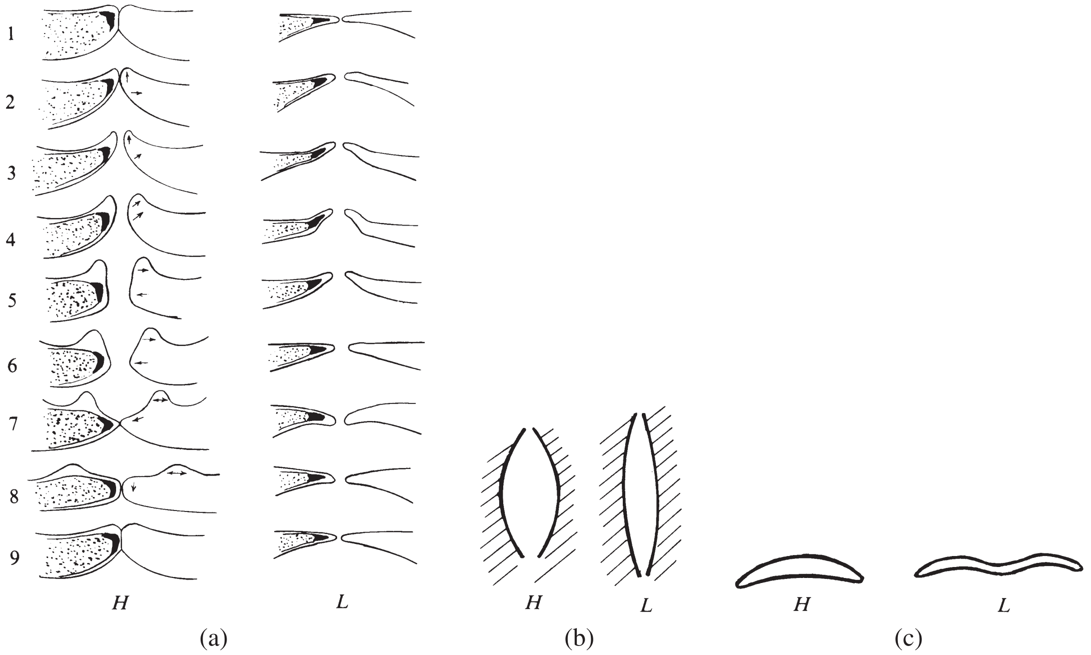


FIGURE 17.16 Schematic diagrams of vocal folds vibrating: (a) side view (from Titze 1973); (b) top view; (c) edge view. In each diagram *H* denotes the heavy mechanism (chest voice) and *L* the light mechanism (head voice).

When vibrating in the light mechanism, the vocal folds are up to 30% longer, are appreciably thinner, and have a smaller effective mass. The folds do not ordinarily close completely during any part of the cycle (compare the *open-phase* speech mode described in Section 15.2). This results in fewer harmonics of the fundamental and also in less effi-

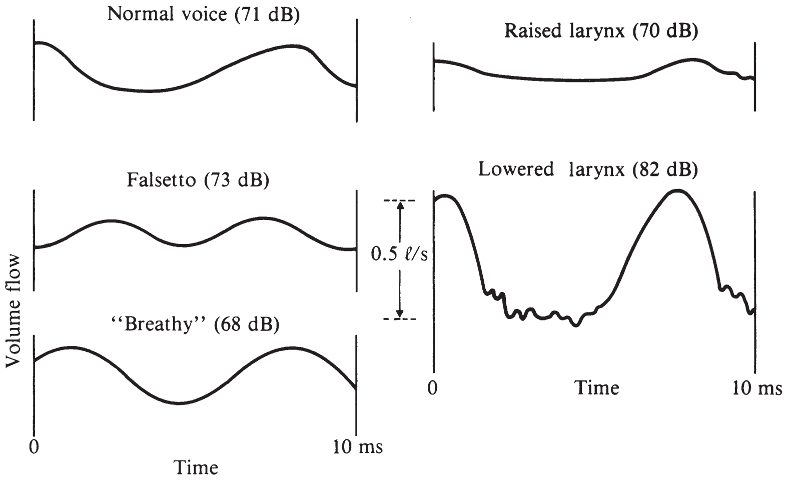


FIGURE 17.17 Waveforms of glottal air flow during various modes of singing. (After Sundberg 1978.)

cient conversion of breath power into sound power. The waveforms of the glottal air flow during speech and their spectra of overtones were shown in Figs. 15.6 and 15.7. Waveforms of glottal air flow during various modes of singing are shown in Fig. 17.17.

Two other more unusual registers deserve to be mentioned as well. One is the *Strohbass* register used by male voices to produce very low bass notes, such as those required in some Russian choir music. This is also called the *vocal fry* register, because it makes use of a loose glottal closure that is termed vocal fry by many speech therapists. Although used mainly to extend the voice below the singer's normal range, it is also used to help train the low notes in the chest or modal register (McKinney 1994).

Some singers denote a special "whistle," or flageolet, register at the top of the female vocal range. Although early writers suggested that the vocal folds actually puckered like lips to form a whistle, it is probably more correct to say the flutelike sound results from air passage through a small opening between the arytenoid cartilages.

17.7 ■ OTHER FACTORS INFLUENCING THE SPECTRA OF SUNG NOTES

It is characteristic of nearly all musical instruments that raising the dynamic level increases the levels of the higher harmonics more rapidly than that of the fundamental (see, for example. Figs. 11.13(c) and 12.16). The same effect is observed in singing, as can be seen in Fig. 17.18. In loud singing a greater fraction of the total sound energy appears in the higher harmonics as compared to soft singing.

The reason for this gain in energy in the higher harmonics can be seen by comparing the glottal air-flow waveforms (*glottograms*) in Fig. 17.19. As the loudness of phonation is increased, the rate of closure of the glottis (indicated by the slopes of the heavy lines drawn along the trailing edges of the waveforms) increases. Fourier analysis shows that waveforms with rapid rates of rise or fall have spectra rich in harmonics (compare Fig. 7.11).

Normally, a male voice tends to have weaker fundamental and stronger harmonics than a female voice, as shown in Fig. 17.20(a). Also, a male voice singing falsetto has a stronger fundamental and weaker harmonics than when singing the same note in the modal register, as shown in Fig. 17.20(b). In Fig. 17.20 the vertical axis shows only the deviation from

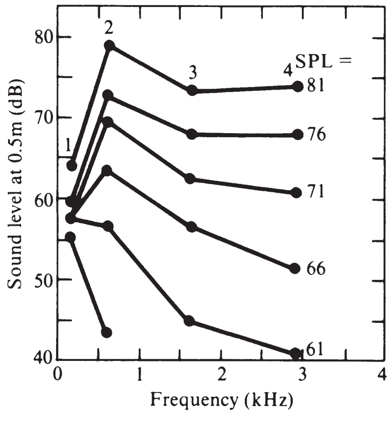


FIGURE 17.18 Sound pressure level in the first four harmonics at different total sound pressure levels. In soft phonation, the fundamental dominates, but the higher harmonics take on increasing importance as the loudness increases. (From Sundberg 1987.)

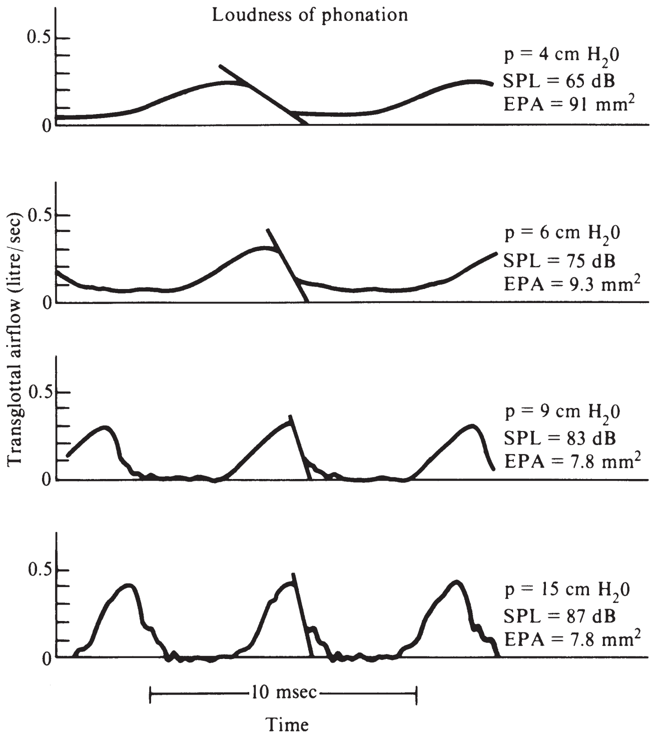


FIGURE 17.19 Glottal waveforms for four different levels of phonation. The rate of glottal closure increases as the phonation level increases. (From Sundberg 1987.)

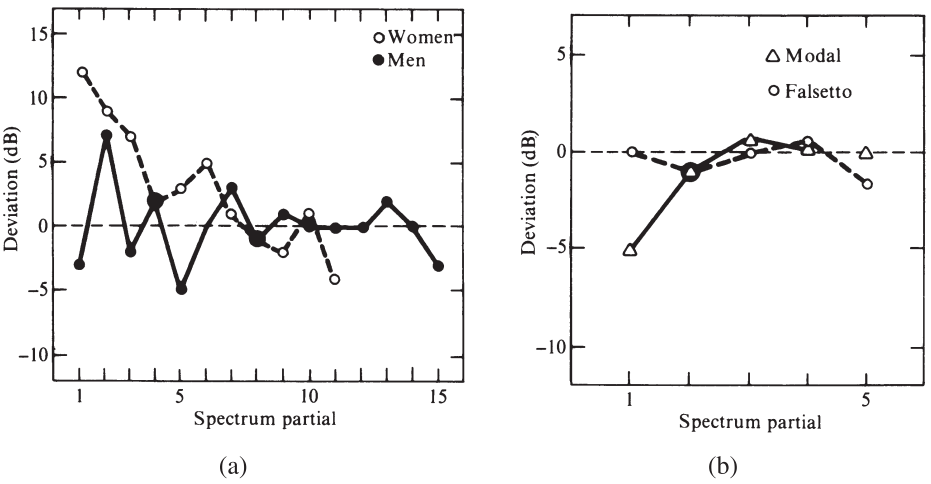


FIGURE 17.20 (a) Relative strengths of harmonics in male and female voices. (b) Relative strengths of harmonics in a male voice in the modal and falsetto registers. In both cases the vertical axis shows the deviation from the overall decrease of 12 dB/octave that characterizes voice source. (From Sundberg 1987.)

the overall decrease of 12 dB/octave that characterizes the voice source in both speech and singing.

A certain number of tones can be sung using either chest or head voice. At a given frequency, the chest voice source offers greater resistance to glottal flow than the head voice. Thus, it is possible to increase the subglottal pressure and thereby achieve greater loudness. Furthermore, the chest voice has greater harmonic content, which makes it possible to sound louder at low fundamental frequencies.

17.8 ■ CHOIR SINGING

Choral singing and solo singing are two distinctly different modes of musical performance, making different demands on the singers. Most research on the acoustics of singing has been directed at solo singing, and so less is known about the voice use in choir singing.

The first author had the opportunity of participating in some experiments at the Royal Institute of Technology (KTH) in Stockholm, which compared identical passages sung by experienced singers in solo and choir modes. A number of differences were noted, in both male and female singers.

Male singers tended to employ a more prominent singer's formant in the solo mode, as can be seen in Fig. 17.21, while the fundamental is emphasized more in the choir mode, as might be expected (Rossing, Sundberg, and Ternström 1986). It appeared that this was accomplished through adjustments in both articulation (adjustment of formant frequencies) and phonation (change in the glottal waveform).

Female singers also tend to produce more energy in the range 2 to 4 kHz in the solo mode, as shown in Fig. 17.22, although different subjects appear to differ substantially in spectral characteristics in this frequency range. It is more difficult to obtain accurate glottal waveforms from female singers, so it is difficult to distinguish changes in articulation from voice source changes. The extent of vibrato appeared to be greater in the solo mode (Rossing, Sundberg, and Ternström 1987).

Other studies by the Stockholm group have observed the degree of unison and accuracy of intervals in choir singing, both under normal conditions and when singers were deprived of feedback from other singers. In a good amateur choir, the standard deviation in the notes

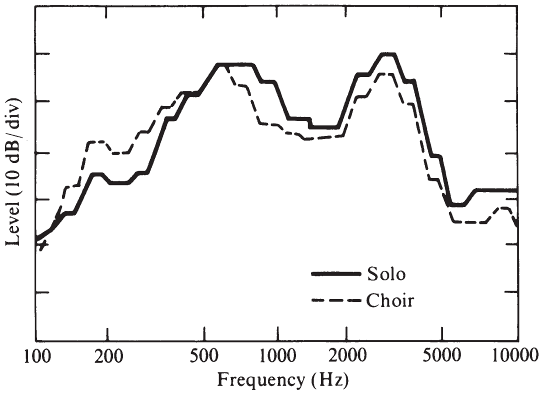


FIGURE 17.21 Average spectrum envelopes for a male singer who sang a phrase as a solo singer and as a choral singer. In the latter case his lowest partials are somewhat stronger and his singer's formant is slightly weaker. (From Rossing, Sundberg, and Ternström 1986.)

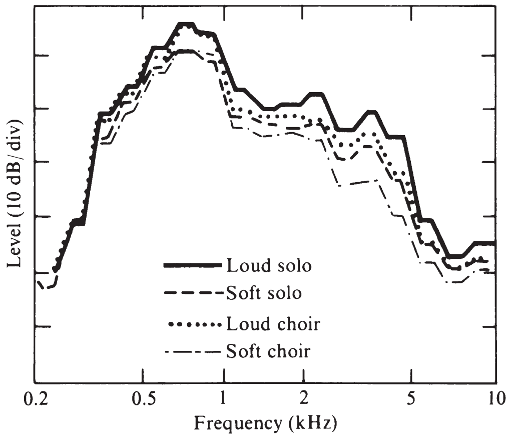


FIGURE 17.22 Average spectrum envelopes for a female singer who sang the same phrase at two different sound levels as a solo singer and as a choral singer. Choral singing gives slightly weaker high partials. (From Rossing et al. 1987.)

sung by members of the bass section was found to be 16 cents (Ternström and Sundberg 1988). In another experiment, singers were asked to sing a note at constant sound level and with the same pitch, as reference tones presented at different levels. When the reference tone was the vowel /*a*/, they were able to do this quite well, but when the vowel was /u/, the singers sang about 25 cents sharp with the softest reference tone and about 45 cents flat with the loudest reference tone. Apparently the relatively low number of harmonics in the /u/ tone (due to the low frequencies of the first and second formants) was the cause of this (Sundberg 1987).

17.9 ■ POPULAR SINGING AND OTHER STYLES

Because there are more popular ("pop") singers than classical singers in contemporary society and their media exposure is considerably greater, more young singers emulate the stars of popular entertainment than of opera and classical music. A wide variety of singing styles exist in nonclassical music, and relatively few of them have been studied scientifically. We will, however, attempt a brief discussion of a few of them.

There are a few easily recognized features of most popular singing styles. First of all, the texts of songs play a very important role in the total effect, more so than in most classical singing. The texts are often witty or carry an emotional message; it is important that they be understandable, even on first hearing. Therefore, the singer is permitted less of the vowel modification used by classical singers to enhance the instrumental beauty of the vocal line.

Second, a high value is put on naturalness of the sound, even at the expense of beauty. The use of some vocal techniques cultivated in classical singing are avoided in order not to make the voice sound well trained. The dark, or covered, voice seems particularly offensive to the impression of naturalness. On the other hand, unevenness and certain features of individual voices are readily tolerated (Schutte and Miller 1993).

Third, the performer is generally considered more important than the composition, and so the song is freely changed in order to show off the singer's voice to best advantage.

However, this can easily be carried too far, as in some of the modifications made to the national anthem by some popular singers.

We have already discussed how registers overlap and how many singers can sing a number of notes in two different registers. *Belting* describes a manner of loud singing used by female popular singers to extend the chest register above its normal range. It is characterized by a raised larynx and matching the first formant with the second harmonic on open (high F_1) vowels.

Schutte and Miller (1993) compare the spectra of the same vowel sung by a versatile mezzo soprano in classical, popular, and belt styles. In the classical style, the F_1 and F_2 formants are lower, making the first two harmonics about equal in amplitude (giving the sound a darker quality), whereas in the popular (also called "legit," or Broadway) style, the formants are higher, so that the first harmonic dominates. In the belt style, F_1 and F_2 are raised so much that the second and third harmonics are 27 dB above the first harmonic, resulting in a loud, bright, "edgy" sound.

Professional male country singers were found by Stone, Cleveland, and Sundberg (1999) to demonstrate characteristics different from those found in classical singers. The inspiratory and expiratory patterns of breathing, as well as the voice source properties and formant frequencies, were found to be quite distinctive in country singers. Whereas the classically trained singers exhibit different characteristics in their singing and speaking voices, country singers showed similar features, including a "speaker's formant" (a prominent F_4) in both. On the other hand the "singer's formant" found in classical singers, was generally missing. This probably reflects the emphasis on text in country singing as well as the fact that they use electronic voice amplification to be heard over an orchestra, and hence do not require a singer's formant to be heard.

Ordinarily male singers do not deliberately tune their vocal tract resonances (formants) to the vocal-fold vibration frequency (as a soprano does). If a male singer unknowingly does so, he may be in for a surprise. The strong acoustic feedback from the vocal tract to the larynx disrupts the national motion of the vocal folds and a *voice break* occurs (Sundberg 1981). The situation is similar to the "wolf note" on a string instrument due to strong interaction between string and body resonances.

By tuning their vocal tract resonances to a *harmonic* of the vocal-fold vibration frequency, however, Tibetan monks produce a very interesting sound. Using this technique while singing a very low note, a single monk can accentuate certain harmonics and produce what sounds like a chord (Smith, Stevens, and Tomlinson 1967). At a higher fundamental frequency, the tuned harmonic may sound more like a whistle. The listener hears a tonal vowel with a pitch corresponding to the fundamental frequency of the voice, accompanied by the whistling of the tuned harmonic. The whistle tracks the fundamental tone.

This technique, sometimes called harmonic singing or throat singing, has been used by David Hykes, who has developed a remarkable control of frequency and sharpness of vocal tract resonance. On a single sustained vowel, he is able to slide the resonance up and down the harmonic series, causing individual harmonics to "pop out" in succession. Alternatively, he can hold a steady whistle frequency while moving the fundamental up and down in pitch. From a perceptual point of view, the difference between emphasizing a single harmonic this way, as shown in Fig. 17.23(a), and the more familiar emphasis of several harmonics by vowel formants, as shown in Fig. 17.23(b), is interesting. The fifth

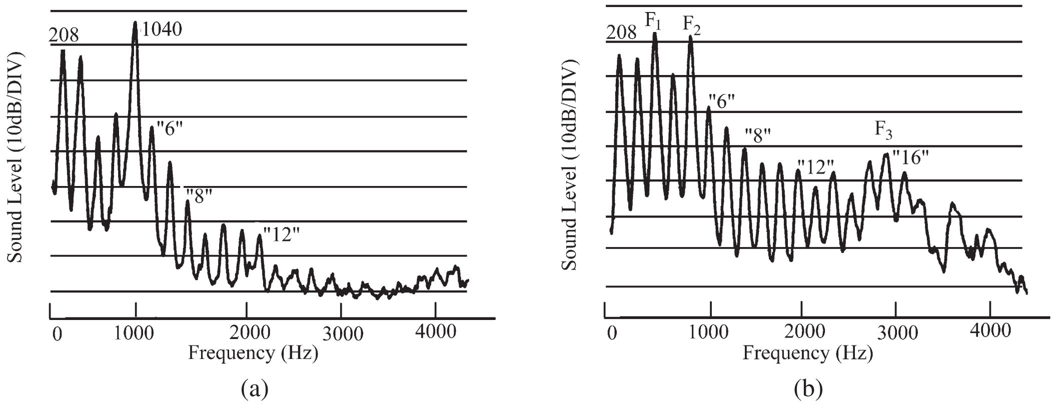


FIGURE 17.23 The spectra of the sung vowel /a/ with fundamental frequency of 208 Hz. (a) A single harmonic (fifth) stands out in harmonic chant. (b) Normal vowel formant emphasizes several harmonics that blend together. (After Hartmann 1997.)

harmonic in Fig. 17.23(a) is considerably stronger than the other harmonics and thus it pops out and is clearly heard (Hartmann 1997).

17.10 ■ SUMMARY

In singing, as in speaking, the vocal folds act as a source of sound, which is filtered by the vocal tract. The resonances (formants) of the vocal tract determine the vowel sounds as well as the timbre of the sung tone. Sung vowels and their formants are slightly different than spoken vowels, one of the most important differences being the appearance of a *singer's formant* around 2500–3000 Hz. One of the important results of vocal training is to learn to lower the larynx and open the pharynx to create this extra formant. Sopranos often sing at pitches above their normal formants, and therefore must "tune" these formants if they are to reinforce the sung notes.

There appear to be two mechanisms for singing: In one, the vocalis muscles are active; in the other, they tend to be passive. They can be referred to as *heavy mechanism* (chest voice) and *light mechanism* (head voice). During normal breathing, about 500 cm^3 of air is moved per breath. In a trained singer, both air flow rate and pressure increase with sound level.

In loud singing, a greater fraction of the total sound energy appears in the higher harmonics, partly due to the higher rate of closure of the glottis. Singers tend to concentrate more energy in the range 2 to 4 kHz in solo singing, whereas they emphasize the fundamental more in choir singing.

REFERENCES AND SUGGESTED READINGS

Appelman, D. R. (1967). *The Science of Vocal Pedagogy*. Bloomington, Ind.: Indiana University Press.

Bartholomew, W. T. (1940). "The Paradox of Voice Teaching," *J. Acoust. Soc. Am.* **11**: 446.

Benade, A. H. (1976). *Fundamentals of Musical Acoustics*. New York: Oxford. (See Chapter 19.)

Bloothooft, G., and R. Plomp (1984, 1985, 1986). "Spectral Analysis of Sung Vowels. I, II, and III," *J. Acoust. Soc. Am*. **75**: 1259; **77**: 1580; **79**: 852.

Bjorklund, A. (1961). "Analyses of Soprano Voices," *J. Acoust. Soc. Am*. **33**: 575.

Bouhuys, A., J. Mead, D. F. Proctor, and K. N. Stevens (1968). "Pressure-Flow Events During Singing," *Annals N. Y. Acad. Sci*. **155**: 165.

Cleveland, T., and J. Sundberg (1983). "Acoustic Analysis of Three Male Voices of Different Quality," in *Proc. Stockholm Music Acoustics Conference (SMAC 83)*, ed. A. Askenfelt, S. Felicetti, E. Jansson, and J. Sundberg. Stockholm: Royal Swedish Acad. of Music.

Haasemann, F., and J. M. Jordan (1991). *Group Vocal Technique*. Chapel Hill, N.C.: Hinshaw Music.

Hartmann, W. H. (1997). *Signals, Sound, and Sensation*. Woodbury, N.Y.: American Inst. Physics. 124.

Hollien, H. (1974). "On Vocal Registers," *J. Phonetics* **2**: 125–143.

Large, J. (1972). "Towards an Integrated Physiologic-Acoustic Theory of Vocal Registers," *Nat. Assn. Teachers of Singing, Bull.* **28**: 18, 30.

Leanderson, R., J. Sundberg, and C. von Euler (1987). "Role of Diaphragmatic Activity During Singing: A Study of Transdiaphragmatic Pressures," *J. Appl. Physiol*. **62**: 259.

Mackworth-Young, G. (1953). *What Happens in Singing*. London: Neame.

Marchesi, M. (1970). *Bel Canto: A Theoretical and Practical Vocal Method*. (Dover reproduction of original undated publication by Enoch & Sons, London.)

McKinney, J. C. (1994). *The Diagnosis and Correction of Vocal Faults*. Nashville, Tenn.: Genevox Music Group.

Miller, D. G. (2000). *Registers in Singing*. Roden, Netherlands: author copyright.

Peterson, G. E., and H. L. Barney (1952). "Control Methods Used in a Study of the Vowels," *J. Acoust. Soc. Am*. **24**: 104.

Rossing, T. D., J. Sundberg, and S. Ternström (1986). "Acoustic Comparison of Voice Use in Solo and Choir Singing," *J. Acoust. Soc. Am.* **79**: 1975.

Rossing, T. D., J. Sundberg, and S. Ternström (1987). "Acoustic Comparison of Soprano Solo and Choir Singing," *J. Acoust. Soc. Am*. **82**: 830.

Schutte, H. K., and D. G. Miller (1993). "Belting and Pop, Non-Classical Approaches to the Female Middle Voice: Some Preliminary Considerations," *J. Voice* **7**: 329–334.

Seymour, J. (1972). "Acoustic Analysis of Singing Voices, Parts I, II, III," *Acustica* **27**: 203, 209, 218.

Shipp, T. (1977). "Vertical Laryngeal Position in Singing," *J. Research in Singing* **1**: 16. (abstract in *J. Acoust. Soc. Am.* **58**: S95.)

Smith, H., K. N. Stevens, and R. S. Tomlinson (1967). "On an Unusual Mode of Chanting by Certain Tibetan Lamas," *J. Acoust. Soc. Am*. **41**: 1262–1264.

Stone, R. E., T. Cleveland, and J. Sundberg (1999). "Formant Frequencies in Country Singers' Speech and Song," *J. Voice* **13**: 161–167.

Strong, W. J., and G. R. Plitnik (1977). *Music, Speech and High Fidelity*, Provo, Utah: Brigham Young University Press. (See Chapter 6B.)

Sundberg, J. (1974). "Articulatory Interpretation of the 'Singing formant,' " *J. Acoust. Soc. Am*. **55**: 838.

Sundberg, J. (1975). "Formant Technique in a Professional Female Singer," *Acustica* **32**: 8.

Sundberg, J. (1977a). "The Acoustics of the Singing Voice," *Sci. Am*. **236**(3).

Sundberg, J. (1977b). "Singing and Timbre," in *Music Room and Acoustics*. Stockholm: Royal Academy of Music.

Sundberg, J. (1978). "Waveform and Spectrum of the Glottal Voice Source," Report STL-QPSR 2–3, 35. Stockholm: Speech Transmission Lab., Royal Institute of Technology.

Sundberg, J. (1981). "Formants and Fundamental Frequency Control in Singing. An Experimental Study of Coupling Between Vocal Tract and Voice Source," *Acustica* **49**: 47–54.

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

Ternström, S., and J. Sundberg (1988). "Intonation Precision of Choir Singers," *J. Acoust. Soc. Am*. **84**: 59.

Titze, I. R. (1973). "The Human Vocal Cords: A Mathematical Model, Part I," *Phonetica* **28**: 129. (See also Part II, *Phonetica* **29**: 1.)

van den Berg, Jw. (1968). "Register Problems," *Ann. New York Academy of Sciences* **155**: 129.

van den Berg, Jw. and W. Vennard (1959). "Toward an Objective Vocabulary for Voice Pedagogy," *N. A. T. S. Bull*., Feb. 1959.

Vennard, W. (1967). *Singing: The Mechanism and the Technic*. New York: Carl Fischer.

GLOSSARY

belting A manner of loud singing used by female popular singers to extend their chest register above its normal range.

chest (modal) voice Mode of singing associated with a heavy mechanism or active vocalis muscles.

cricoarytenoids The muscles of the larynx that help to apply tension to the vocal folds.

cricoid Lower cartilage of the larynx.

cricothyroids The muscles of the larynx that determine the relative position of cricoid and thyroid cartilages and thus affect vocal fold tension.

diaphragm The dome-shaped muscle that forms a floor for the chest cavity.

external (inspiratory) intercostals Intercostal muscles used to breathe air into the lungs.

flow rate The volume of air that flows past a point and is measured per second.

formant A resonance of the vocal tract.

functional reserve capacity (FRC) Volume of air in the lungs at the end of a quiet expiration.

harmonic singing Tuning vocal tract resonances to a harmonic of the vocal-fold vibration frequency to produce single or multiple tones.

head voice Mode of singing associated with a light mechanism, passive vocalis muscle, and elongated, thin vocal folds.

heavy mechanism Mode of vocal fold vibration in which the vocalis muscles are active, and the vocal folds or cords are thick.

intercostal muscles Muscles joining the ribs that are used for breathing.

internal (expiratory) intercostals Intercostal muscles used to breathe air out from the lungs.

larynx The source of sound for speaking or singing.

light mechanism Mode of vocal fold vibration in which the vocalis muscles are relaxed, and the vocal folds elongated and thin.

middle register A combination of light and heavy mechanism that lies between the chest and head registers.

pharynx The lower part of the vocal tract connecting the larynx and the oral cavity.

singer's formant A resonance around 2500 to 3000 Hz in male (and low female) voices that adds brilliance to the tone.

speaker's formant A prominent fourth formant in the speaking and singing voice of country singers.

Strohbass (vocal fry) register Register used for very low bass notes; makes use of a loose glottal closure termed *vocal fry*.

subglottal pressure Amount by which the air pressure in the lungs exceeds atmospheric pressure.

thyroarytenoids (vocalis muscles) The muscles that form part of the vocal folds.

thyroid The upper cartilage of the larynx.

tidal volume The volume of air moved in and out of the lungs during a normal breath.

vital capacity The volume of air that can be moved in and out of the lungs with maximum effort during inhalation and exhalation.

vocal fry Loose glottal closure that allows air to bubble through with a frying-pan sound.

vocal tract The tube connecting the larynx to the mouth consisting of the pharynx and the oral cavity.

vocalis muscle The thyroarytenoid muscle.

whistle register Very high register in the female voice in which the arytenoids form a whistle.

REVIEW QUESTIONS

1. What are the four main parts of the vocal organ?
2. In singing, the vocal tract is tuned by changing its length. (T or F)
3. What vowel has a low F_1 and a high F_2?
4. A soprano can match first formants of what two vowels with her fundamental frequency?
5. What vowel has the highest F_1?
6. What is a *singer's formant*?
7. How is a singer's formant formed?
8. Why does a soprano tune the first formant to match the fundamental frequency?
9. What is meant by subglottal pressure?
10. How is the subglottal pressure measured?
11. How much air is moved in and out of the lungs in normal breathing?

12. Doubling the subglottal pressure produces about how much change in the sound level?

13. What is the approximate range of air flow during singing?

14. Describe the vocal folds when singing in *chest* voice.

15. What is vocal fry? For what type of singing is it employed?

16. How does glottal closure rate change as the phonation level increases?

17. What are two differences in the male singing voice in solo and choir singing?

18. What are two differences in the female singing voice in solo and choir singing?

19. What is *belting*?

20. Describe the technique used by a Tibetan monk to sing what sounds like a chord.

QUESTIONS FOR THOUGHT AND DISCUSSION

1. Try to sing as many notes as possible in both chest and head registers. Can you sing in both registers? How much overlap is there in your voice?

2. Is a stress of 10^6 N/m^2 (See Fig. 17.15) a large stress? What is the breaking stress of a piece of cotton cord? nylon thread?

3. Normal speaking is done in chest voice. Is it possible to speak in a head voice? Is speech intelligibility affected?

4. Place either a cardboard tube, a length of pipe, or your cupped hands around your lips to extend the vocal tract and lower the formant frequencies. Describe the tone produced. What is often called a dark, or covered, tone is produced by extending the vocal tract at the lower end. Is this equivalent to what you have done?

EXERCISES

1. Find the frequencies that correspond to the three singing registers designated in Fig. 17.14.

2. What harmonics of G$_2$ ($f = 98$ Hz) are enhanced by the formants of /i/? of /u/?

3. Compare the first three formant frequencies in Fig. 17.4 to those in Table 17.1 for the sung vowels /u/, /ɑ/, and /i/.

4. Find the lengths of closed pipes that would resonate at 2500 and at 3000 Hz. Are these reasonable lengths for the cavity formed by the (closed) glottis and the (open) pharynx?

5. The power (in watts) used to move air in or out of the lungs is equal to the pressure (in N/m^2) multiplied by the flow rate (in m^3/s). Find the power for:

(a) Quiet breathing ($p = 100$ N/m^2, flow rate $=$ 100 cm^3/s);

(b) Soft singing ($p = 1000$ N/m^2, flow rate $=$ 100 cm^3/s);

(c) Loud singing ($p = 4000$ N/m^2, flow rate $=$ 400 cm^3/s).

6. According to Fig. 17.13, a pressure of 4000 N/m^2 will produce a sound level of about 120 dB.

(a) Find the intensity and sound pressure that correspond to this sound level (see Chapter 6).

(b) Compare the sound pressure at the mouth to the steady subglottal air pressure.

(c) Assuming a mouth opening of 20 cm^2, calculate the total radiated sound power.

(d) What portion of the total power calculated in Exercise 5 is converted into sound? (*Answer:* About 0.1%.)

EXPERIMENTS FOR HOME, LABORATORY, AND CLASSROOM DEMONSTRATION

Home and Classroom Demonstration

1. *Waveforms of vowel sounds* By connecting a microphone to an oscilloscope, display the waveforms for different vowel sounds. A male voice singing "oo" in falsetto at E$_4$ (near the first formant frequency) produces nearly a sine wave with few overtones, for example. Singing "ee" at the same frequency adds small wiggles due to the upper harmonics (mainly the sixth and seventh), which are near the second formant. Finally singing "ah" in a normal chest voice at about

that same pitch can produce a tone in which the second and third harmonics exceed the fundamental because of the high first-formant frequency.

2. *Darkened vowel sound* Produce something akin to dark, or covered, vowel sounds by singing with a short length of tubing surrounding your lips. This lengthens your vocal tract and lowers the formant frequencies. (Of course, in actual covered singing the length of additional tubing is at the other end of the vocal tract, in the vicinity of the vocal folds). This works the best when your mouth opening is made large, as in singing "ah" or "ee." (Strong and Plitnik 1997).

3. *Formant frequencies* Use an FFT analyzer to record spectra of as many vowels as possible. Try to determine the formant frequencies and compare them to the frequencies for spoken vowels in Table 15.3 and in Fig. 17.2. Compare the same vowel spoken and sung.

4. *Identifying vowels at high tessitura* Have a trained soprano sing various vowel sounds near the top of her singing range and try to identify each vowel sound. (It may be difficult to do out of context if the fundamental frequency is higher in frequency that the first formant.)

5. *Scaled formants* Record a series of sung vowels on tape and play them back at different speeds. Note that some of the vowel sounds change to others ("ah" changes to "oh" at half-speed, for example).

6. *Different ways of breathing* Attempt to breathe with your ribs alone, keeping your abdominal wall stationary (try to "freeze" it in the belly-in and belly-out positions as well as in the normal position). Breathe by keeping your ribs as stationary as possible (in both the raised and lowered positions) and moving your abdominal wall to activate your diaphragm.

7. *Vibrato rate* Display the waveform of a good singer on an oscilloscope so that you can see the vibrato rate. Ask the singer to increase and decrease the vibrato rate and amplitude.

8. *Voice source analyzed by inverse filtering* This videotape from the Royal Institute of Technology in Stockholm includes several find demonstration experiments on the voice source in singing as well as speaking.

Laboratory Experiments

The singing voice (Experiment 25 in *Acoustics Laboratory Experiments*)