5

Human Voice

The acoustic communication mode specific to human beings is *speech*. It is the original type of linguistic communication, substantially important for everyday life. Before the emergence of written language, speech was also a way to keep information alive through oral repetition, from one subject to another and from one generation to the next.

The speech communication chain is characterized in Figure 5.1. This model pays attention to the *source* → *channel* → *receiver* structure of communication, but also to the generative process within the speaker and the analysis chain at the receiver. The speaker combines a linguistic message (such as words to be said) with non-linguistic information (rhythm, stress, and intonation of the words) at neural processing levels, constructs motoric control signals for the speech organs, and produces a spoken message conveyed by an acoustic waveform. The communication channel can be any medium, such as a pressure wave in the air, a wired or wireless telephone channel, voice over internet (VoIP), radio, or storage (= temporal transfer) by a recording device. Finally, the receiver is a subject who, in favourable conditions, is able to uncover a meaningful representation from the content of the message. Peripheral hearing is first used to extract a general auditory representation, and then more speech-specific processes are involved to decode the linguistic contents. Sometimes the non-linguistic content can be more important than the linguistic one.

Note that not only the channel but also the source and the receiver may be technical systems based on speech synthesis and recognition (see Sections 16.2 and 16.3), as long as some true speech communication takes place. In this chapter, the focus is first on speech production from both physical and signal processing points of view. The receiving function, implying the perception of speech, as well as the influence of the channel on it, will be discussed in later chapters.

5.1 Speech Production

Speech, as a means of linguistic communication (O'Shaughnessy, 1987), is a uniquely human process that is not found among other species of the world. It is so self-evident an ability that we often don't notice how complex and delicate it is until something goes wrong with it.

Communication Acoustics: An Introduction to Speech, Audio, and Psychoacoustics, First Edition.
Ville Pulkki and Matti Karjalainen.

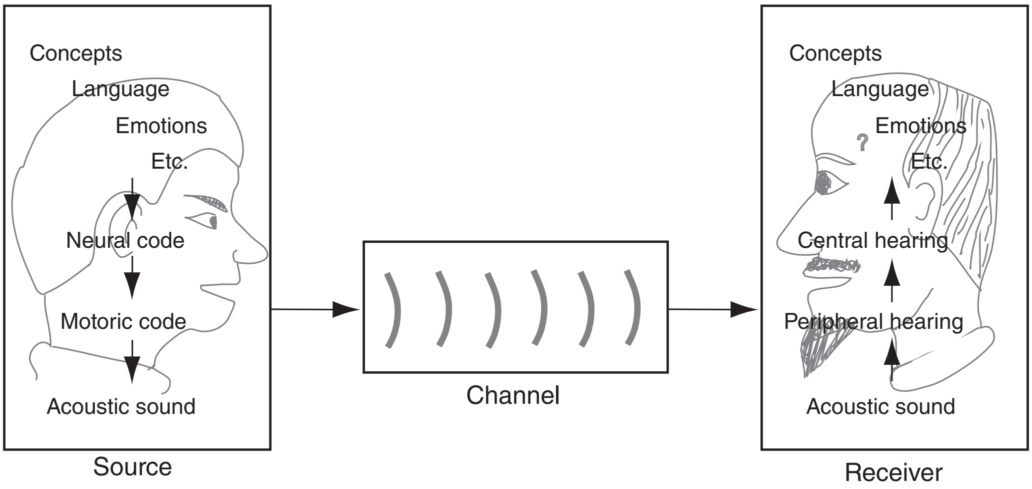

Figure 5.1 Speech communication chain: *source* → *channel* → *receiver*.

The production of a speech signal starts from a need to express the source subject's conceptual, emotional, or other internal representations, as shown in Figure 5.1. A linguistic structure (if it exists) is encoded, adding non-linguistic code (if any). This is a neural process, which results in motoric actions of the speech production organs. Finally, a speech signal is the acoustic output from the speech production organs.

Not much is known about the detailed brain processes at the highest levels of speech formation. What is known much better is the lowest level of a source subject's function, the acoustic theory of speech production (Fant, 1970).

5.1.1 Speech Production Mechanism

The acoustics of speech production have been subjected to extensive research. As early as at the end of the 18th century, it was shown experimentally by Christian Kratzenstein and Wolfgang von Kempelen that the generation of speech sounds can be explained using an acoustic–mechanic model (Flanagan, 1972; Schroeder, 1993).

Figure 5.2 illustrates a cross-sectional view of the human speech organs. The names of most of the important elements and positions used to characterize speech sounds by place of articulation are also given, and they are described below. The speech production mechanism will also be discussed in Section 5.3 from the point of view of modelling. In the next subsections, a brief review of the main functions of the speech organs is presented.

5.1.2 Vocal Folds and Phonation

The source of voiced speech is the airflow that is generated by the vocal folds located in the *larynx* (see Figure 5.2). The *vocal folds* are two horizontal tissues with an elastic membrane coating called mucosa (see Figure 5.3). The positioning and tension of the vocal folds are adjusted by the attached muscles. The orifice between the vocal folds is called the *glottis*. The area of the glottis is at its widest during breathing. During speech production, the glottal area varies temporally between zero and its maximum value.

When air is forced to flow from the lungs through the glottis, the vocal folds start to vibrate, closing and opening the glottis almost periodically. The formation of voiced sounds through

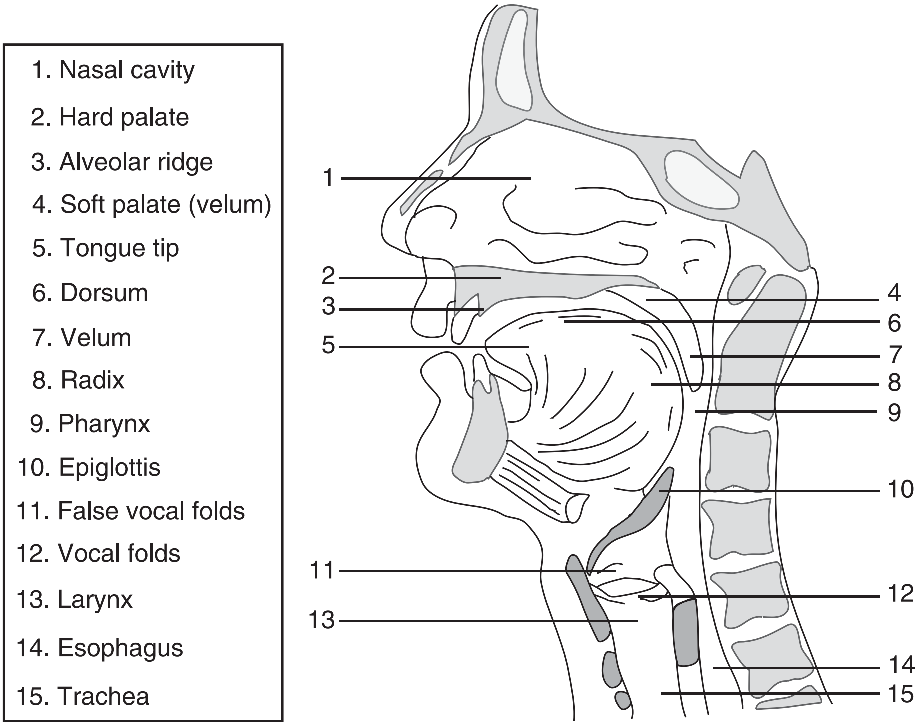

Figure 5.2 Cross-sectional view of the human speech production organs with anatomic names of the organs.

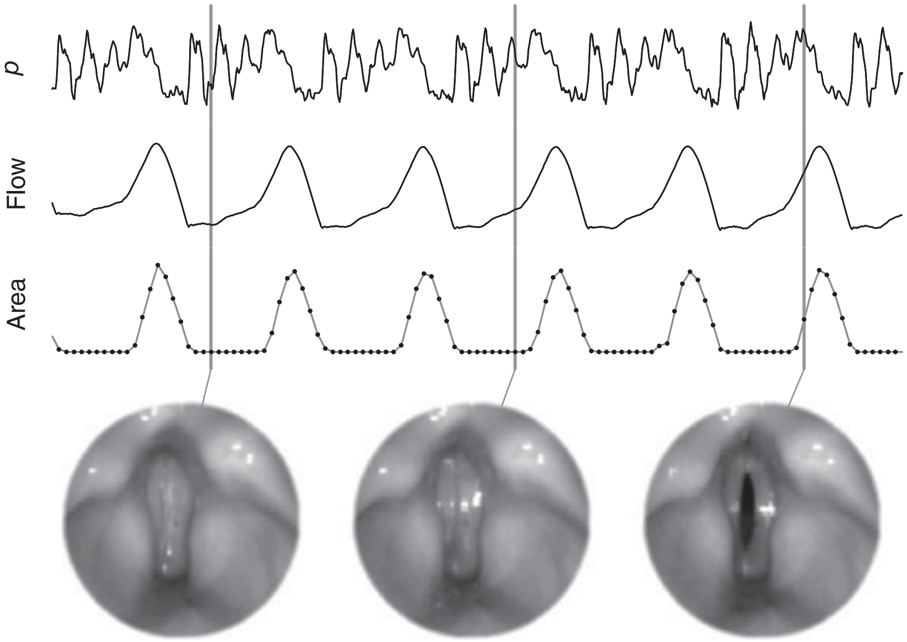

Figure 5.3 The vocal folds during phonation of a vowel imaged with a high-speed camera. The area of the orifice has been measured graphically and is shown as a time-dependent signal. The glottal flow estimated with inverse filtering from the pressure *p* signals is also shown. The vertical lines denote the temporal positions of image captures. Courtesy of Hannu Pulakka.

this vibration is called *phonation*. During speech production, humans are able to control the vibration mode of the vocal folds, hence varying the characteristics of the airflow pulses that are generated at the glottis as an acoustic excitation of voiced speech. Phonation types can be divided into three different categories: breathy, modal (or normal), and pressed (Alku and Vilkman, 1996). In *breathy phonation*, the area of the glottis is a smooth, almost sinusoidal function of time. The corresponding airflow pulse is also smooth in time, with a low-frequency emphasis in its spectrum. Breathy phonation typically has a non-zero leakage of air through the vocal folds over the entire glottal period. In *modal phonation*, the glottal area and the corresponding flow pulse are more asymmetric between the opening and closing of the glottis (see Figure 5.3 for an example of glottal flow during phonation). In *pressed phonation*, the closing phase shortens further in time, hence producing an excitation pulse that has more high frequencies in its spectrum. In addition to these three basic phonation types, humans are also capable of adjusting their glottal function to produce, for example, a *creaky voice* and *whispering*. In the former, the vibration mode of the vocal folds is irregular, involving two different modes that alternate in time, a phenomenon called *diplophonia*, and the duration of the closed phase is typically relatively long. *Whispering* is a speech production mode in which the vocal folds do not vibrate at all, and the sound excitation is produced by an exhaled airflow from the lungs.

The frequency of glottal oscillation is one of the most important acoustic parameters of voiced speech, the fundamental frequency f_0. In conversational speech, the average value of f_0 is about 120 Hz for males and about 200 Hz for females. Humans are, however, capable of producing much larger f_0 values: for a soprano singer, for example, f_0 can go up to 1500 Hz (Klingholz, 1990). The glottal waveform in voiced speech is discussed later in this chapter.

5.1.3 Vocal and Nasal Tract and Articulation

The wave from the glottis during phonation does not radiate directly out – it will propagate through the *pharynx* right above the larynx and then through the *oral cavity* above the tongue and possibly through the *nasal cavity*. These paths are called the *vocal tract* and the *nasal tract*. These acoustic tubes or cavities have an important impact on the final voice that radiates through the lips or the nostrils of a speaker. The process of controlling this acoustic process is called *articulation*.

The vocal tract from the glottis to the lips is about 17 cm in length for an adult male (Rabiner and Schafer, 1978), and about 14.5 cm for females (Klatt and Klatt, 1990). In the middle of the vocal tract, the tip of the soft palate called the *velum* opens or closes the sideway through the nasal tract to the nostrils. In normal production of speech this is open only for nasalized sounds. The effect of these two tracts is to determine the acoustic transfer function in such a way that the spectral properties of the radiating voice can convey speech information through a rich set of distinct sounds. From a signal processing point of view, the vocal and nasal tracts act as filters with controllable resonances that emphasize specific frequencies. These resonances are called *formants*, which are one of the most important cues in speech. The tracts may also create *antiformants*, dips in the spectrum; in other words, zeros in the transfer function.

The articulation position is primarily determined by the tongue, but also by the lips, the jaw, and by the opening or closing of the port to the nasal tract. The tongue is a complex and delicately controlled bundle of muscles that can move forwards and backwards as well as up and down, and its tip has a moving role of its own. Based on the positions of these speech production organs the resulting cross-sectional shape of the vocal tract (and the coupling or

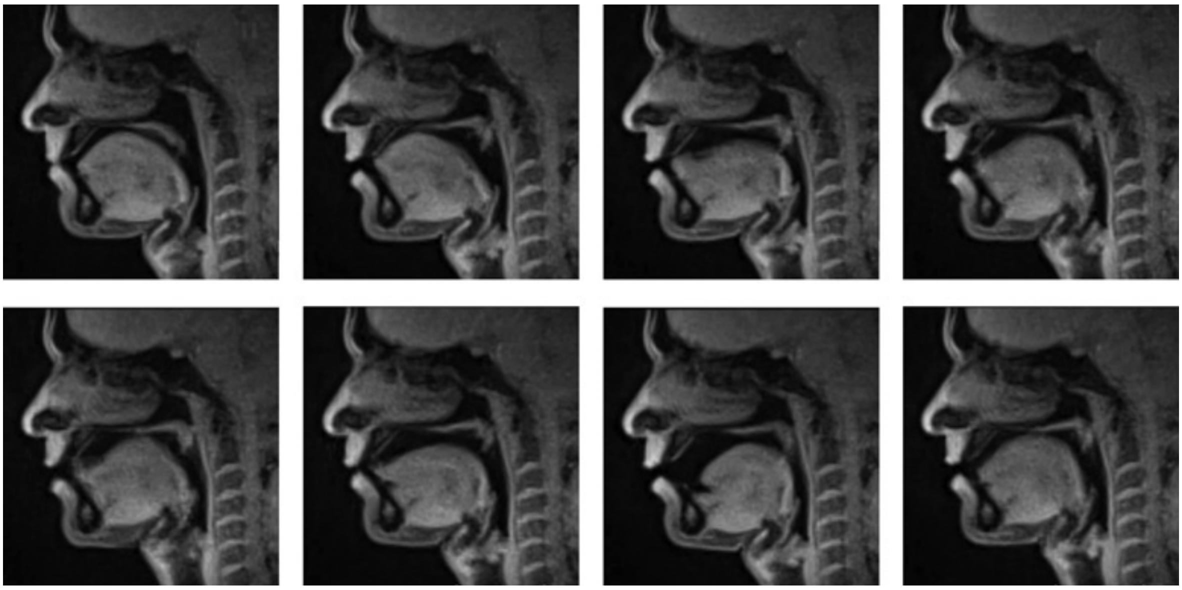

Figure 5.4 Snapshots of the movements of the speech organs during natural speaking. The original video (Niebergall *et al.*, 2011) has been measured using real-time magnetic resonance imaging (Uecker *et al.*, 2010). Licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. Copyright: Max-Planck Institute.

uncoupling of the nasal tract) determine the acoustic transfer function and the filtering effect on the glottal waveform to yield the final *voiced speech* sound. Figure 5.4 shows snapshots of the movements of the speech organs recorded using real-time magnetic resonance imaging. The movements of the tongue between the snapshots are large. In many cases the position of the tongue is close to some part of the roof of the mouth, and notice also the opening and closing of the nasal tract.

In addition to voiced sounds, speech signals contain *unvoiced speech* and mixed sounds with both voiced and unvoiced components. One type of unvoiced sound is created by frication, where the vocal tract has a *constriction*, a nearly closed point, where the airflow velocity is forced to become very high and turbulent, resulting in noise signal generation. Another case is the generation of explosive transient-like sounds where the vocal tract is suddenly opened with a sudden release of pressure. These sounds are called *plosives*, which may be either unvoiced or voiced; consider the difference between /p/ and /b/.

The processes, of human phonation and articulation can be studied by several means. One is the direct observation of the movements of the articulatory organs. With functional magnetic resonance imaging, it is possible to estimate how the cross-sectional shape of the vocal tract behaves during speech, as shown in Figure 5.4. The phonatory function, or the vibration of the vocal folds, is difficult to see due to their position, but, for example, with high-speed digital imaging it is possible to obtain a relatively accurate observation of how they vibrate, as shown in Figure 5.3. Other means of study are, for example, to use a *contact microphone* attached externally to the larynx or an *electroglottograph* to measure the external electrical conductivity changes in the skin close to the larynx caused by the opening and closing of the vocal folds (Colton and Conture, 1990). A further possibility is to analyse the speech waveform, as registered by a high-quality microphone, and to use inverse modelling techniques to derive an estimate of the the glottal waveform (Alku, 2011).

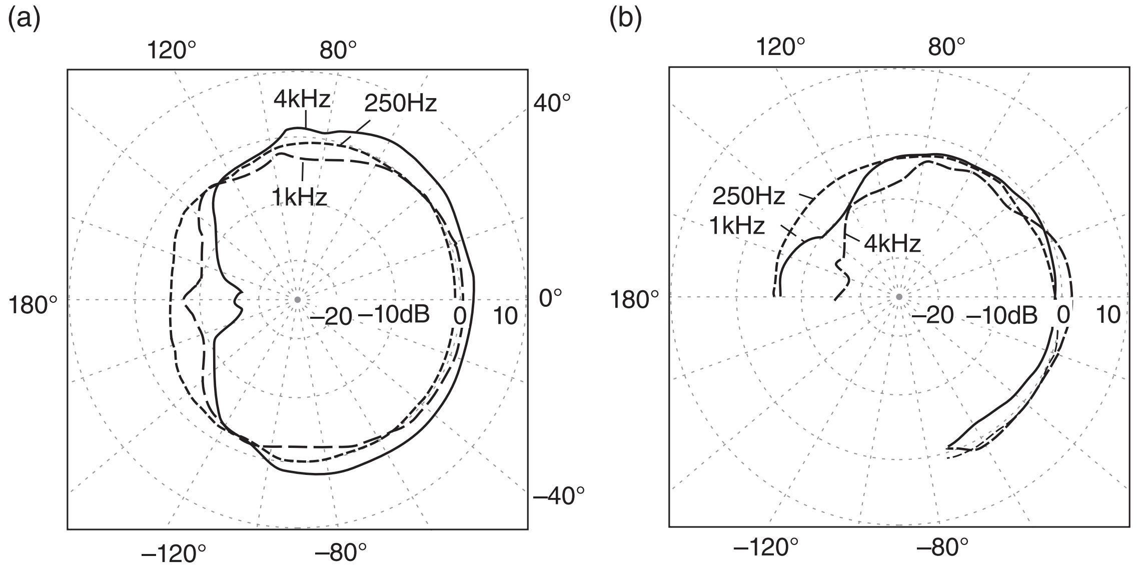

Figure 5.5 The directional patterns of the mouth of a dummy head measured at different frequencies (a) in the horizontal plane and (b) in the median plane.

5.1.4 Lip Radiation Measurements

The radiation of sound from the lips to the environment is also interesting in the context of speech communication (Flanagan, 1960). As the range of frequencies that humans can produce ranges from below 100 Hz to several kilohertz, it can be assumed that the directional pattern of a speaker changes with frequency prominently. At frequencies where the wavelength is much larger than the size of the head, the directional pattern of a human speaker obviously is a unitary sphere. The patterns have been measured with a dummy head with a mouth, and they are illustrated in Figure 5.5. As can be seen, the patterns are relatively constant for all frontal directions. In rear directions, the directions that exceed about 100° from the frontal direction, the sound is attenuated by a few dB for the 250-Hz case and somewhat more at higher frequencies.

5.2 Units and Notation of Speech used in Phonetics

Different languages vary in the form and number of basic units that are used to generate speech messages. Since there is temporal continuity of articulatory motion of the speech production mechanism, each unit is influenced by the preceding and following units. Therefore, speech is never composed of units in isolation but of a continuous flow of signal states. On the other hand, the structure of spoken language cannot be understood without some kind of discrete units that are concatenated into speech signals conveying linguistic information. This unit structure is reflected, more or less regularly, in written language. In languages such as English and French the written form (orthography) and phonetic transcription may often be quite different, while in languages at the other extreme, such as Finnish and Turkish, there is almost a one-to-one correspondence between the written and spoken forms.

Spoken languages exhibit an enormous variation in speech units and their combination. Languages can be grouped into *language families*, such as *Indo-European*, *Dravidian*, and *Uralic* languages (SIL, 2014). In different countries and areas, a language, such as English, may show remarkable variation, so that different *dialects* may vary radically within a single country.

Individual speakers have unique features and their own spoken language undergoes variation and evolution. This makes it difficult to describe speech formally as well as to develop systems and applications for spoken language technology.

Phonetics is the science that has developed ways to analyse and describe speech units and their features and structures composed of speech units (Ashby and Maidment, 2005). Various textual or symbolic notations exist to characterize how the units in speech are pronounced, both for practical guidance in dictionaries and for scientific purposes. An attempt to make a universal notation for any speech sound in any language has been taken in the *International Phonetic Alphabet* (IPA) (International Phonetic Association, 2014). The IPA notation is used in the examples within this book.

Since the symbols used in the IPA – including a large set of symbols for phonetic units, markers for so-called suprasegmental features, additional notational features called diacritics, etc. – are not well suited to computerized representation, different modifications have been developed with simplified notation. For example, IPA-ASCII applies only to ASCII characters. Other such symbol sets are SAMPA (Wells, 1997), where speech units specific to a single language are designed individually, and Speech Synthesis Markup Language (W3C, 2004), which is an XML-based (extensive markup language) markup language for assisting the generation of synthetic speech in Web and other applications.

The units of a spoken language can be defined at different structural and abstraction levels and based on different categorization criteria:

- *Phoneme*: an abstract unit of speech, a class that includes all such instances that don't cause meaning to change if used instead of another member of the same phoneme class. In this sense, the phoneme is a minimal linguistic unit. When a phoneme is referred to in text, the symbols are enclosed in a pair of slashes, as in /i/.
- *Phone* (speech sound): a concrete unit of speech sound, including details of producing the sound. Based on context, the properties of phones vary due to *coarticulation*, which is the continuity of articulatory movements and ease of pronounciation causing the change. Also based on context, a phoneme can be pronounced differently, and the alternative pronunciations for a phoneme are called *allophones*. Sequences of phones constitute syllables that are further composed into words, phrases, and so on. When the text discusses phones, they are referred to by enclosing them within square brackets, as in [i].
- *Diphone*: a unit providing another view to a sequence of phones. While a phone covers the time span from the previous phone boundary to the next, a diphone is understood as the time span from the middle of a phone to the middle of the next one, including the phone transition in the centre of the diphone. A phone boundary is typically the moment of the fastest transition between phones or the beginning of a transition. The duality of phones and diphones reflects the fact that some phones are quite static, carrying information in their central part, while in other cases the transition (diphone nucleus) carries the main phonetic information.
- *Triphone*: a temporal unit that covers two diphones. Longer units include more coarticulation within them, which is favourable in speech technology applications (recognition and synthesis), but on the other hand the combinatory explosion of the number of possible units makes them more problematic than short units. Triphones are utilized in automatic speech recognition, as will be discussed in Section 16.3.
- Phones can often be divided further into shorter *speech segments*, such as the transition phase, steady state phase, silence, burst of noise, and sudden onset.

Languages can typically be transcribed using a few tens of phoneme classes. The basic division of the classes is into the categories of *vowels* and *consonants*. A vowel is a speech sound where the vocal tract is relatively open without major constrictions. Vowels are produced relatively independently of their context. Consonants vary more in the manner of production. These two groups of phonemes are characterized below for the case of the English language.

5.2.1 Vowels

Vowels and consonants are combined in a language in a way that creates its rhythmic pattern, such as syllabic structure, where vowels often are kinds of anchor points to which consonants or consonant clusters are attached. Vowels can be classified using a few articulatory features:

- *Front–back position* of articulation as a measure of tongue position. For example, in the words *bet* and *but*, the only change in articulation is that the position of the tongue moves from the front in *bet* to the back in *but*.
- *Open–close dimension* (openness) of articulation specifies the up–down position of the tongue. The closer the tongue is to the palate, the more the vowel is 'closed'.
- *Rounded–unrounded* shape of the opening of the lips. If a vowel is pronounced first with lips [laterally] wide and then with rounded lips, the perceived vowel changes, as for the vowels in the words *but* and *caught*.

The number of vowel categories varies in different languages, ranging from two to a number higher than ten, depending somewhat on the classification system used (WALS, 2014). Table 5.1 shows 12 vowels in American English and Figure 5.6 shows a mapping of them with the features of articulation. The table does not include *diphthongs*, which are also known as gliding vowels. They refer to two adjacent vowel sounds occurring within the same syllable. Examples of vowel spectra are shown in Figure 5.7 for a male speaker. Notice the clear formant structure in the spectra and also the time-domain signals with a clear repetitive structure.

5.2.2 Consonants

Consonant sounds of a language contrast to vowels by being, in general, less intense, more context-dependent (more coarticulated), and produced with a more constricted vocal tract or by special sound-generation mechanisms. The manner of articulation is the configuration and interaction of the articulators when making a speech sound. The articulators include speech organs, such as the tongue, lips, and palate. The consonants in the English language can

Table 5.1 The most common vowels in American English.

IPA symbol	Examples	IPA symbol	Examples
i	beat	ɪ	bit (busy)
e	bait	ɛ	bet
æ	bat	ɑ	cot
ɔ	caught	o	coat
ʊ	book	u	boot
ʌ	but	ə	*a*bout

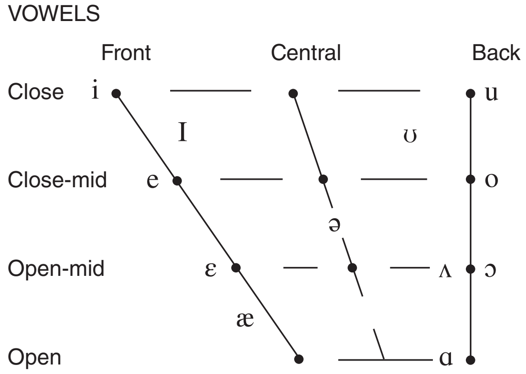

Figure 5.6 IPA chart of American English vowels (not including diphthongs). The front–central–back dimension denotes the position of the tongue, and the close–open dimension indicates how close the tongue is to the roof of the mouth. The symbols on the left of the vertical line have the mouth in an unrounded opening of the lips and those on the right in a rounded one. Adapted from http://www.langsci.ucl.ac.uk/ipa/ipachart.html, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright ©2005 International Phonetic Association.

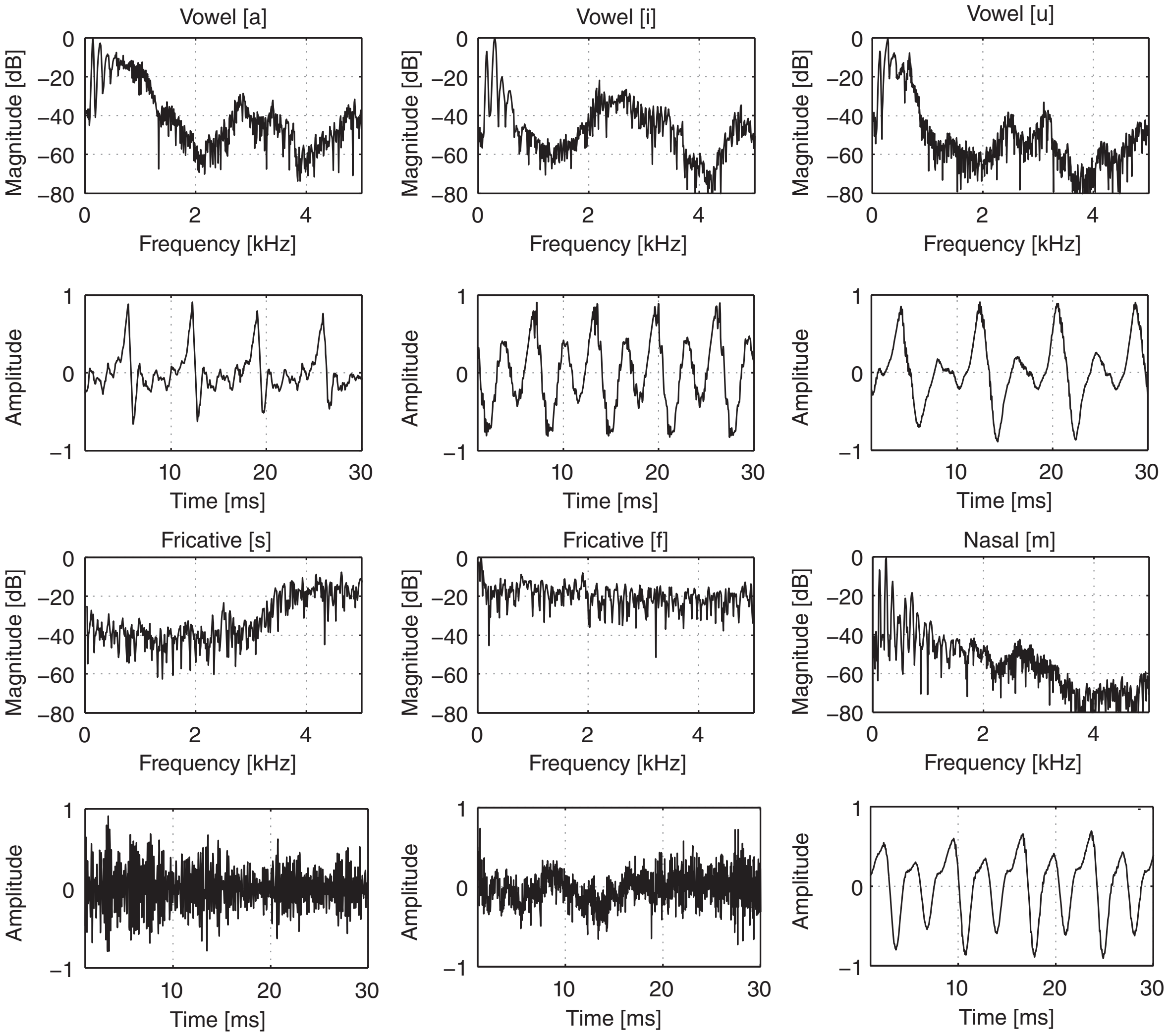

Figure 5.7 Spectra and waveforms of three vowels and three consonants.

be classified, based on the manner of articulation, into the following classes, which are not necessarily orthogonal to each other:

1. *Plosives* /p, b, t, d, k, g/: The vocal tract and nasal tract are completely closed causing a short silence /p, t, k/, or almost completely closed causing a faint voiced sound /b, d, g/. When the tract is opened, a short impulsive sound, or burst, is generated. The impulsive sound can be either noise-like or an exponentially decaying harmonic tone complex. The term *stop* consonant is often used interchangeably with *plosive consonant*.
2. *Nasals* /n, m, ŋ/: The vocal tract is closed, but the soft palate opens the nasal tract. The acoustic characteristics of both nasal and vocal tracts affect the sound. In most languages, including English, nasals are voiced.
3. *Trills* /r/: In trills the articulator vibrates rapidly (with a frequency of about 20–25 Hz) against the place of articulation. The only trill in English is /r/ (such as in *roar*), where the tongue vibrates against the alveolar ridge for about two to three vibrations. The known trills are voiced, and the vibrations cause an effect resembling amplitude modulation.
4. *Fricatives* /f, v, θ, ð, s, z, ʃ, ʒ, h/: The vocal tract is almost closed, because the articulator is brought close to the position of articulation. The narrow passage causes turbulant airflow, which is called frication. The frication causes a noisy sound, which is modified by the resonances of the vocal or nasal tract. The fricatives may be voiced or unvoiced.
5. *Approximants* /j, w, ɹ/: The approximants are similar to fricatives, but the articulators do not come close enough to generate frication.
6. *Laterals*: In lateral consonants the airstream proceeds along the sides of the tongue. English has only one lateral, /l/ (lateral approximant), where the tongue comes close to the alveolar ridge with voiced excitation.

In addition to the manner of articulation, the consonants can also be classified according to the position of articulation. As already mentioned, during the articulation the vocal tract is either totally or partially closed by the articulator, and the consonants articulated at the same position belong to the same class. The most common consonants in American English are shown in Table 5.2 with both their manner and position of articulation. The positions of articulation are shown in Figure 5.8. Examples of consonant spectra are shown in Figure 5.7 for a male speaker. Notice the clear patterns in the spectra and also the time-domain signals with noisy characteristics for unvoiced [s] and [f], and the repetitive structure with voiced [m].

5.2.3 Prosody and Suprasegmental Features

Speech signals exhibit features that are not strictly bound to the segmental structure. Such features are called *suprasegmental* or *prosodic* features. The concept *prosody* refers to the behaviour of prosodic features in general. The set of prosodic features comprises the following:

- *Intonation*. Many phonemes have voiced excitation, which always has the fundamental frequency f_0. The speakers actively vary f_0, or pitch, to convey the utterance information. These variations are called intonation. It can be used for a range of purposes, depending on language, such as indicating emotions and attitudes or signalling the difference between statements and questions. Intonation is also used to focus attention on important words of the spoken message and to help regulate conversational interaction. In tone languages, such as

Table 5.2 Table of common American English consonants.

IPA symbol	Example	Manner	Voiced	Position
j	you	approximant	yes	palatal
w	wow	approximant	yes	labial–velar
ɹ	red (American dialect)	approximant	yes	alveolar
l	lull	approximant	yes	lateral
r	roar	trill	yes	alveolar
m	my	nasal	yes	bilabial
n	none	nasal	yes	alveolar
ŋ	hang	nasal	yes	velar
f	fine	fricative	no	labiodental
v	valve	fricative	yes	labiodental
θ	thigh	fricative	no	dental
ð	though	fricative	yes	dental
s	say	fricative	no	alveolar
z	zoo	fricative	yes	alveolar
ʃ	show	fricative	no	postalveolar
ʒ	measure	fricative	yes	postalveolar
h	how	fricative	no	glottal
p	pot	plosive	no	labial
b	bib	plosive	yes	labial
t	tot	plosive	no	alveolar
d	did	plosive	yes	alveolar
k	kick	plosive	no	velar
g	gig	plosive	yes	velar
tʃ	church	affricate	no	alveopalatal
dʃ	judge	affricate	yes	alveopalatal

Chinese, pitch changes associated with syllables carry prominent linguistic information. This means that changing the pitch changes the phone, although other acoustic features remain constant.

- *Stress*. Some words, or syllables of words, are often given greater emphasis in spoken language. Stress can be implemented in different ways, depending on the speaker and on the language. The stressed syllable may have a higher or lower pitch than non-stressed syllables, which is called *pitch accent*. Other features for stress include *dynamic accent*, where the level of sound is raised during the stress, *qualitative accent* meaning differences in place or manner of articulation, and *quantitative accent*, where the length of a syllable is increased.
- *Rhythm* and *timing*. A differentiating factor between spoken languages is how words and phonemes are divided in time. Some languages allocate similar time for each syllable, while others divide the time based on other linguistic units. Quantitative stresses also affect the rhythm. On the other hand, some languages differentiate between short and long sounds, and these are called *quantity languages*. For example, the Finnish words /tuli/ and /tuuli/ (fire and wind) differ primarily in the duration of the stressed vowel.

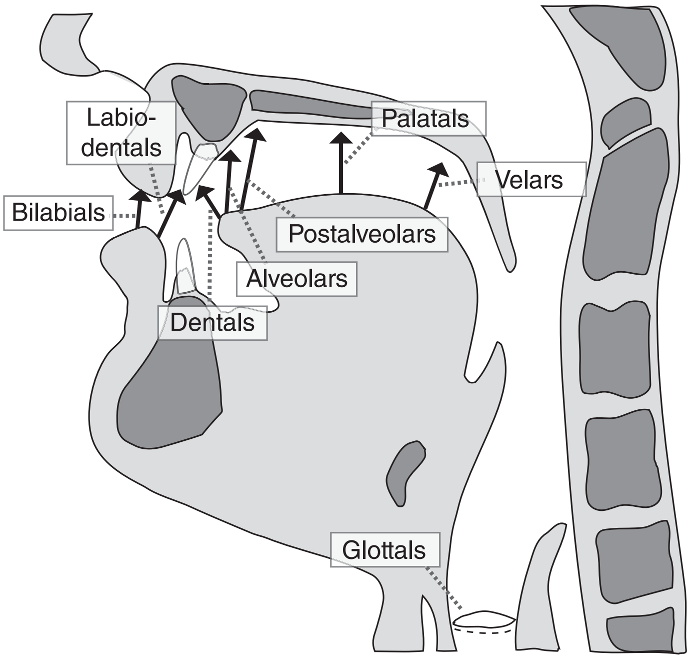

Figure 5.8 The positions of articulation of consonants and the corresponding movement of articulators.

5.3 Modelling of Speech Production

The acoustic production of speech signals can be approximated and modelled mathematically fairly easily (Fant, 1970). Based on such modelling, it is possible to derive signal processing algorithms that are found to be very useful in speech technology, like in speech synthesis and coding (Rabiner and Schafer, 1978). It also helps us to understand speech mechanisms.

Figure 5.9 illustrates an ‘engineering-oriented’ mechanical–acoustic model of speech production. The lungs are seen as a ‘pressure compressor’ that is the driving force for the glottal mass–spring oscillator or for frication and transient generation in vocal tract constrictions. The vocal and nasal tracts form an acoustic resonator or filter system, a kind of complex modulator or encoder of speech sounds. From the lips and the nostrils the signal may radiate and propagate to a listener or a microphone in a technical speech communication system.

The acoustically abstracted presentation in Figure 5.9 can be reduced further into the *circuit analogy* of Figure 5.10 that consists of three cascaded (= coupled in series) subsystems: the *excitation* (generator), a *transmission line* (a *two-port*), and a *radiation impedance* (acoustic radiation load). Acoustic variables that describe the functioning of the system may be the *sound pressure p* and the *volume velocity u* at each point of the system. In Figure 5.10, the variables are p_s (pressure from the lungs), u_g (volume velocity at the glottis), as well as u_m and u_n (volume velocities at the lips and the nostrils, respectively). If the behaviour within the vocal or nasal tract is of interest, these blocks must be described as distributed wave propagation subsystems.

In Figure 5.11, speech production is finally reduced into a *signal model*, the so-called *source–filter model*. In voiced signals (in phonation), the source is the glottal oscillation and in unvoiced sounds the source is frication noise created in a constriction or a transient sound generated by rapid release of pressure when a vocal tract closure is opened.

When the acoustic variables in Figure 5.9 are selected to be the sound pressure and the volume velocity, the natural choices for their electric analogies in this case are voltage and

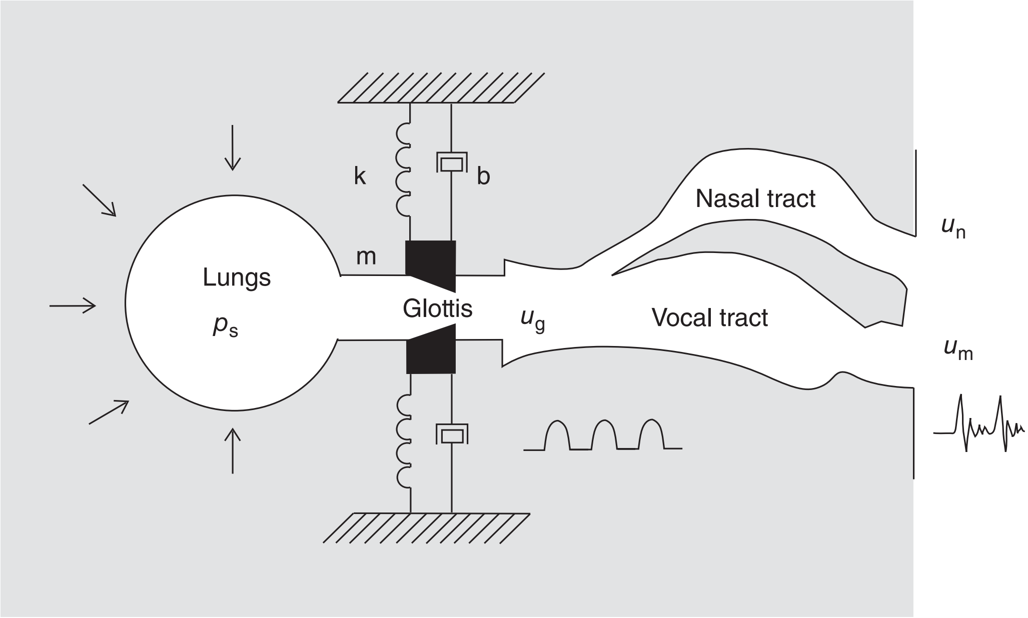

Figure 5.9 The mechanical–acoustic model of speech production.

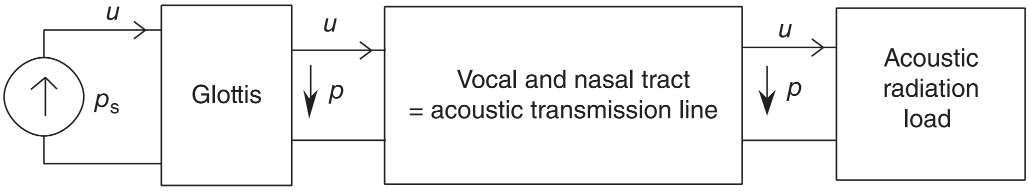

Figure 5.10 A circuit model of speech production: p denotes pressure (p_S is the source pressure from the lungs) and u is volume velocity.

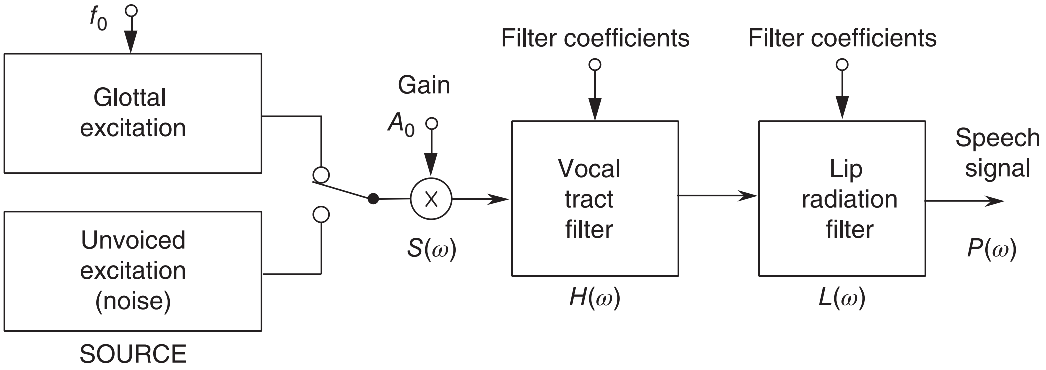

Figure 5.11 A signal model (source–filter model) of speech production.

current, respectively. In signal modelling (Figure 5.11), we do away with this dual variable modelling and consider only a single transfer function, for example, with the Fourier transform formulation:

$$P(\omega) = S(\omega) \cdot H(\omega) \cdot L(\omega), \tag{5.1}$$

where $P(\omega)$ is the pressure of the radiating waveform, $S(\omega)$ is the glottal excitation (for example, volume velocity), $H(\omega)$ is the corresponding transfer function of the vocal/nasal tract system, and $L(\omega)$ is the transfer function of mouth/nose radiation. Note that the input signal is the volume velocity and the output signal is the radiated pressure signal.

A similar formulation is possible when the excitation is, for example, frication noise from a constriction in the vocal tract, and the transfer function is formed by the proper part of the vocal tract and radiation acoustics. In both cases, the simple model of Figure 5.11 is applicable if the excitation source is selected according to the type of excitation. For voiced fricatives a mixed-source model is needed where voicing and noise are mixed properly.

Now the acoustic system is reduced to a simple, mathematically formulated transfer function. The fact that such models are relatively simple makes them practical in terms of realizing synthetic speech generation – speech synthesizers – in the form of electronic circuits or digital signal processing algorithms. In the following, the models in Figures 5.10 and 5.11 are discussed in more detail.

5.3.1 Glottal Modelling

Glottal oscillation, the almost periodic opening and closing of the vocal folds, is the main excitation in voiced sounds. Glottal functioning has been modelled mathematically in several ways, for example as a simple mass–spring system as in Figure 5.9 (Flanagan, 1972), as a spatially distributed mass–spring system (surface-wave transmission line), or just as experimental functions of time for volume velocity (Fant *et al.*, 1985). Figure 5.3 on page 81 plots the volume velocity at the glottis during static phonation of a vowel. The glottal waveform typically resembles a half-wave rectified sinusoidal signal with asymmetry, so that the opening phase is relatively slow and the closing phase is more abrupt, as seen in the figure.

Figure 5.3 also plots the pressure wave at the lips for the vowel that is generated. Notice that the main excitation is due to the closing event of the glottis, since this contains most of the excitatory energy at the formant frequencies. Recognize also that the glottal function is a non-LTI (not linear and time-invariant) system since its parameters – for example, the acoustic impedance in the glottal orifice – vary with glottal closing and opening. Thus, Equation (5.1) is not valid in a strict sense. Since the interaction of the vocal tract and the vocal folds is relatively weak, such a simplified and linearized model is useful as a first approximation.

There is no phonation in unvoiced sounds; that is, the vocal folds don't vibrate. Frication or pressure release bursts are the excitation signals in such speech sounds.

5.3.2 Vocal Tract Modelling

The vocal tract acts as an acoustic transmission line where the glottal excitation propagates to the lips. While the uvular link to the nasal cavity is open, the coupling of the nasal tract creates nasal or nasalized sounds. The shape of the tract is determined by the position of the tongue and its tip, the opening of the jaw, and the shape of the lip opening. From an acoustic point of view, the transmission properties of the vocal tract are due to the cross-sectional area function of the tract $A(x)$, where x is the position in the glottis–lips dimension. According to Equation (2.23), the acoustic impedance at each point is $Z_{\mathrm{a}}(x) = \rho c / A(x)$; that is, it is inversely proportional to the area.

Figure 2.10 and Equation (2.22) on page 27 show us that a plane wave travels in the tube without modification as long as the area is constant, but at any discontinuity of $A(x)$ a part of the wave is reflected back and the rest propagates onwards. The simplest tract shape is one

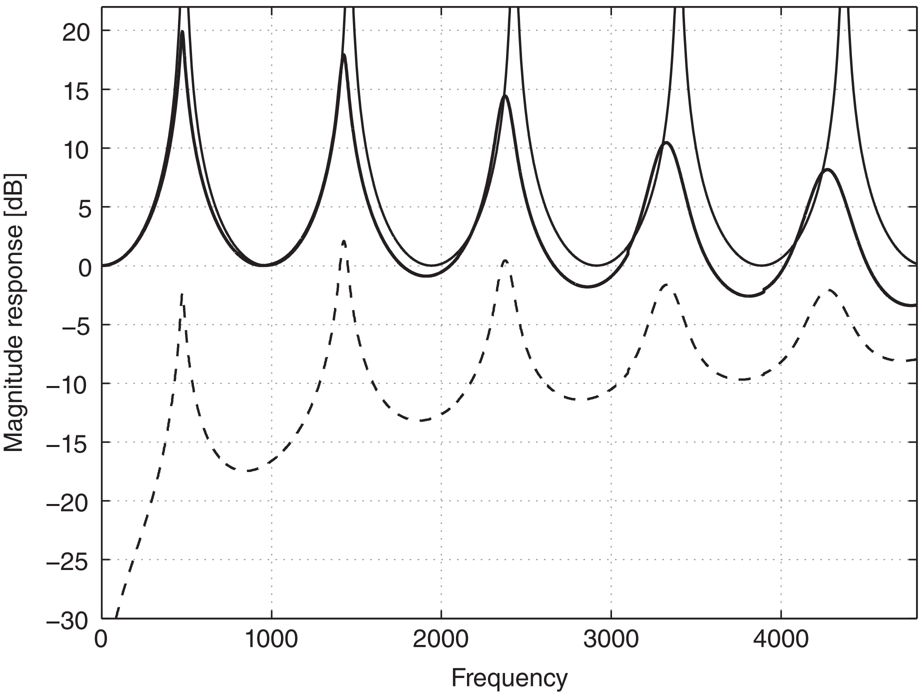

Figure 5.12 Solid lines: the transfer function of the neutral vocal tract (magnitude response of the volume velocity transfer from the glottis to the lips). An idealized, lossless case is plotted by the thin line, where the lines continue to positive infinity at resonance frequencies. The thick line represents a simulation with losses due to yielding walls, friction, and thermal loss. The dashed line shows the transfer function from the glottal velocity to pressure after lip radiation. Adapted from Rabiner and Schafer (1978).

where $A(x)$ is constant from the glottis to the lips. This is approximately true for the *neutral vowel*, something like /ə/. While the glottal termination has a high acoustic impedance and the lip opening a low impedance, the acoustic system for the neutral vowel corresponds to the quarter-wavelength resonator characterized in Figure 2.12b on page 29. The magnitude response $H(\omega)$ of an ideal lossless tract, that is, the ratio of the glottal volume velocity $U_\mathrm{g}(\omega)$ to the lip volume velocity $U_\mathrm{l}(\omega)$, is of the form

$$H(\omega) = \frac{U_\mathrm{l}(\omega)}{U_\mathrm{g}(\omega)} = \frac{1}{\cos(\omega l/\mathrm{c})}, \tag{5.2}$$

where l is the length of the tract and c is the speed of sound. $H(\omega)$ is characterized in Figure 5.12 whereby the formant resonances for a 17-cm long vocal tract (typical for an adult male) are located at frequencies

$$f_n = 500 \cdot (2n - 1), \qquad n = 1, 2, \ldots \tag{5.3}$$

that is, at $f_1 = 500\,\mathrm{Hz}, f_2 = 1500\,\mathrm{Hz}, f_3 = 2500\,\mathrm{Hz}$, and so on. The formant resonances of the ideal tract in Equation (5.2) have infinite amplitude (the thin line in Figure 5.12), while with frequency-dependent losses due to both vocal tract and lip radiation, the formants follow the bold line, as discussed by (Rabiner and Schafer, 1978). Finally, the transfer function from the glottal flow to the pressure signal after lip radiation is shown in the same figure with a dashed line.

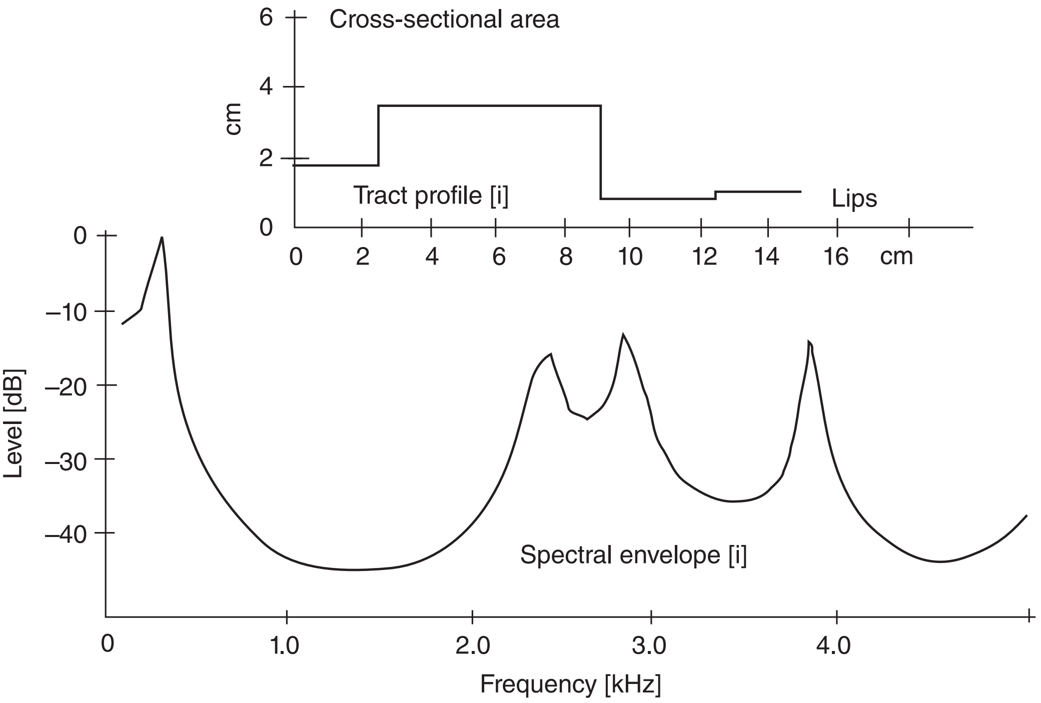

Figure 5.13 An approximation of the geometry of the vocal tract during the phonation of /i/ made with four tubular sections. The lower diagram shows the simulated magnitude spectrum of the response of the approximated tract.

For real speech signals the tract can take many shapes, and this results in a multitude of vocal tract transfer functions. For each shape the positions of the formant resonances take specific positions. The spectral shape, and especially the frequencies of the lowest formants, then carry the information for the identity of the speech sound to the listener. As an example, the shape of the vocal tract when pronouncing the vowel [i] is modelled. The tract is approximated with four concatenated cylinders of different lengths and diameters, as shown in Figure 5.13. In reality, the shape of the tract is, naturally, continuous. However, this approximation already gives quite a realistic result. The modelling takes into account the spectrum of the glottis excitation, the magnitude response of the model of the tract, and also the radiation from the lips. The magnitude spectrum of the resulting transfer function is shown in the figure, which matches nicely with the magnitude spectrum of [i] shown in Figure 5.7.

5.3.3 Articulatory Synthesis

Since the vocal tract can be considered an acoustic transmission line, it can be simulated computationally based on the idea of Figure 5.11 and acoustically on that of Figure 5.9. The vocal tract (and the nasal tract for nasalized sounds) can be approximated by a cascade of transmission line segments, as shown in Figure 5.13. This is called the *Kelly–Lochbaum model* (KL-model) (Kelly and Lochbaum, 1962), where the vocal tract cross-sectional profile $A(n)$ is mapped onto reflection coefficients $k(n)$:

$$k(n) = \frac{A(n+1) - A(n)}{A(n+1) + A(n)} \tag{5.4}$$

The principal advantage of articulatory synthesis is that the dynamics of the speech production systems are included in a natural way. There are, however, practical problems associated with getting data on vocal tract parameters from real speakers and with getting the complex, non-linear behaviour of the glottal oscillation and its coupling to the vocal tract acoustics. Articulatory modelling and synthesis remains an important research topic for understanding the underlying phenomena, but it may not be the most practical way to realize practical speech synthesizers.

5.3.4 Formant Synthesis

The source–filter model of Figure 5.11 can be used to approximate the generation of any speech sound. Thus, there is no necessity to model the details of the vocal tract, instead a black-box model realizing the spectral structure and related formant resonances of speech sounds is applicable. Each formant is created by a second-order filter section as a part of the overall synthesis filter structure.

The basic configuration of formant speech synthesis is the parallel filter structure shown in Figure 5.14. A finite number of second-order low-pass filters (typically 4–5) are in parallel in the structure, and it can be used to synthesize any phones of a language. A rich set of control parameters is necessary. Each formant has to be controlled by the frequency, the amplitude level, and the Q-value (sharpness of the resonance). Voiced and unvoiced excitation sources are needed, as in Figure 5.11.

Although the method is general in principle and can be used to produce any speech sound, it is not widely used in speech synthesis. The task of generating the control information for the synthesizer is demanding, and other solutions have replaced the technique in many applications. The synthesis methods which are in wide use will be discussed in Section 16.2.

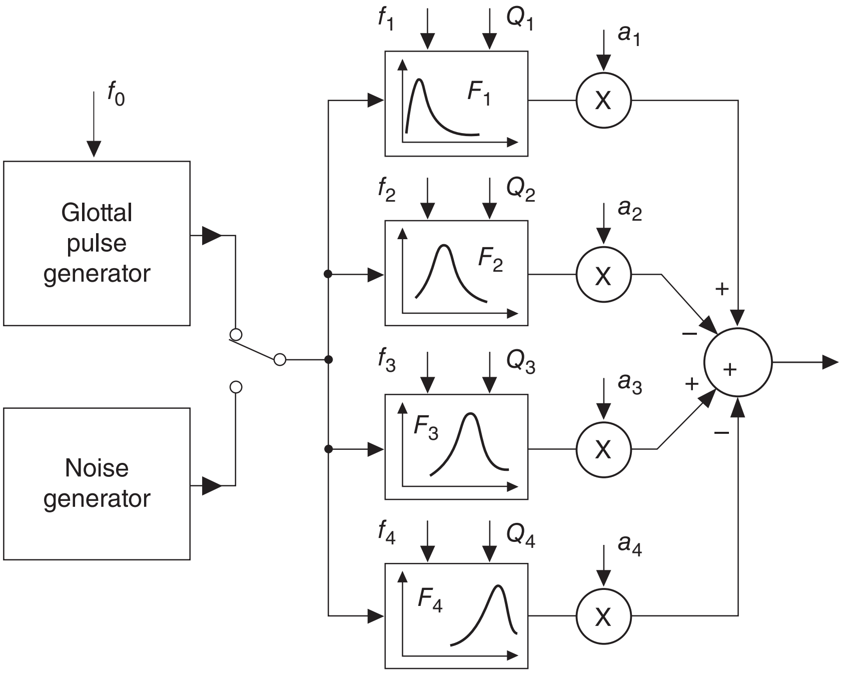

Figure 5.14 Formant synthesis with a parallel formant synthesizer. F_1-F_4 are second-order resonators for individual formants, f_0 is the fundamental frequency of voiced excitation, and f_1-f_4, Q_1-Q_4, and a_1-a_4 show the frequencies, Q-factors, and gains for each formant, respectively.

5.4 Singing Voice

There are a vast number of singing styles, which very creatively use the speech organs to generate different sounds. The styles vary from opera to reggae, from rap to lied, and from joik to chanson. There is no clear border between speech and singing, but some general remarks can be made about the differences (Rossing *et al.*, 2001).

- Singing has, in most cases, a controlled and easily definable pitch, which shifts often in rapid shifts from one frequency to another according to the melody and rhythm of the musical piece. In speech prosody, the changes in pitch can also be rapid, but the pitch typically changes smoothly all the time, while in singing the pitch is often relatively constant during each note. Quite often, each phone corresponds to a single note, although it can be divided to last for several notes.
- The spectrum of the vowels is quite often different from normal speech. The first harmonic or some other harmonics may be louder in singing than in speech. This effect is accomplished with such changes as lowering the larynx, opening the jaws more, advancing the tip of the tongue, and protruding the lips (Rossing *et al.*, 2001).
- Vibrato is used in singing much more often than in speech. When singing with vibrato, the pitch and/or the level is modulated at a rate close to 7 Hz (Sundberg, 1977). The depth and modulation frequency depend on the artist and on the musical style.
- The fundamental frequency may range more widely in singing than in speech. When the range of pitches in speech prosody is about one octave, the range may extend two octaves in singing (Rossing *et al.*, 2001).

Classical singing has perhaps been studied the most of all singing styles (Rossing *et al.*, 2001). The style has evolved from before the use of electrical reinforcement of sound, and the primary task of the singer is to produce a esthetic singing that can be heard when the accompanying symphony orchestra plays loudly. Singers are able to reinforce some of the formants of the vocal tracts into the *singer's formant*, which boosts by more than 20 dB the frequencies near 2–3 kHz (Sundberg, 1977). This makes the singing audible in such a scenario.

Another commonly known phenomenon is that trained soprano singers may produce such high pitches that the harmonics are sparse in frequency, so that the formant structure is not present in the signal. This makes high-pitched vocals unintelligible. Some female singers are also able to change the frequencies of the formants to match those of the harmonics of the note (Sundberg, 1977). This does not make the vocals more intelligible, but it raises the total level of sound, which might be desirable, for example, in opera singing.

An extreme example of the human ability to use speech organs artistically is *overtone singing* (a.k.a. throat singing), where the vocal folds produce a low-pitched sound with relatively high-level harmonics. The vocal tract is then used to produce very sharp resonances, which selectively amplify some of the harmonics (Bloothooft *et al.*, 1992). By changing the resonance frequencies, the singer creates melodies.

Summary

This chapter has been devoted to human voice as a communication signal; how it is produced by humans and technical systems. The modelling aspect is used here to characterize the general principles that are, in most cases, common to human voice production, the acoustic generation of sound, and to the electronic and computer means of generating voice in technical devices.

Further Reading

These topics are well studied in the existing research, and literature on them is rich, so in this chapter only the surface has been scratched, providing a tutorial-like presentation and a handbook-style overview. Further reading can be found easily in the following references. Recommended references for further reading on speech communication and technology are, for example, Deller *et al.* (2000); Fant (1970); Flanagan (1972); Markel and Gray (1976); O'Shaughnessy (1987); Rabiner and Schafer (1978); Santen *et al.* (1997); Stevens (1998) and Titze (1994). A more concise introduction to phonetics can be found in Ashby and Maidment (2005).

References

Alku, P. (2011) Glottal inverse filtering analysis of human voice production: a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, **36**(5), 623–650.

Alku, P. and Vilkman, E. (1996) A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. *Folia phoniatrica et logopaedica*, **48**(5), 240–254.

Ashby, M. and Maidment, J. (2005) *Introducing Phonetic Science*. Cambridge University Press.

Bloothooft, G., Bringmann, E., Van Cappellen, M., Van Luipen, J.B., and Thomassen, K.P. (1992) Acoustics and perception of overtone singing. *J. Acoust. Soc. Am.*, **92**, 1827–1836.

Colton, R.H. and Conture, E.G. (1990) Problems and pitfalls of electroglottography. *J. Voice*, **4**(1), 10–24.

Deller, J.R., Proakis, J.G., and Hansen, J.H. (2000) *Discrete-Time Processing of Speech Signals*. IEEE.

Fant, G. (1970) *Acoustic Theory of Speech Production*. Mouton de Gruyter.

Fant, G., Liljencrants, J., and Lin, Q.C. (1985) A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm*, **4**, 1–13.

Flanagan, J.L. (1960) Analog measurements of sound radiation from the mouth. *J. Acoust. Soc. Am.*, **32**, 1613–1620.

Flanagan, J.L. (1972) *Speech Analysis, Synthesis and Perception*. Springer.

International Phonetic Association (2014) Homepage. `http://www.arts.gla.ac.uk/IPA/ipa.html`.

Kelly, J.L. and Lochbaum, C.C. (1962) Speech synthesis *Proc. Fourth Int. Congr. Acoust.*

Klatt, D.H. and Klatt, L.C. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, **87**, 820–857.

Klingholz, F. (1990) Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels. *J. Acoust. Soc. Am.*, **87**, 2218–2224.

Markel, J.D. and Gray, A.H. (1976) *Linear Prediction of Speech Signals*. Springer.

Niebergall, A., Uecker, M., Zhang, S., Voit, D., Merboldt, K.D., and Frahm, J. (2011) Real-time MRI–speaking `http://commons.wikimedia.org/wiki/File:Real-time_MRI_-_Speaking_(English).ogv`.

O'Shaughnessy, D. (1987) *Speech Communication–Human and Machine*. Addison-Wesley.

Rabiner, L.R. and Schafer, R.W. (1978) *Digital Processing of Speech Signals*. Prentice-Hall.

Rossing, T.D., Moore, F.R., and Wheeler, P.A. (2001) *The Science of Sound*, 3rd edn. Addison-Wesley.

Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (eds) (1997) *Progress in Speech Synthesis.* Springer.

Schroeder, M.R. (1993) A brief history of synthetic speech. *Speech Communication*, **13**(1), 231–237.

SIL, (2014) Ethnologue – languages of the world `http://www.ethnologue.com`.

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

Sundberg, J. (1977) *The Acoustics of the Singing Voice*. Scientific American.

Titze, I.R. (1994) *Principles of Voice Production*. Prentice-Hall.

Uecker, M., Zhang, S., Voit, D., Karaus, A., Merboldt, K.D., and Frahm, J. (2010) Real-time MRI at a resolution of 20 ms. *NMR in Biomedicine*, **23**(8), 986–994.

W3C, (2004) Speech synthesis markup language (SSML) `http://www.w3.org/TR/speech-synthesis/`.

WALS, (2014) Ethnologue – Languages of the World. Technical report, The World Atlas of Language Structures Online. `http://wals.info/chapter/2`.

Wells, J. (1997) SAMPA computer readable phonetic alphabet. *Handbook of Standards and Resources for Spoken Language Systems*.