Production and Classification of Speech Sounds

3.1 Introduction

A simplified view of speech production is given in Figure 3.1, where the speech organs are divided into three main groups: the lungs, larynx, and vocal tract. The lungs act as a power supply and provide airflow to the larynx stage of the speech production mechanism. The larynx modulates airflow from the lungs and provides either a periodic puff-like or a noisy airflow source to the third organ group, the vocal tract. The vocal tract consists of oral, nasal, and pharynx cavities, giving the modulated airflow its “color” by spectrally shaping the source. Sound sources can also be generated by constrictions and boundaries, not shown in Figure 3.1, that are made within the vocal tract itself, yielding in addition to noisy and periodic sources, an impulsive airflow source. We have here idealized the sources in the sense that the anatomy and physiology of the speech production mechanism does not generate a perfect periodic, impulsive, or noise source.\(^1\) Following the spectral coloring of the source by the vocal tract, the variation of air pressure at the lips results in a traveling sound wave that the listener perceives as speech.

There are then three general categories of the source for speech sounds: periodic, noisy, and impulsive, although combinations of these sources are often present. Examples of speech sounds generated with each of these source categories are seen in the word “shop,” where the “sh,” “o,” and “p” are generated from a noisy, periodic, and impulsive source, respectively. The reader should speak the word “shop” slowly and determine where each sound source is occurring, i.e., at the larynx or at a constriction within the vocal tract.

\(^1\) This idealization also assumes a flat (white) noise spectrum. Noise and its white subclass are defined formally in a stochastic signal framework in Chapter 5.
Figure 3.1 Simple view of speech production. The sound sources are idealized as periodic, impulsive, or (white) noise and can occur in the larynx or vocal tract.
Such distinguishable speech sounds are determined not only by the source, but by different vocal tract configurations, and how these shapes combine with periodic, noisy, and impulsive sources. These more refined speech sound classes are referred to as phonemes, the study of which is called phonemics. A specific phoneme class provides a certain meaning in a word, but within a phoneme class, as we will see in a moment, there exist many sound variations that provide the same meaning. The study of these sound variations is called phonetics. Phonemes, the basic building blocks of a language, are concatenated, more or less, as discrete elements into words, according to certain phonemic and grammatical rules. This chapter provides a qualitative description of the speech production mechanism and the resulting variety of phonetic sound patterns, and, to a lesser extent, how these sound patterns differ among different speakers. Implications for the design of digital signal processing algorithms will be illustrated. In Chapter 4, we refine this qualitative description with more quantitative mathematical models.

In Section 3.2, we first describe the anatomy and physiology of the different organ groups and show how these organ groups result in source inputs and vocal tract configurations that contribute generally to making different speech sounds. Time- and frequency-domain properties of the source and its spectral shaping by the vocal tract are illustrated, and these result in a number of important definitions, such as the pitch and harmonics of a periodic source and the formants of the vocal tract. In this section, we also elaborate on sound categorization based on source only: periodic, noisy, and impulsive sound sources. In Section 3.3, we deviate and develop the spectrogram, which is a means to illustrate the spectral evolution of a sound; in Chapter 7, the spectrogram will be studied more formally. Having four tools in hand—the time-waveform, spectrogram, source classification, and vocal tract configurations—we then embark in Section 3.4 on the study of phonetics. In Section 3.5, we take a wider temporal view of the speech waveform, i.e., across phonetic boundaries of individual speech sounds, and study the prosodies of speech, which is the rhythm (timing of the phonemes) and intonation (changing pitch of the source) over phrases and sentences. In Section 3.6, we give a flavor for the perceptual aspect of phonetics, i.e., how the auditory system might perceive a speech sound, and how various properties of sound production are important in the distinguishing of different speech phonemes. We will see in later chapters how characteristics of speech production, used as perceptual cues, can drive the development and selection of signal processing algorithms.

### 3.2 Anatomy and Physiology of Speech Production

Figure 3.2 shows a more realistic view of the anatomy of speech production than was shown in Figure 3.1. We now look in detail at this anatomy, as well as at the associated physiology and its importance in speech production.

#### 3.2.1 Lungs

One purpose of the lungs is the inhalation and exhalation of air. When we inhale, we enlarge the chest cavity by expanding the rib cage surrounding the lungs and by lowering the diaphragm that sits at the bottom of the lungs and separates the lungs from the abdomen; this action lowers the air pressure in the lungs, thus causing air to rush in through the vocal tract and down the trachea.
into the lungs. The trachea, sometimes referred to as the “windpipe,” is about a 12-cm-long and 1.5–2-cm-diameter pipe which goes from the lungs to the epiglottis. The epiglottis is a small mass, or “switch,” which, during swallowing and eating, deflects food away from entering the trachea. When we eat, the epiglottis falls, allowing food to pass through a tube called the esophagus and into the stomach. When we exhale, we reduce the volume of the chest cavity by contracting the muscles in the rib cage, thus increasing the lung air pressure. This increase in pressure then causes air to flow through the trachea into the larynx. In breathing, we rhythmically inhale to take in oxygen, and exhale to release carbon dioxide.

During speaking, on the other hand, we take in short spurts of air and release them steadily by controlling the muscles around the rib cage. We override our rhythmic breathing by making the duration of exhaling roughly equal to the length of a sentence or phrase. During this timed exhalation, the lung air pressure is maintained at approximately a constant level, slightly above atmospheric pressure, by steady slow contraction of the rib cage, although the air pressure varies around this level due to the time-varying properties of the larynx and vocal tract.

### 3.2.2 Larynx

The larynx is a complicated system of cartilages, muscles, and ligaments\(^2\) whose primary purpose, in the context of speech production, is to control the vocal cords or vocal

---

\(^2\) Some definitions useful throughout this chapter are: (1) muscles are tissue that contract when stimulated; (2) cartilage is rigid, yet elastic bony tissue, but not as hard as bone, helping to support organs in place; (3) ligaments are tough bands of tissue that connect bones to bones and also support organs in place.
The vocal folds are two masses of flesh, ligament, and muscle, which stretch between the front and back of the larynx, as illustrated in Figure 3.3. The folds are about 15 mm long in men and 13 mm long in women. The glottis is the slit-like orifice between the two folds. The folds are fixed at the front of the larynx where they are attached to the stationary thyroid cartilage. The thyroid cartilage is located at the front (or Adam’s apple) and sides of the larynx. The folds are free to move at the back and sides of the larynx; they are attached to the two arytenoid cartilages that move in a sliding motion at the back of the larynx along with the cricoid cartilage. The size of the glottis is controlled in part by the arytenoid cartilages, and in part by muscles within the folds. Another important property of the vocal folds, in addition to the size of the glottis, is their tension. The tension is controlled primarily by muscle within the folds, as well as the cartilage around the folds. The vocal folds, as well as the epiglottis, close during eating, thus providing a second protection mechanism. The false vocal folds, above the vocal folds (Figure 3.2), provide a third protection. They also extend from the Adam’s apple to the arytenoids. They can be closed and they can vibrate, but they are likely open during speech production [4]. We see then that a triple barrier is provided across the windpipe through the action of the epiglottis, the false vocal folds, and the true vocal folds. All three are closed during swallowing and wide open during breathing.

There are three primary states of the vocal folds: breathing, voiced, and unvoiced. In the breathing state, the arytenoid cartilages are held outward (Figure 3.3b), maintaining a wide glottis, and the muscles within the vocal folds are relaxed. In this state, the air from the lungs flows freely through the glottis with negligible hindrance by the vocal folds. In speech production, on the other hand, an obstruction of airflow is provided by the folds. In the voicing state, as, for example, during a vowel, the arytenoid cartilages move toward one another (Figure 3.3a). The vocal folds tense up and are brought close together. This partial closing of the glottis and increased fold tension cause self-sustained oscillations of the folds. We can describe how this oscillation comes about in three steps [10] (Figure 3.4a).

3 The more accurate term is “vocal folds,” since the masses are actually not cords. The term “vocal cords” originated with an early erroneous anatomical study [30]. Although we use the term “vocal folds” more often, we apply the two terms interchangeably throughout the text.
Figure 3.4 Bernoulli’s Principle in the glottis: (a) basic horizontal open/close voicing cycle; (b) refinement of (a) with vertical vocal fold motion. Vertical lines represent airflow in the direction of the arrows.

Suppose the vocal folds begin in a loose and open state. The contraction of the lungs first results in air flowing through the glottis. According to a fluid dynamic property called Bernoulli’s Principle, as the airflow velocity (i.e., the velocity of air particles) increases, local pressure in the region at the glottis decreases. At the same time, tension in the vocal folds increases. This increase in tension of the folds, together with the decrease in pressure at the glottis, causes the vocal folds to close shut abruptly. Air pressure then builds behind the vocal folds as the lungs continue to contract, forcing the folds to open. The entire process then repeats and the result is periodic “puffs” of air that enter the vocal tract.

Thus far, we have illustrated the vocal folds as vibrating horizontally, perpendicular to the tracheal wall. The vocal fold movement, however, is generally not so simple. For example, both horizontal and vertical movement of the folds may occur simultaneously, as illustrated in Figure 3.4b. During the time when the glottis is open, because the lower parts of the fleshy folds are more flexible than the upper parts, there is a time delay between the closing of the two regions, as seen in Steps 1–3 of Figure 3.4b. Additional vertical movement then occurs because there is also a time delay between the opening of the two regions. When the air pressure below the glottis increases during the time when the glottis closes, the lower region of the folds is first pushed up, followed by the upper region, as seen in Steps 4–6. Such complexity has led to a nonlinear two-mass model [11] (Figure 3.5), as well as more elaborate nonlinear multi-component models describing various modes of vibration along the folds themselves [39]. The masses $m_k$, nonlinear spring constants $s_k$, and damping constants $\tau_k$ in such mechanical models correspond, respectively, to the masses, tensions, and resistances within the vocal folds and the surrounding cartilage.

According to our description of the airflow velocity in the glottis, if we were to measure the airflow velocity at the glottis as a function of time, we would obtain a waveform approximately
3.2 Anatomy and Physiology of Speech Production

Figure 3.5 Two-mass mechanical model of Flanagan and Ishizaka with masses $m_1$ and $m_2$, resistances $\tau_1$ and $\tau_2$, and spring constants $s_1$ and $s_2$.


similar to that illustrated in Figure 3.6 that roughly follows the time-varying area of the glottis. Typically, with the folds in a closed position, the flow begins slowly, builds up to a maximum, and then quickly decreases to zero when the vocal folds abruptly shut. The time interval during which the vocal folds are closed, and no flow occurs, is referred to as the glottal *closed phase*; the time interval over which there is nonzero flow and up to the maximum of the airflow velocity is referred to as the glottal *open phase*, and the time interval from the airflow maximum to the time of glottal closure is referred to as the *return phase*. The specific flow shape can change with the speaker, the speaking style, and the specific speech sound. In some cases, the folds do not even close completely, so that a closed phase does not exist. For simplicity throughout this text, we will often refer to the glottal airflow velocity as simply the *glottal flow*.

The time duration of one glottal cycle is referred to as the *pitch period* and the reciprocal of the pitch period is the corresponding *pitch*, also referred to as the *fundamental frequency*. The term “pitch” might lead to some confusion because the term is often used to describe the subjectively perceived “height” of a complex musical sound even when no single fundamental frequency exists. In this text, however, we use the term in the above strict sense, i.e., pitch is synonymous with fundamental frequency. In conversational speech, during vowel sounds, we
might see typically one to four pitch periods over the duration of the sound, although, as we will see in the discussion of prosodics, the number of pitch periods changes with numerous factors such as stress and speaking rate. The rate at which the vocal folds oscillate through a closed, open, and return cycle is influenced by many factors. These include vocal fold muscle tension (as the tension increases, so does the pitch), the vocal fold mass (as the mass increases, the pitch decreases because the folds are more sluggish), and the air pressure behind the glottis in the lungs and trachea, which might increase in a stressed sound or in a more excited state of speaking (as the pressure below the glottis increases, so does the pitch). The pitch range is about 60 Hz to 400 Hz. Typically, males have lower pitch than females because their vocal folds are longer and more massive.

A simple mathematical model of the glottal flow is given by the convolution of a periodic impulse train with the glottal flow over one cycle. The following example shows glottal flow waveforms with different shapes and pitch periods, as well as how the simple convolutional model lends insight into the spectral nature of the glottal airflow.

**EXAMPLE 3.1** Consider a glottal flow waveform model of the form

\[ u[n] = g[n] * p[n] \]  

where \( g[n] \) is the glottal flow waveform over a single cycle and \( p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP] \) is an impulse train with spacing \( P \). Because the waveform is infinitely long, we extract a segment by multiplying \( x[n] \) by a short sequence called an *analysis window* or simply a *window*. The window, denoted by \( w[n, \tau] \), is centered at time \( \tau \), as illustrated in Figure 3.7, and the resulting waveform segment is written as

\[ u[n, \tau] = w[n, \tau](g[n] * p[n]). \]

Using the Multiplication and Convolution Theorems of Chapter 2, we obtain in the frequency domain

\[ U(\omega, \tau) = \frac{1}{P} W(\omega, \tau) \otimes \left[ \sum_{k=-\infty}^{\infty} G(\omega) \delta(\omega - \omega_k) \right] \]
Figure 3.7 Illustration of periodic glottal flow in Example 3.1: (a) typical glottal flow and its spectrum; (b) same as (a) with lower pitch; and (c) same as (a) with “softer” or more “relaxed” glottal flow.

\[
\frac{1}{P} \sum_{k=\infty} G(\omega_k) W(\omega - \omega_k, \tau)
\]

where \( W(\omega, \tau) \) is the Fourier transform of \( w[n, \tau] \), where \( G(\omega) \) is the Fourier transform of \( g[n] \), where \( \omega_k = \frac{2\pi}{P} k \), and where \( \frac{2\pi}{P} \) is the fundamental frequency or pitch. As illustrated in Figure 3.7, the Fourier transform of the window sequence is characterized by a narrow main lobe centered at \( \omega = 0 \) with lower surrounding sidelobes. The window is typically selected to trade off the width of the mainlobe and attenuation of the sidelobes. Figure 3.7 illustrates how the Fourier transform magnitude of the waveform segment changes with pitch and with characteristics of the glottal flow. As the pitch period decreases, the spacing between the frequencies \( \omega_k = \frac{2\pi}{P} k \), which are referred to as the harmonics of the glottal waveform, increases, as can be seen by comparing Figures 3.7a and 3.7b. The first harmonic is also the fundamental frequency, and the other harmonics occur at integer
multiples of the fundamental frequency. Located at each harmonic is a translated window Fourier transform $W(\omega - \omega_k)$ weighted by $G(\omega_k)$; as the pitch changes, the harmonics can be thought of as sliding under $G(\omega)$. As the glottal flow over a cycle becomes more smooth, i.e., a gradual rather than an abrupt closing, then the "spectral shaping" by $G(\omega)$ of the harmonically-spaced window Fourier transforms becomes more lowpass, as seen by comparing Figures 3.7a and 3.7c. We can see, based on these sliding and spectral shaping properties, why the magnitude of the spectral shaping function, in this case $|G(\omega)|$, is sometimes referred to as a spectral envelope of the harmonics.

We saw in the previous example that the Fourier transform of the periodic glottal waveform is characterized by harmonics. Typically, the spectral envelope of the harmonics, governed by the glottal flow over one cycle, has, on the average, a $-12$ dB/octave rolloff, although this changes with the specific nature of the airflow and the speaker characteristics. With more forceful speaking, for example, the glottal closure may be more abrupt (e.g., Figure 3.7a, b) with perhaps an average $-9$ dB/octave slope being more typical [29]. In more "relaxed" voicing, the vocal folds do not close as abruptly, and the glottal waveform has more rounded corners (e.g., Figure 3.7c), with an average $-15$ dB/octave rolloff, typically. Exercise 3.18 explores some specific cases. The model in Example 3.1 is ideal in the sense that, even for sustained voicing—i.e., a vowel uttered by a speaker trying to hold steady pitch and vocal tract shape—a fixed pitch period is almost never maintained in time but can randomly vary over successive periods, a characteristic referred to as pitch “jitter.” In addition, the amplitude of the airflow velocity within a glottal cycle may differ across consecutive pitch periods, even in a sustained vowel, a characteristic called amplitude “shimmer.” These variations are due, perhaps, to time-varying characteristics of the vocal tract and vocal folds. Pitch jitter and shimmer, however, have also been speculated to be due to nonlinear behavior in the speech anatomy whereby successive cyclic variations may alternate on each glottal cycle [38] or may appear random while being the result of an underlying deterministic (chaotic) system [15]. The jitter and shimmer over successive pitch periods help give the vowel its naturalness, in contrast to a monotone pitch and fixed amplitude that can result in a machine-like sound. In addition to naturalness, however, the extent and form of jitter and shimmer can contribute to voice character. A high degree of jitter, for example, results in a voice with a hoarse quality which can be characteristic of a particular speaker or can be created under specific speaking conditions such as with stress or fear. The time- and frequency-domain properties of this condition are further studied in Exercise 3.2.

We have described two states of the vocal folds: breathing and voicing. The last state of the vocal folds is unvoicing. This state is similar to the breathing state in there being no vocal fold vibration. In the unvoiced state, however, the folds are closer together and more tense than in the breathing state, thus allowing for turbulence to be generated at the folds themselves. Turbulence at the vocal folds is called aspiration. Aspiration occurs in normal speech as with “h” in the word “he.” Such sounds are sometimes called “whispered” sounds because turbulence is also created at the vocal folds when we whisper. Whispering is not simply a reduction in volume, because when we whisper the vocal folds do not oscillate. In certain voice types, aspiration occurs normally simultaneously with voicing, resulting in the breathy voice, by maintaining part of the vocal folds nearly fixed and somewhat open to produce turbulence and part of the vocal folds in oscillation. Nevertheless, aspiration occurs to some extent in all speakers and the amount of aspiration may serve as a distinguishing feature. The physiological change, then, in
creating the breathy voice is distinctly different from that of the hoarse voice which, as we saw earlier, is associated with pitch jitter. Figure 3.8 shows a comparison of vocal fold configurations for aspiration (whispering), voicing, and aspirated voicing.

There are also other forms of vocal fold movement that do not fall clearly into any of the three states of breathing, voicing, or unvoicing. We point out these different voice types because, as we will see, they can pose particularly large challenges in speech signal processing and, contrary to being “idiosyncratic,” they occur quite often. One such state of the vocal folds is the creaky voice where the vocal folds are very tense, with only a short portion of the folds in oscillation, resulting in a harsh-sounding voice with a high and irregular pitch. (Look ahead to Figure 10.15b.) In vocal fry, on the other hand, the folds are massy and relaxed with an abnormally low and irregular pitch [27],[40], which is characterized by secondary glottal pulses close to and overlapping the primary glottal pulse within the open phase, as illustrated in Figure 3.9a. We use the term “glottal pulse” loosely in this chapter to mean a glottal airflow velocity waveform over a single glottal cycle. In vocal fry, the true vocal folds may couple with the false vocal folds, producing the secondary glottal pulses. Vocal fry occurs even in the normal voice at the end of a phrase or word where the muscles of the larynx relax and the lung pressure is decreasing. Another atypical voice type is the diplophonic voice where again secondary glottal pulses occur between the primary pulses but within the closed phase, away from the primary pulse [18], as illustrated in Figure 3.9b. Diplophonia often occurs as extra flaps in low-pitch speakers and, as with vocal fry, in normal voices at the end of a phrase or word. An example of a low-pitch diplophonic voice is provided later, in Figure 3.16. In the diplophonic and vocal fry voice types, a simple model in discrete time for the occurrence of a secondary glottal pulse is given by the modified glottal flow waveform \( \tilde{g}[n] = g[n] + \alpha g[n - n_o] \), where \( g[n] \) is the primary glottal pulse, where \( n_o \) is the spacing between the primary and secondary glottal pulses, and \( \alpha \) is an attenuation factor on the secondary pulse. We assume here the same shape of the

![Figure 3.8](image)

Figure 3.8 Sketches of various vocal fold configurations: (a) aspiration (whispering), (b) voicing, and (c) aspirated voicing. Arrows indicate vocal fold vibration, while ragged lines indicate turbulence.

---

4 Pitch period doubling also often occurs at the end of a phrase or word where the vocal cords are relaxed. In Figure 3.13, the “o” in “to” shows this phenomenon.

5 We assume that the spacing in continuous time is \( t_o = n_o T \) (with \( T \) being the sampling interval) so that in discrete time the spacing is represented by the integer \( n_o \).
secondary and primary glottal pulses, but generally they may differ. The presence of \( \alpha g[n - n_o] \) introduces zeros into the z-transform of the glottal waveform (Exercise 3.3). Other abnormal voice types include the falsetto voice where there occurs extreme swings in vocal fold tension and relaxation allowing for abnormally large pitch fluctuations. Some speakers, especially in the singing voice, may regularly induce a rapid pitch modulation, referred to as pitch \textit{vibrato}, over a smaller range to give the utterance more flavor or musicality. We will see examples of some of these voice types throughout the text.

### 3.2.3 Vocal Tract

The vocal tract is comprised of the oral cavity from the larynx to the lips and the nasal passage that is coupled to the oral tract by way of the velum. The oral tract takes on many different lengths and cross-sections by moving the tongue, teeth, lips, and jaw and has an average length of 17 cm in a typical adult male and shorter for females, and a spatially-varying cross section of up to 20 cm\(^2\). If we were to listen to the pressure wave at the output of the vocal folds during voicing, we would hear simply a time-varying buzz-like sound which is not very interesting. One purpose of the vocal tract is to spectrally “color” the source, which is important for making perceptually distinct speech sounds. A second purpose is to generate new sources for sound production.
Spectral Shaping — Under certain conditions, the relation between a glottal airflow velocity input and vocal tract airflow velocity output can be approximated by a linear filter with resonances, much like resonances of organ pipes and wind instruments. The resonance frequencies of the vocal tract are, in a speech science context, called formant frequencies or simply formants. The word “formant” also refers to the entire spectral contribution of a resonance so we often use the phrases “formant bandwidth” and “formant amplitude” (at the formant frequency). Formants change with different vocal tract configurations. With different vowels, for example, the jaw, teeth, lips, and tongue, are generally in different positions. Panel (a) of Figure 3.10 shows the tongue hump high in the front and back of the palate (upper wall of mouth), each position corresponding to different resonant cavities and thus different vowels.

The peaks of the spectrum of the vocal tract response correspond approximately to its formants. More specifically, when the vocal tract is modeled as a time-invariant all-pole linear system then, as we will see in Chapter 4, a pole at $z_o = r_o e^{j\omega_o}$ corresponds approximately to a vocal tract formant. The frequency of the formant is at $\omega = \omega_o$ and the bandwidth of the formant is determined by the distance of the pole from the unit circle ($r_o$). Because the poles of a real sequence typically occur in complex conjugate pairs (except for the case of a pole falling on the real axis), only the positive frequencies are used in defining the formant frequencies, and the formant bandwidth is computed over positive frequencies using, for example, the definitions of bandwidth in Chapter 2. Under the linear time-invariant all-pole assumption, each vocal tract shape is characterized by a collection of formants. Because the vocal tract is assumed stable with poles inside the unit circle, the vocal tract transfer function can be expressed either in product or partial fraction expansion form:

$$H(z) = \frac{A}{\prod_{k=1}^{N_i} (1 - c_k z^{-1}) (1 - c_k^* z^{-1})} = \sum_{k=1}^{N_i} \tilde{A} \frac{(1 - c_k z^{-1}) (1 - c_k^* z^{-1})}{(1 - c_k z^{-1}) (1 - c_k^* z^{-1})}$$

(3.2)

![Figure 3.10](image-url)  
Figure 3.10 Illustration of changing vocal tract shapes for (a) vowels (having a periodic source), (b) plosives (having an impulsive source), and (c) fricatives (having a noise source).
where \((1 - c_k z^{-1})\) and \((1 - c_k^* z^{-1})\) are complex conjugate poles inside the unit circle with \(|c_k| < 1\). The formants of the vocal tract are numbered from the low to high formants according to their location; the first formant is denoted by \(F_1\), the second formant by \(F_2\), and so on up to the highest formant. Generally, the frequencies of the formants decrease as the vocal tract length increases; as a consequence, a male speaker tends to have lower formants than a female, and a female has lower formants than a child. Under a vocal tract linearity and time-invariance assumption, and when the sound source occurs at the glottis, the speech waveform, i.e., the airflow velocity at the vocal tract output, can be expressed as the convolution of the glottal flow input and vocal tract impulse response, as illustrated in the following example:

**Example 3.2** Consider a periodic glottal flow source of the form

\[
u[n] = g[n] * p[n]
\]

where \(g[n]\) is the airflow over one glottal cycle and \(p[n]\) is the unit sample train with spacing \(P\). When the sequence \(u[n]\) is passed through a linear time-invariant vocal tract with impulse response \(h[n]\), the vocal tract output is given by

\[
x[n] = h[n] * (g[n] * p[n]).
\]

A window centered at time \(\tau\), \(w[n, \tau]\), is applied to the vocal tract output to obtain the speech segment

\[
x[n, \tau] = w[n, \tau] [h[n] * (g[n] * p[n])].
\]

Using the Multiplication and Convolution Theorems of Chapter 2, we obtain in the frequency domain the Fourier transform of the speech segment

\[
X(\omega, \tau) = \frac{1}{P} W(\omega, \tau) \ast \left[ H(\omega)G(\omega) \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_k) \right]
\]

\[
= \frac{1}{P} \sum_{k=-\infty}^{\infty} H(\omega_k)G(\omega_k)W(\omega - \omega_k, \tau)
\]

where \(W(\omega, \tau)\) is the Fourier transform of \(w[n, \tau]\), where \(\omega_k = \frac{2\pi}{P} k\), and where \(\frac{2\pi}{P}\) is the fundamental frequency or pitch. Figure 3.11 illustrates that the spectral shaping of the window transforms at the harmonics \(\omega_1, \omega_2, \ldots \omega_N\) is determined by the spectral envelope \(|H(\omega)G(\omega)|\) consisting of a glottal and vocal tract contribution, unlike in Example 3.1, where only the glottal contribution occurred. The peaks in the spectral envelope correspond to vocal-tract formant frequencies, \(F_1, F_2, \ldots F_M\). The general upward or downward slope of the spectral envelope, sometimes called the *spectral tilt*, is influenced by the nature of the glottal flow waveform over a cycle, e.g., a gradual or abrupt closing, and by the manner in which formant tails add. We also see in Figure 3.11 that the formant locations are not always clear from the short-time Fourier transform magnitude \(|X(\omega, \tau)|\) because of sparse sampling of the spectral envelope \(|H(\omega)G(\omega)|\) by the source harmonics, especially for high pitch.
3.2 Anatomy and Physiology of Speech Production

This example illustrates the important difference between a formant, or resonance, frequency and a harmonic frequency. A formant corresponds to the vocal tract poles, while the harmonics arise from the periodicity of the glottal source. The spectrum of the vocal tract, for a perfectly periodic source, is, in essence, sampled at the harmonic frequencies; with this idealized perfect periodicity, there is spectral information only at the harmonics. In the development of signal processing algorithms that require formants, this sparcity of spectral information can perhaps be a detriment to formant estimation. In some situations, on the other hand, the spectral sampling at harmonics can be exploited to enhance perception of a sound, as in the singing voice.

**EXAMPLE 3.3** A soprano singer often sings a tone whose first harmonic (fundamental frequency $\omega_1$) is much higher than the first formant frequency ($F_1$) of the vowel being sung [37]. As shown in Figure 3.12, when the nulls of the vocal tract spectrum are sampled at the harmonics, the resulting sound is weak, especially in the face of competing instrumentals. To enhance the sound, the singer creates a vocal tract configuration with a widened jaw which increases the first formant frequency (Exercise 3.4), and can match the frequency of the first harmonic, thus generating a louder sound [37] (Figure 3.12). In training, the singer is asked to “Hear the next tone within yourself before you start to sing it” because a widening of the jaw requires some articulatory anticipation [37].

We have seen that the nasal and oral components of the vocal tract are coupled by the velum. When the vocal tract velum is lowered, introducing an opening into the nasal passage, and the oral tract is shut off by the tongue or lips, sound propagates through the nasal passage and out through the nose. The resulting nasal sounds, e.g., “m” as in “meet,” have a spectrum that is dominated by low-frequency formants of the large volume of the nasal cavity. Because the nasal cavity, unlike the oral tract, is essentially constant, characteristics of nasal sounds may

---

$^6$ A singer will also lower his/her larynx in such a way as to introduce a new high-frequency formant between 2500–3000 Hz, a frequency region where the background instrumental is low, to help further enhance the sound [37].
be particularly useful in speaker identifiability. The velum can be lowered even when the oral tract is open. When this coupling occurs, we obtain a nasalized vowel. One effect of the nasal passage is that the formant bandwidths of the oral tract become broader because of loss of energy through the nasal passage. A second effect is the introduction of anti-resonances, i.e., zeros, in the vocal tract transfer function due to the absorption of energy at the resonances of the nasal passage [29].

The previous discussion has assumed a linear time-invariant vocal tract. Formants, however, are time-varying because the vocal tract changes in time. Although the vocal tract is almost time-invariant for steady-state sounds, as with a sustained vowel, in normal conversational speech the vocal tract is continuously and significantly changing. This time-variation will influence signal analysis techniques. We will return to this speech production characteristic in the discussion of transitional speech sounds.

**Source Generation** — We have seen that different vocal tract shapes correspond to different resonant cavities; different vocal tract shapes can also result in different sound sources. The panel (b) of Figure 3.10 shows a complete closure of the tract, the tongue pressing against the palate, required in making an impulsive sound source. There is a build-up of pressure behind the closure and then an abrupt release of pressure. Panel (c) shows another sound source created
with the tongue close to the palate, but not completely impeded, for the generation of turbulence and thus a noise source. As with a periodic glottal sound source, a spectral shaping similar to that described in Example 3.2 also occurs for either type of input, i.e., an impulsive or noise source; this spectral shaping is performed by a resonant vocal tract cavity whose formants change with different vocal tract configurations, such as those illustrated in panels (b) and (c) of Figure 3.10. There is not, however, harmonic structure in the impulsive or noise source spectrum, but rather the source spectrum is shaped at all frequencies by \( |H(\omega)| \). Keep in mind that we have idealized the impulsive and noise sources to have flat spectra; in practice, these sources will themselves have a non-flat spectral shape.

There is yet one other source type that is generated within the vocal tract, but is less understood than noisy and impulsive sources occurring at oral tract constrictions. This source arises from the interaction of vortices with vocal tract boundaries such as the false vocal folds, teeth, or occlusions in the oral tract [1],[38]. The reader may have intuition about the nature of a vortex; for the moment, let’s think of a vortex in the oral tract as a tiny rotational airflow. For voiced speech, the vortices move possibly as a train from the glottis to the lips along the oral tract and are predicted to initiate from the air jet emanating from the glottis during vocal fold vibration [1],[38]. Vortices can also arise during fricative sounds with resulting sources distributed along the oral tract [19]. There is evidence that sources due to vortices influence the temporal and spectral, and perhaps perceptual, characteristics of speech sounds [1],[19],[38]. We delay further discussion of these vortical sound sources until Chapter 11.

### 3.2.4 Categorization of Sound by Source

There are various ways to categorize speech sounds. For example, we can categorize speech sounds based on different sources to the vocal tract; we have seen that different sources are due to the vocal fold state, but are also formed at various constrictions in the oral tract. Speech sounds generated with a periodic glottal source are termed **voiced**; likewise, sounds not so generated are called **unvoiced**. There are a variety of unvoiced sounds, including those created with a noise source at an oral tract constriction. Because the noise of such sounds comes from the friction of the moving air against the constriction, these sounds are sometimes referred to as **fricatives** (Figure 3.10c). An example of frication is in the sound “th” in the word “thin” where turbulence is generated between the tongue and the upper teeth. The reader should hold the “th” sound and feel the turbulence. A second unvoiced sound class is **plosives** created with an impulsive source within the oral tract (Figure 3.10b). An example of a plosive is the “t” in the word “top.” The location of the closed or partial constriction corresponds to different plosive or fricative sounds, respectively. We noted earlier that a barrier can also be made at the vocal folds by partially closing the vocal folds, but without oscillation, as in the sound “h” in “he.” These are **whispered** unvoiced speech sounds. These voiced and unvoiced sound categories, however, do not relate exclusively to the source state because a combination of these states can also be made whereby vocal fold vibration occurs simultaneously with impulsive or noisy sources. For example, with “z” in the word “zebra,” the vocal folds are vibrating and, at the same time, noise is created at a vocal tract constriction behind the teeth against the palate. Such sounds are referred to as **voiced fricatives** in contrast to **unvoiced fricatives** where the vocal folds do not vibrate simultaneously with frication. There also exist **voiced plosives** as counterparts to **unvoiced plosives** as with the “b” in the word “boat.” Examples of some of these sound classes are shown in Figure 3.13 in the sentence, “Which tea party did Baker go to?”
This loose classification provides a stepping stone to Section 3.4 where distinctive features of these sound classes will be further studied and where we will combine this source categorization with different vocal tract configurations to form the more complete classification of elements of a language.

3.3 Spectrographic Analysis of Speech

We have seen that a speech waveform consists of a sequence of different events. This time-variation corresponds to highly fluctuating spectral characteristics over time. For example, in the word “to,” the plosive “t” is characterized by high-frequency energy corresponding to a vocal
3.3 Spectrographic Analysis of Speech

The “t” is followed by the vowel “o,” which is characterized by low-frequency energy corresponding to a vocal tract configured as a long cavity along the oral tract. We will show quantitatively in Chapter 4 how such spectral energy shifts occur with different cavity lengths and cross sections. A single Fourier transform of the entire acoustic signal of the word “to” cannot capture this time-varying frequency content. In contrast, the short-time Fourier transform (STFT) consists of a separate Fourier transform of pieces of the waveform under a sliding window. We have already introduced this sliding window in Examples 3.1 and 3.2 and denoted it by \( w[n, \tau] \), where \( \tau \) is the position of the window center. The window is typically tapered at its end (Figure 3.14) to avoid unnatural discontinuities in the speech segment and distortion in its underlying spectrum. The Hamming window, for example, is given by the sequence \( w[n, \tau] = 0.54 - 0.4 \cos \left( \frac{2\pi(n-\tau)}{N_w-1} \right) \) for \( 0 \leq n \leq N_w - 1 \) and zero otherwise, with \( N_w \) as the window duration. As we mentioned earlier, the window is typically selected to trade off the width of its mainlobe and attenuation of its sidelobes. The effect of specific window shapes will be further discussed in Chapter 7. In practice, the window does not necessarily move one sample at a time, but rather moves at some frame interval consistent with the temporal structure one wants to reveal.

The Fourier transform of the windowed speech waveform, i.e., the STFT, is given by

\[
X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x[n, \tau] \exp(-j\omega n) \tag{3.3}
\]

where

\[
x[n, \tau] = w[n, \tau] x[n]
\]

represents the windowed speech segments as a function of the window center at time \( \tau \). The spectrogram is a graphical display of the magnitude of the time-varying spectral characteristics and is given by

\[
S(\omega, \tau) = |X(\omega, \tau)|^2
\]

which can be thought of as a two-dimensional (2-D) “energy density,” i.e., a generalization of the one-dimensional (1-D) energy density associated with the Fourier transform, describing the relative energy content in frequency at different time locations, i.e., in the neighborhood of \( (\omega, \tau) \), as we move, for example, from plosive to voiced to fricative sounds. We will have more to say about \( S(\omega, \tau) \) as a 2-D energy density in following chapters.\(^7\) We could plot \( S(\omega, \tau) \) for each window position \( \tau \) to represent the spectral time variations, but we would soon run out of space. A more compact time-frequency display of the spectrogram places the spectral magnitude measurements vertically in a three-dimensional mesh or two-dimensionally with intensity coming out of the page. This later display is illustrated in Figure 3.14 where the Fourier transform magnitudes of the segments \( x[n, \tau] \) are shown laid out on the 2-D time-frequency grid. The figure also indicates two kinds of spectrograms: narrowband, which gives good spectral resolution, e.g., a good view of the frequency content of sinewaves with closely

\(^7\) The notion of \( S(\omega, \tau) \) as a 2-D energy density follows from the relation \( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |S(\omega, \tau)|^2 d\omega \) that holds under certain conditions on the window \( w[n, \tau] \).
Figure 3.14 Formation of (a) the narrowband and (b) the wideband spectrograms.
3.3 Spectrographic Analysis of Speech

spaced frequencies, and wideband, which gives good temporal resolution, e.g., a good view of the temporal content of impulses closely spaced in time.\(^8\) We introduce the two classes of spectrograms using speech sounds with a voiced source as an example.

For voiced speech, we have approximated the speech waveform as the output of a linear time-invariant system with impulse response \(h[n]\) and with a glottal flow input given by the convolution of the glottal flow over one cycle, \(g[n]\), with the impulse train \(p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP]\). This results in the windowed speech waveform expressed as

\[
x[n, \tau] = w[n, \tau] \{ (p[n] * g[n]) * h[n] \} = w[n, \tau] (p[n] * \tilde{h}[n])
\]

where we have written the glottal waveform over a cycle and vocal tract impulse response as lumped into \(\tilde{h}[n] = g[n] * h[n]\). Using the result of Example 3.2, the spectrogram of \(x[n]\) can therefore be expressed as

\[
S(\omega, \tau) = \frac{1}{P^2} \sum_{k=-\infty}^{\infty} |\tilde{H}(\omega_k) W(\omega - \omega_k, \tau)|^2
\]  

(3.4)

where

\[
\tilde{H}(\omega) = H(\omega) G(\omega)
\]

and where \(\omega_k = \frac{2\pi}{P} k\) and \(\frac{2\pi}{P}\) is the fundamental frequency.

**Narrowband Spectrogram** — The difference between the narrowband and wideband spectrogram is the length of the window \(w[n, \tau]\). For the narrowband spectrogram, we use a “long” window with a duration of typically at least two pitch periods. Under the condition that the main lobes of shifted window Fourier transforms are non-overlapping and that corresponding transform sidelobes are negligible, Equation (3.4) leads to the approximation

\[
S(\omega, \tau) \approx \frac{1}{P^2} \sum_{k=-\infty}^{\infty} |\tilde{H}(\omega_k)|^2 |W(\omega - \omega_k, \tau)|^2.
\]  

(3.5)

This approximation is left as an exercise (Exercise 3.8). We see then that using a long window gives a short-time Fourier transform of voiced speech that consists of a set of narrow “harmonic lines,” whose width is determined by the Fourier transform of the window, which are shaped by the magnitude of the product of the glottal flow Fourier transform and vocal tract transfer function. The narrowband spectrogram gives good frequency resolution because the harmonic lines are “resolved”; these harmonic lines are seen as horizontal striations in the time-frequency plane of the spectrogram. The long window, however, covers several pitch periods and thus is unable to reveal fine periodicity changes over time; it also smears closely spaced temporal

---

\(^8\) More precise definitions of temporal and spectral resolution are given later in the text. For our purpose here an intuition for the concepts is sufficient.
events and thus gives poor time resolution, as with a plosive that is closely spaced to a succeeding voiced sound.

**Wideband Spectrogram** — For the *wideband* spectrogram, we choose a “short” window with a duration of less than a single pitch period (Figure 3.14); shortening the window widens its Fourier transform (recall the uncertainty principle). The wide Fourier transform of the window, when translated to harmonics, will overlap and add with its neighboring window transforms and smear out the harmonic line structure, roughly tracing out the spectral envelope $|\tilde{H}(\omega)|$ due to the vocal tract and glottal flow contributions. In an alternative temporal perspective, since the window length is less than a pitch period, as the window slides in time it “sees” essentially pieces of the periodically occurring sequence $\tilde{h}[n]$ (assuming tails of previous responses have died away). For the steady-state voiced sound, we can therefore express the wideband spectrogram (very) roughly (Exercise 3.9 asks the reader to complete the argument) as

$$S(\omega, \tau) \approx \beta |\tilde{H}(\omega)|^2 E[\tau] \quad (3.6)$$

where $\beta$ is a constant scale factor and where $E[n]$ is the energy in the waveform under the sliding window, i.e., $E[\tau] = \sum_{n=-\infty}^{\infty} |x[n, \tau]|^2$, that rises and falls as the window slides across the waveform. In this case, where the window $w[n, \tau]$ is short, and less than a pitch period, the spectrogram shows the formants of the vocal tract in frequency, but also gives vertical striations in time every pitch period, rather than the harmonic horizontal striations as in the narrowband spectrogram. These vertical striations arise because the short window is sliding through fluctuating energy regions of the speech waveform.

In our description of the narrowband and wideband spectrograms, we have used the example of voiced speech. Similar reasoning can be made for fricative and plosive sounds. With regard to fricatives, the squared STFT magnitude of noise sounds is often referred to as the *periodogram*, which is characterized by random wiggles around the underlying function $|\tilde{H}(\omega)|^2$. The periodogram is developed formally in a stochastic process framework later in the text. For plosives, the spectrogram reveals the general spectral structure of the sound as the window $w[n, \tau]$ slides across the signal. For these sound classes, both the narrowband and wideband spectrograms show greater intensity at formants of the vocal tract; neither, however, typically shows horizontal or vertical pitch-related striations because periodicity is not present except when the vocal folds are vibrating simultaneously with these noise or impulsive sounds.

With plosive sounds, the wideband spectrogram is often preferred because it gives better temporal resolution of the sound’s components, especially when the plosive is closely surrounded by vowels.

Figure 3.15 compares the narrowband (20-ms Hamming window) and wideband (4-ms Hamming window) spectrograms for a particular utterance. The spectrograms were computed with a 512-point FFT. For the narrowband spectrogram, the 20-ms Hamming window was shifted at a 5-ms frame interval, and for the wideband spectrogram, the 4-ms Hamming window was shifted at a 1-ms frame interval. Both spectrograms reveal the speech spectral envelope $|\tilde{H}(\omega)| = |H(\omega)G(\omega)|$ consisting of the vocal tract formant and glottal contributions. Notice, however, the distinctive horizontal and vertical striations in the narrowband and wideband spectrograms, respectively. Observe, however, that occasionally the vertical striations are barely visible in the wideband spectrogram when the pitch is very high. Observe also a difference in
3.4 Categorization of Speech Sounds

In Section 3.2, we described the anatomy of speech production, the vocal folds and vocal tract being the two primary components, and described the mechanism of speech production, i.e., how we generate sounds with our speech anatomy and physiology. We saw that a sound source can be created with either the vocal folds or with a constriction in the vocal tract, and, based on

gain time and frequency resolution between the two spectrograms; for example, the short-time spectrum of the short-duration speech sound “t” in the words “tea” and “to,” across time, is blurry in the narrowband spectrogram while sharp in the wideband spectrogram. Figure 3.16 gives a similar comparison for an utterance that transitions from normal voicing into diplophonic voicing as the pitch becomes very low. In this case, the pitch is so low that horizontal striations are barely visible in the narrowband spectrogram, in spite of an increased window length of 40-ms to improve resolution of harmonic lines. In the wideband spectrogram, one clearly sees vertical striations corresponding to both the primary glottal pulses and secondary diplophonic pulses.

Figure 3.15 Comparison of measured spectrograms for the utterance, “Which tea party did Baker go to?”: (a) speech waveform; (b) wideband spectrogram; (c) narrowband spectrogram.
the various sound sources, we proposed a general categorization of speech sounds. Section 3.3 then deviated from the flow of this chapter to describe spectrographic analysis for the study of time-varying spectral characteristics of speech. We are now in a position to study and classify speech sounds from the following different perspectives:

1. The nature of the source: periodic, noisy, or impulsive, and combinations of the three;

2. The shape of the vocal tract. The shape is described primarily with respect to the place of the tongue hump along the oral tract and the degree of the constriction of the hump, sometimes referred to as the *place and manner-of-articulation*, respectively. The shape of the vocal tract is also determined by possible connection to the nasal passage by way of the velum;

3. The time-domain waveform which gives the pressure change with time at the lips output;

4. The time-varying spectral characteristics revealed through the spectrogram.
With these four speech descriptors, we embark on a brief study of the classification of speech sounds. We focus on the English language, but from time to time discuss characteristics of other languages.

### 3.4.1 Elements of a Language

A fundamental distinctive unit of a language is the phoneme; the phoneme is distinctive in the sense that it is a speech sound class that differentiates words of a language [29]. For example, the words “cat,” “bat,” and “hat” consist of three speech sounds, the first of which gives each word its distinctive meaning, being from different phoneme classes. We saw earlier, and we will discuss further below, that many sounds provide this distinctive meaning, and such sounds represent a particular phoneme. To emphasize the distinction between the concept of a phoneme and sounds that convey a phoneme, the speech scientist uses the term *phone* to mean a particular instantiation of a phoneme. As we discussed in this chapter’s introduction, this distinction is also seen in the different studies of phonemics and phonetics.

Different languages contain different phoneme sets. Syllables contain one or more phonemes, while words are formed with one or more syllables, concatenated to form phrases and sentences. Linguistics is the study of the arrangement of speech sounds, i.e., phonemes and the larger speech units built from phonemes, according to the rules of a language. Phonemes can differ across languages, but certain properties of the grammatical rules combining phonemes and larger units of a language may be common and instinctual [30]. There are various ways to study speech sounds that make up phoneme classes; the use of the above first two descriptors in this study is sometimes referred to as *articulatory phonetics*, while using the last two is referred to as *acoustic phonetics*. One broad phoneme classification for English is in terms of vowels, consonants, diphthongs, affricates, and semi-vowels. Figure 3.17 shows this classification, along with various subgroups, where each phoneme symbol is written within slashes according to both the International Phonetic Alphabet and an orthographic (alphabetic spelling) representation. An insightful history of the various phoneme symbol representations is described in [6]. In the remainder of this text, we use the orthographic symbols.

Phonemes arise from a combination of vocal fold and vocal tract articulatory *features*. Articulatory features, corresponding to the first two descriptors above, include the vocal fold state, i.e., whether the vocal folds are vibrating or open; the tongue position and height, i.e., whether it is in the front, central, or back along the palate and whether its constriction is partial or complete; and the velum state, i.e., whether a sound is nasal or not. It has been hypothesized that the first step in the production of a phone is to conceive in the brain the set of articulatory features that correspond to a phoneme. A particular set of speech muscles is responsible for “activating” each feature with certain relative timing. It is these features that we may store in our brain for the representation of a phoneme. In English, the combinations of features are such to give 40 phonemes, while in other languages the features can yield a smaller—e.g., 11 in Polynesian, or a larger, e.g., 141 in the “click” language of Khosian—phoneme set [30]. The rules of a language string together its phonemes in a particular order; for example, in Italian,

---

9 A click used in the Khosian language is made by the lips and tongue body and with air drawn into the oral tract. The positions of the lips and tongue are features of the language that combine with other features, such as whether the vocal folds are vibrating or not, to form the Khosian phoneme set.
consonants are not normally allowed at the end of words. The ordering of the phonemes is also determined in part by the underlying articulatory features of the phones; for example, vibration of the vocal folds or a particular vocal tract shape can constrain or influence the following sound.

A phoneme is not strictly defined by the precise adjustment of articulators; for example, the tongue hump forming a 0.1-mm constriction with the palate, 3 cm along the oral tract, will likely correspond to the same phoneme when these specifications are changed by a few percent. The articulatory properties are influenced by adjacent phonemes, rate and emphasis in speaking, and the time-varying nature of the articulators. The variants of sounds, or phones, that convey the same phoneme are called the allophones of the phoneme [29]. Consider, for example, the words “butter,” “but,” and “to,” where the /t/ in each word is somewhat different with respect to articulation, being influenced by its position within the word. Therefore, although the allophones of a phoneme do have consistent articulatory features, the fine details of these features vary in different conditions. In this sense, then, the concept of a phoneme as a distinctive unit of a language is abstract.
In speech production, the articulatory features ultimately lead to the speech waveform and its acoustic temporal and spectral features, corresponding to the above third and fourth descriptors, such as the time delay of a plosive before a voiced sound and vocal tract formants. In the motor theory of perception [3], such acoustic properties are measured by the auditory system and ultimately are mapped in the brain to the set of articulatory features that define the phoneme, i.e., in perceiving the phoneme the listener reconstructs the set of articulatory features for that phoneme. Later in this chapter, we return to this paradigm, as well as to a different view where articulatory features are not the end perceptual representation. We now begin a short study of the classification of speech sounds, using both articulatory and acoustic characterizations. For each phoneme class, we describe source and system (vocal tract) articulators, and the resulting spectral and waveform characteristics that give a phoneme its distinction.

3.4.2 Vowels

The largest phoneme group is that of vowels. Vowels contain three subgroups defined by the tongue hump being along the front, central, or back part of the palate.

**Source:** The source is quasi-periodic puffs of airflow through the vocal folds vibrating at a certain fundamental frequency. We use the term “quasi” because perfect periodicity is never achieved; henceforth, the term “periodic” will be used in this sense. A simple model of the source waveform and spectrum and its modification by the vocal tract was given in Examples 3.1 and 3.2. In English, the pitch of the periodic source does not distinguish phonemes as in some languages such as Chinese.

**System:** Each vowel phoneme corresponds to a different vocal tract configuration. The vocal tract shape is a function of the tongue, the jaw, the lips, and the velum which is closed in non-nasalized vowels, i.e., the nasal passage is not coupled to the oral tract. In addition to their degree of openness, the lips can contribute to the vocal tract configuration by being rounded, which can increase the effective vocal tract length. Recite the phoneme /u/ in the word “boot” and you will feel the lips become rounded and protruded. The tongue, which is the primary determinant of vocal tract shape, has three general places of articulation: front, center, or back of the oral cavity. The degree of constriction by the tongue is another shape determinant. A comparative example is given with the vowel /a/ as in “father” and with the vowel /i/ as in “eye” [32]. For the vowel /a/ the vocal tract is open at the front, the tongue is raised at the back, and there is a low degree of constriction by the tongue against the palate. For the vowel /i/ the vocal tract is open at the back, the tongue is raised at the front, and there is a high degree of constriction of the tongue against the palate. These examples are included in Figure 3.18, which illustrates the vocal tract profiles for all English vowels in terms of tongue position and degree of constriction [31]. Keep in mind that Figure 3.18 shows the oral cavity and does not include the pharynx, the region just above the glottis, which can also influence formant locations. X-ray studies of the complete vocal tract for different phonemes are found in the early work of Fant [8], as well as in more recent magnetic resonance imaging studies [35].

**Spectrogram:** The particular shape of the vocal tract determines its resonances. Qualitative rules based on physical principles have been developed by Stevens [33] for mapping changes in vocal tract shape to formant movement. Perturbations in cross-section at various points of a uniform reference tube (approximately modeling the vowel /A/), by narrowing of the front,
Figure 3.18 Vocal tract profiles for vowels in American English. The two horizontal lines denote voicing.


central, or back of the oral cavity by the tongue and jaws, are mapped to certain changes in formant location. In Chapter 4, we will study quantitatively the relation between vocal tract shape and formants using a concatenated acoustic tube model. The wideband spectrograms and spectral slices of the narrowband spectrograms of the two vowels /a/ and /i/ are shown in Figure 3.19. The first formant of /a/ is dominant and falls at roughly 800 Hz, while the second and third weaker formants are at roughly 1200 Hz and 2300 Hz, respectively. For the vowel /i/, the first formant is at about 400 Hz and the second and third formants are at about 2000 Hz and 3000 Hz, respectively, with the third being greater in amplitude than its counterpart in /a/. The wideband spectrograms in these and following examples are obtained with a 4-ms window and a 1-ms frame interval. The narrowband spectral slices are obtained with a 20-ms and 30-ms window for the /a/ and /i/, respectively, and a 5-ms frame interval.

Waveform: Certain vowel properties seen in the spectrogram are also seen in the speech waveform within a pitch period. As illustrated in Figure 3.19, for the vowel /a/ the dominant first
3.4 Categorization of Speech Sounds

Figure 3.19 Waveform, wideband spectrogram, and spectral slice of narrowband spectrogram for two vowels: (a) /i/ as in “eye”; (b) /a/ as in “father.” The first three formants $F_1$, $F_2$, and $F_3$ are marked on the spectral slices.

formant gives a low-frequency damped oscillation while the second and third weaker formants give no visible high-frequency energy. In contrast, for the vowel /i/, the first formant gives a very low-frequency damped oscillation and the third strong formant gives a visible high-frequency oscillation superimposed on the low-frequency formant.

In spite of the specific properties of different vowels, there is much variability of vowel characteristics among speakers. We noted earlier that articulatory differences in speakers is one cause for allophonic variations. The place and degree of constriction of the tongue hump and cross-section and length of the vocal tract, and therefore the vocal tract formants, will vary with the speaker. Peterson and Barney [28],[32] measured the first ($F_1$) and second ($F_2$) formants from a spectrogram for a large range of speakers. Vowels deemed to be “perceptually equivalent” were used. A plot of $F_1$ and $F_2$ on a 2-D grid reveals approximate elliptical clusters corresponding to the different vowels and shows a large range of variation in $F_1$ and $F_2$ for each vowel group. This variability presents a challenge to speech recognition algorithms that
rly on invariance of vowel spectral properties across speaker, but aids in speaker recognition where spectral variability with speaker is required.

### 3.4.3 Nasals

The second large phoneme grouping is that of consonants. The consonants contain a number of subgroups: nasals, fricatives, plosives, whispers, and affricates. We begin with the nasals since they are closest to the vowels.

**Source:** As with vowels, the source is quasi-periodic airflow puffs from the vibrating vocal folds.

**System:** The velum is lowered and the air flows mainly through the nasal cavity, the oral tract being constricted; thus sound is radiated at the nostrils. The nasal consonants are distinguished by the place along the oral tract at which the tongue makes a constriction (Figure 3.20). The two nasals that we compare are /m/ as in “mo” and /n/ as “no.” For /m/, the oral tract constriction occurs at the lips and for /n/ the constriction is with the tongue to the gum ridge.

**Spectrogram:** The spectrum of a nasal is dominated by the low resonance of the large volume of the nasal cavity. The resonances of the nasal cavity have a large bandwidth because viscous losses are high as air flows along its complexly configured surface, quickly damping its impulse response. The closed oral cavity acts as a side branch with its own resonances that change with the place of constriction of the tongue; these resonances absorb acoustic energy and thus are anti-resonances (zeros) of the vocal tract. The anti-resonances of the oral tract tend to lie beyond the low-frequency resonances of the nasal tract; a result of this is that for nasals there is little high-frequency energy passed by the vocal tract transfer function. For the /m/ in Figure 3.21b, there is a low $F_1$ at about 250 Hz with little energy above this frequency. A similar pattern is seen for the /n/ in Figure 3.21a. Observe that at the release of the constriction of the nasal there is an abrupt change in the spectrogram when the sound is radiated from the mouth. The formant transitions that follow the release are quite different for the nasals /m/ and /n/; these transitions, which reflect the manner in which the oral cavity transitions into its steady vowel position, are an important perceptually distinguishing characteristic of the two nasals [33].

![Figure 3.20](image-url) Vocal tract configurations for nasal consonants. Oral tract constrictions occur at the lips for /m/, with the tongue tip to the gum ridge for /n/, and with the tongue body against the palate near the velum for /ng/. Horizontal lines denote voicing.

3.4 Categorization of Speech Sounds

![Waveform](image)

**Waveform:** The waveforms for both the /m/ and /n/ are dominated by the low, wide-bandwidth $F_1$ formant; within each glottal cycle, there is seen a rapidly damped oscillation. Other resonances are not high enough in energy to be seen.

A phenomenon we referred to earlier as nasalization of vowels is related to the generation of nasals in that the velum is partially open. The nasal cavity couples with the oral cavity and introduces anti-resonances (zeros) into the vocal tract system function. The open nasal cavity acts as a side chamber that introduces the anti-resonances by absorbing energy at certain frequencies, thus serving the same role as the oral tract for a nasal consonant. There is also some broadening of the bandwidth of the resonances of the oral cavity due to the oral-nasal tract coupling. In vowel nasalization, the speech sound is primarily due to the sound at the lips and not the sound at the nose output, which is very weak. Vowels adjacent to nasal consonants tend to be nasalized. Certain speakers characteristically nasalize their vowels by keeping their velum partially open. In English, unlike some languages such as French, Polish, and Portuguese, vowel nasalization is not used to differentiate phonemes [30].

3.4.4 Fricatives

Fricative consonants are specified in two classes: voiced and unvoiced fricatives.

**Source:** In unvoiced fricatives, the vocal folds are relaxed and not vibrating. Noise is generated by turbulent airflow at some point of constriction along the oral tract, a constriction that is narrower than with vowels. The degree of the constriction somewhat colors the spectral character of the noise source, although this is a secondary effect, the vocal tract spectral coloring being primary.
System: The location of the constriction by the tongue at the back, center, or front of the oral tract, as well as at the teeth or lips, influences which fricative sound is produced. The constriction separates the oral tract into front and back cavities with the sound radiated from the front cavity. Although the front cavity dominates the spectral shaping of the sound, the back cavity introduces anti-resonances in the transfer function, absorbing energy at approximately its own resonances. Because the front cavity is shorter than the full oral cavity and because anti-resonances of the back cavity tend to be lower in frequency than the resonances of the front cavity, the resulting transfer function consists primarily of high-frequency resonances which change with the location of the constriction.

Voiced fricatives have a similar noise source and system characteristic to unvoiced fricatives; for voiced fricatives, however, the vocal folds often vibrate simultaneously with noise generation at the constriction and a periodicity of the noisy airflow is seen. Recite the voiced fricative /z/, as in “zebra,” and you will feel the vocal folds vibrating while noise is generated. Generally, fricatives occur in voiced/unvoiced pairs. We compare the unvoiced fricative /f/ as in “for” and the matching voiced fricative /v/ as in “vote.” In /f/, the vocal folds are not vibrating and the constriction occurs by the teeth against the lips. In contrast, for /v/ the vocal folds are vibrating and again the constriction is formed by the teeth against the lips (Figure 3.22).

When the vocal folds vibrate in a voiced fricative, the periodic airflow from the glottis passes through the back oral cavity to the constriction. At the constriction, frication takes place only when the airflow velocity of the periodic puffs is “high enough.” According to fluid dynamical principles, the airflow velocity must exceed a constant called the Reynolds number, which is a function of the density and viscosity of the air medium as well as the geometry of the constriction [20]. This implies that frication is approximately synchronized with airflow velocity. The glottal waveform shape therefore can be thought of as modulating a noise source. This leads to a simplified model of voiced frication given in the following example:

**Example 3.4** A voiced fricative is generated with both a periodic and noise source. The periodic glottal flow component can be expressed as

\[ u[n] = g[n] \ast p[n] \]

where \( g[n] \) is the glottal flow over one cycle and \( p[n] \) is an impulse train with pitch period \( P \). In a simplified model of a voiced fricative, the periodic signal component \( u[n] \) is passed through a linear time-invariant vocal tract with impulse response \( h[n] \). The output at the lips due to the periodic glottal source is given by

\[ x_g[n] = h[n] \ast (g[n] \ast p[n]). \]

In the model of the noise source component of the voiced fricative, the vocal tract is constricted along the oral tract and air flows through the constriction, resulting in a turbulent airflow velocity source at the constriction that we denote by \( q[n] \). In this simplified model, the glottal flow \( u[n] \) modulates this noise function \( q[n] \) (assumed white noise). The modulated noise then excites the front oral cavity that has impulse response \( h_f[n] \). The output flow at the lips due to the noise source is expressed as

\[ x_q[n] = h_f[n] \ast (q[n]u[n]). \]
We assume in our simple model that the results of the two airflow sources add, so that the complete output of the lips is given by

\[
x[n] = x_g[n] + x_q[n]
= h[n] * u[n] + h_f[n] * (q[n]u[n]).
\]

The spectral characteristics of \( x[n] \) are studied in Exercise 3.10.

In this simple model, we have ignored that the modulating function \( u[n] \) is modified by the oral cavity and that the noise response \( x_q[n] \) can be influenced by the back cavity. We have also not accounted for sources from nonlinear effects other than the modulation process, one possibility being distributed sources due to traveling vortices.
In voiced fricatives, however, voicing does not always occur simultaneously with noise generation. Simultaneous voicing may occur only early on or not at all during the frication. A voiced fricative can also be distinguished from its unvoiced counterpart by a shorter duration of frication prior to the onset of voicing in a following vowel. The timing of the onset of voicing after frication thus provides a cue in the distinction of these sounds. The formant transitions from the frication into the following vowel also serve to distinguish between voiced and unvoiced fricative counterparts; for voiced fricatives, the voicing occurs sooner into the transition, thus accentuating the transition relative to the weaker noise excitation of its unvoiced counterpart during the initial part of the transition.\(^{10}\) Generally then, there are multiple cues that help in these distinctions.

**Spectrogram:** Unvoiced fricatives are characterized by a “noisy” spectrum while voiced fricatives often show both noise and harmonics. The spectral nature of the sound is determined by the location of the tongue constriction. For example, with an /S/ the frication occurs at the palate, and with an /f/ at the lips. The /S/ has a highpass spectrum corresponding to a short upper oral cavity. For the /f/ there is little front cavity, so its spectrum is almost flat with a mild upward trend. A comparison of the unvoiced fricative /f/ with the voiced fricative /v/ is given in Figure 3.23. The noise component of each fricative has a wide spectrum focused in the high-frequency region (1000–5000 Hz). The voiced fricative, however, is characterized by the additional harmonic structure due to the oscillating vocal folds, as revealed in the spectral slices as well as the spectrograms of Figure 3.23. The influence of the surrounding vowels on the formant transitions to and from the fricatives can also be seen in the spectrograms.

**Waveform:** For the unvoiced/voiced fricative pair, the waveform of the unvoiced fricative contains noise, while that of the voiced fricative contains noise superimposed on periodicity during the fricative region, as seen in Figure 3.23.

**Whisper:** Although the whisper is a consonant similar in formation to the unvoiced fricative, we place the whisper in its own consonantal class. We saw earlier that with a whisper the glottis is open and there is no vocal fold vibration. Turbulent flow is produced, however, at the glottis, rather than at a vocal tract constriction. The spectral characteristics of the whisper depend on the size of the glottis, which influences the spectrum of the noise source, and the resonant cavity at the onset of the vowel. An example is /h/, the sole whisper in English, as in “he.” Other whispers exist outside the English language.

### 3.4.5 Plosives

As with fricatives, plosives are both unvoiced and voiced.

**Source and System:** With unvoiced plosives, a “burst” is generated at the release of the buildup of pressure behind a total constriction in the oral tract. We have idealized this burst as an impulsive source, although in practice there is a time spread and turbulent component to this source. The constriction can occur at the front, center, or back of the oral tract (Figure 3.24). There is no vibration of the folds. The sequence of events is: (1) complete closure of the oral tract and buildup of air pressure behind closure; during this time of closure, no sound is radiated from the lips; (2) release of air pressure and generation of turbulence over a very short-time duration.

\(^{10}\) Within an unvoiced/voiced consonant pair, the formant transitions are similar but differ in their excitation. Across consonant pairs, on the other hand, differences in formant transitions from a consonant to a following vowel help in the distinction of consonants.
i.e., the burst ("impulsive") source, which excites the oral cavity in front of the constriction; (3) generation of aspiration due to turbulence at the open vocal folds (before onset of vibration) as air rushes through the open oral cavity after the burst; and (4) onset of the following vowel about 40–50 ms after the burst. The voiced onset time is the difference between the time of the burst and the onset of voicing in the following vowel. The length of the voice onset time and the place of constriction vary with the plosive consonant.

With voiced plosives, as with unvoiced plosives, there is a buildup of pressure behind an oral tract constriction, but the vocal folds can also vibrate. When this vibration occurs, although the oral tract is closed, we hear a low-frequency vibration due to its propagation through the walls of the throat. This activity is referred to as a "voice bar."\(^{11}\) After the release of the burst, unlike the unvoiced plosive, there is little or no aspiration, and the vocal folds continue to vibrate

---

\(^{11}\) Voice bars are exploited by other species besides humans. Male frogs emit a mating call by forcing air from their lungs through vocal folds into the mouth and nostrils. During this maneuver the mouth and nostrils are closed tightly and a thin wall sac at the base of the mouth is blown up like a balloon. The vibrating vocal folds propagate sound to the sac, which radiates the mating call into the external air [26].
into the following vowel; there is much shorter delay between the burst and the voicing of the vowel onset. Figure 3.25 compares an abstraction of the unvoiced plosive and a voiced plosive with the presence of a voice bar. In Figure 3.26, we compare an actual voiced/unvoiced plosive pair: /g/ as in “go” and /k/ as in “baker.” The phoneme /k/ is unvoiced with a constriction at the velum; the phoneme /g/ is characterized by the same constriction, but with vibrating folds.

Vocal fold vibration, and thus a voice bar, does not always occur during the burst of a voiced plosive, and therefore is not the only distinguishing feature between the two sound classes. The length of the voice onset time can also provide the distinction. Perceptual experiments indicate that if the release of the burst and onset of voicing are within 20 ms of each other, the consonant is considered voiced; otherwise, it is judged as unvoiced. The muscles controlling the vocal fold positions and tension are very precisely activated in time to generate the different voice onset times [10]. The formant transitions from the onset of the burst into the following vowel also help in distinguishing between voiced and unvoiced plosive counterparts; as with voiced fricatives, the voicing in voiced plosives occurs sooner into the transition, thus accentuating the transition relative to the weaker aspiration excitation of its unvoiced counterpart.

**Spectrogram and Waveform:** For the unvoiced plosive /k/, we observe (Figure 3.26) in both the waveform and the spectrogram a gap of near silence, followed by an abrupt burst, and then aspiration noise. The spectrogram at the time of the burst is governed by the shape of the oral cavity in front of the constriction (which is excited by the burst) and the spectral character of the burst itself. The aspiration then acts as an excitation to the vocal tract as it transitions from its constricted state. Finally, when the vocal folds vibrate, the resulting periodic puffs excite the vocal tract as it enters into its steady vowel state. The formant trajectories through aspiration and into voicing reflect this changing vocal tract shape. For the counterpart voiced plosive /g/, a low-frequency voice bar is seen in the spectrogram prior to the burst. Observe also that the voice onset
3.4 Categorization of Speech Sounds

Figure 3.25  A schematic representation of (a) unvoiced and (b) voiced plosives. The voiced onset time is denoted by VOT.

Figure 3.26  Waveform, wideband spectrogram, and narrowband spectral slice of voiced and unvoiced plosive pair: (a) /g/ as in “go”; (b) /k/ as in “key.” Spectral slices are taken in burst regions over a 40-ms window in (a) and a 25-ms window in (b).
time is negligible for the /g/ in contrast to that for the /k/. Given a negligible aspiration stage, the formant trajectories sometimes appear more visible, being excited throughout by voicing. As with voiced fricatives, vocal fold vibration need not be present for a plosive to be “voiced”; the voice onset time and formant transitions can be sufficient to make the distinction.

In the following example, we explore a time-varying linear system model for the voiced plosive.

**Example 3.5** We have seen that a voiced plosive is generated with a burst source and can also have present a periodic source throughout the burst and into the following vowel. Assuming the burst occurs at time \( n = 0 \), we idealize the burst source as an impulse \( \delta[n] \). The glottal flow velocity model for the periodic source component is given by

\[
u[n] = g[n] * p[n]
\]

where \( g[n] \) is the glottal flow over one cycle and \( p[n] \) is an impulse train with spacing \( P \). Assume that the vocal tract is linear but time-varying, due to the changing vocal tract shape during its transition from the burst to a following steady vowel. The vocal tract output cannot, therefore, be obtained by the convolutional operator. Rather, the vocal tract output can be expressed using the time-varying impulse response concept introduced in Chapter 2. In our simple model, the periodic glottal flow excites a time-varying vocal tract, with impulse response denoted by \( h[n, m] \), while the burst excites a time-varying front cavity beyond a constriction, denoted by \( h_f[n, m] \). The sequences \( h[n, m] \) and \( h_f[n, m] \), as we described in Chapter 2, represent time-varying impulse responses at time \( n \) to a unit sample applied \( m \) samples earlier at time \( n - m \). We can then write the output with a generalization of the convolution operator as

\[
x[n] = \sum_{m=-\infty}^{\infty} h[n, m]u[n - m] + \sum_{m=-\infty}^{\infty} h_f[n, m]\delta[n - m]
\]

where we assume the two outputs can be linearly combined. In this model, we have not accounted for aspiration that might occur before the onset of voicing, as well as other effects described in Example 3.4.

### 3.4.6 Transitional Speech Sounds

A number of the phone examples we have presented thus far were illustrated as “stationary,” such as the sounds for the phonemes /a/ and /i/ in Figure 3.19, in the sense that their underlying articulators hold an almost fixed configuration and the resulting formants appear nearly constant. We also illustrated the movement of one phone to the next with articulators transiting from one configuration to another, for example, as manifested in the spectral change in going from the /g/ to /o/ in “go” of Figure 3.26. Such speech sounds are “nonstationary” and in some cases the rapid transition across two articulatory states defines the sound. This time-varying nature of articulators is more the norm than the exception. The articulators are almost always in transition between states, often never reaching a desired state in typically spoken speech. Nonstationarity is further imparted through a phenomenon known as **coarticulation**, which involves the anticipation...
of the following sound, sometimes many phones in the future, and therefore a blending of the
articulatory states. Such nonstationarity poses interesting challenges to speech signal processing
algorithms that typically assume (and often require) stationarity over intervals of 10–20 ms.

**Phonemes Defined by Transition — Diphthongs:** Diphthongs have a vowel-like nature with
vibrating vocal folds. Diphthongs, however, cannot be generated with the vocal tract in a steady
configuration; they are produced by varying in time the vocal tract smoothly between two vowel
configurations and are characterized by a *movement* from one vowel *target* to another. The term
target, as implied, refers to seeking a particular vocal tract configuration for a phoneme, but not
necessarily achieving that configuration. Four diphthongs in the American English language are:
/Y/ as in “híde,” /W/ as in “out,” /O/ as in “boy,” and /JU/ as in “new.” As an example, Figure
3.27 shows the spectrogram of the diphthong /O/ in “boy,” where we see a rapid movement of
the formants (especially $F_2$) as the vocal tract changes from its configuration for the /u/ to that
for the /i/ vowel sound. Such formant transitions characterize the diphthong.

**Semi-Vowels:** This class is also vowel-like in nature with vibrating folds. There are two
categories of semi-vowels: glides (/w/ as in “we” and /y/ as in “you”) and liquids (/l/ as in “read”
and /r/ as in “jet”). The glides are dynamic and transitional and often occur before a vowel or
are surrounded by vowels, passing from a preceding vowel and moving toward the following
vowel, thus in this latter case being similar to diphthongs. Glides differ from diphthongs in two
ways. In glides, the constriction of the oral tract is greater during the transition and the speed
of the oral tract movement is faster than for diphthongs. These articulations result in weaker,
but faster formant transitions [29]. An example is /w/ in the phrase “away.” At the onset of
the /w/, the tongue hump is high and the lips are highly constricted. The formants of the /w/
“glide” rapidly from the preceding /a/ to the following /e/. The liquids /l/ and /r/ are similar
to the glides in being voiced, in the speed of movement, and in the degree of oral constriction.
The liquids differ from the glides in the formation of the constriction by the tongue; the tongue

![Figure 3.27](image.png)

*Figure 3.27*  Narrowband spectrogram example for the diphthong /O/ in “boy.”
is shaped in such a way as to form side branches, different from all other oral tract formations thus far [33]. The presence of these side branches can introduce anti-resonances. With /l/ the tongue tip is pressed against the palate, and is shaped to form two openings, one on each side of the contact point with air passing on each side. For the liquid /r/, two different tongue formations are known to exist, one with a tongue tip raised at the palate as shown in Figure 3.28 and creating a space under it, and the other with the tongue tip lowered and the tongue hump bunched near the palate forming a side branch to the primary oral cavity [7],[33]. In addition, the liquids are typically characterized by an effective constriction of the pharynx, quite unusual for American English consonants; magnetic resonance image vocal tract profiles have shown such constrictions [7],[24]. The formant transitions differ for all four semi-vowels; nevertheless, these transitions are smooth except for the liquid /l/, whose spectral evolution is discontinuous because the tongue tip holds and releases its contact with the palate.

**Affricates:** This sound is a counterpart of diphthongs, consisting of consonant plosive-fricative combinations, rapidly transitng from plosives to fricatives. The articulation of affricates is similar to that of fricatives. The difference is that for affricates, the fricative is preceded by a complete constriction of the oral cavity, formed at the same place as for the plosive. An example is the /tS/ in the word “chew,” which is the plosive /t/ followed by the fricative /S/. The voiced counterpart to /tS/ is the affricate /J/ as in “just,” which is the voiced plosive /d/ followed by the voiced fricative /Z/.

**Coarticulation** — Although our vocal fold/vocal tract muscles are programmed to seek a target state or shape, often the target is never reached. Our speech anatomy cannot move to a desired position instantaneously and thus past positions influence the present. Furthermore, to make anatomical movement easy and graceful, the brain anticipates the future, and so the articulators at any time instant are influenced by where they have been and where they are going. **Coarticulation** refers to the influence of the articulation of one sound on the articulation of another sound in the same utterance and can occur on different temporal levels. In “local” coarticulation, articulation of a phoneme is influenced by its adjacent neighbors or by neighbors close in time. Consider,
3.5 Prosody: The Melody of Speech

Long-time variations, i.e., changes extending over more than one phoneme, in pitch (intonation), amplitude (loudness), and timing (articulation rate or rhythm) follow what are referred to as the rules of prosody of a language. These rules are followed to convey different meaning, stress, and emotion. For example, the average pitch contour of a declarative statement tends to increase at the onset of an utterance, decline slowly, then make a rapid drop at the end of the utterance. For a question, on the other hand, the pitch contour increases over time relative to that of the same statement made in declarative form. In addition, a speaker also adds his/her own speaker-dependent prosodic refinements such as a characteristic articulation rate and pitch fluctuation.

Consider changes made in the production mechanism and prosody during the stress of a speech sound. The subglottal pressure, i.e., the pressure just below the glottis, during a normally spoken utterance is typically about constant, but falling near the termination of the

---

13 Short pieces of a phrase can sometimes be replaced by another sound without a change in intelligibility by the listener as, for example, when replacing certain short speech segments by noise. This result may reflect phonological redundancy by the speaker, but also expectation by the listener. It is also interesting to note that, in addition to temporal redundancy, there is evidence of frequency redundancy in that certain spectral bands can be eliminated with little effect on intelligibility of a sound [22].
utterance where there is also often a corresponding decline in pitch. We can impart, however, variations on this average subglottal pressure. When adding stress to a speech sound, we increase pressure within the lungs. The resulting increased subglottal pressure causes an increase in pitch and loudness. This increased pressure will also cause a more abrupt glottal closure by accentuating the Bernoulli effect on airflow through the glottis. A more abrupt closure of the glottis, as seen in Example 3.1, corresponds to more energy in high frequencies of the glottal spectrum, and thus a “harder” voice quality which can also characterize a stressed sound. An increase in pitch is also imparted due to an increase in vocal fold tension. A rise in subglottal pressure and in vocal fold tension is also responsible for a rising pitch over the duration of a question. We can impart these variations globally, as in going from a soft to a shouting voice in argument, or more locally in time to stress or change intonation on a particular syllable, word, or phrase.

**Example 3.6** Figure 3.29 shows a comparison of the utterance, “Please do this today,” which in the first case is spoken normally, while in the second case, the word “today” is stressed. The pitch, spectral content, and loudness are all seen to be generally higher in the stressed case for the word “today” (Exercise 3.13). The reader should speak the utterance in the two voice styles. Feel your lungs contract during the stressing of the word “today.” The increase in lung contraction causes an increase in subglottal pressure. Another change in stressing a sound is its increase in duration as seen in the waveforms and spectrograms of the word “today” in Figure 3.29. Although the speaker planned to stress only the word “today,” the entire sentence is longer, likely due to the anticipation of stressing the last word of the sentence. Stressing a sound is generally accompanied by an increase in the duration of vowels and consonants. For a vowel, the number of pitch periods increases, as we can see in Figure 3.29. For consonants, the time of frication or time prior to closure of a constriction, thus building up greater pressure, will increase. For example, for the /t/ of the stressed word “today,” the preceding silence gap is longer and the burst louder than in its normally spoken counterpart.

We saw in Example 3.6 that an increase in duration is also characteristic of a stressed sound. One theory proposed by Lieberman [21] suggests that subglottal pressure, vocal fold tension, and phoneme duration are set within a time period reference which is the duration of an exhalation. We earlier called the collection of word and phrase groups within a single exhalation a breath group, and saw that the duration of a breath group affects the coarticulation of phonemes. Lieberman’s hypothesis states that, in addition, the durational and timing patterns of phonemes are also determined by their degree of stress within the breath group. The muscles of the larynx that control vocal fold tension and the muscles around the lungs that control subglottal pressure are timed in their contraction relative to the breath group duration and the sounds marked for stress and pitch intonation within the breath group.

We have seen thus far two factors affecting the duration of vowels and consonants: coarticulation and stress within a breath group. There are other factors as well. Vowels are typically longer than consonants. One hypothesis for this difference is that vowels, in addition to relaying intelligibility of the message, by being typically longer carry prosodic components such as rhythm and intonation, while consonants, by being shorter, carry more of the information load.
3.5 Prosody: The Melody of Speech

Figure 3.29 Comparison of “Please do this today,” where “today” is spoken in a normal and stressed style: (a) waveform of normal; (b) waveform of stressed; (c)–(d) spectrograms of (a)–(b).

Although consonants are often shorter than vowels, time is still required to process them by the perceptual system. This processing time may be provided when a vowel follows a consonant. Duration is also affected by other factors, such as position and importance of a word within a breath group or the relative position of vowels and consonants; e.g., a vowel is shorter when followed by a voiceless consonant than a voiced consonant. An understanding of these timing patterns is important because they indicate how speech is temporally organized in the brain and motor system. A better understanding of timing also has the practical benefit of improved speech synthesis techniques.

Related to the timing of speech events is the rate of articulation of a speaker, a topic we will revisit throughout the text when we describe signal processing systems that modify or measure
articulation rate for a variety of applications. Changing articulation rate compresses or expands sound events in time. Pickett [29] has done some extensive studies on the degree of change in vowels and consonants in speaking at a faster or slower rate. Figure 3.30 illustrates spectrograms

**Figure 3.30** Spectrograms of utterance “anticipation of downstream articulators” at different rates of articulation (slow to fast) by Pickett [29]. Articulation rates are estimated in syllables per second (s/s). The lines drawn between successive spectrograms are synchronized with the same points within the utterance and show the extent of compression across different articulation rates.

of a successively time-compressed utterance spoken at different rates. The top spectrogram is of the utterance spoken deliberately slowly, the third spectrogram corresponds to the normal rate of speech, and the last to the fastest rate of speech. We would expect the consonants to be typically time-compressed less than the vowels because of greater limits to consonant compressibility. Plosive consonants require complete vocal tract closure and certain timing relations between the burst and onset of voicing, while fricative consonants require a suitable vocal tract constriction and also durational constraints. Pickett looked at the effect of articulation rate change on short and long vowels, and plosive and fricative consonants. Going from the slowest to the fastest rate, both vowels and consonants were compressed by about 33%, more or less independently of the class. On the other hand, in going from normal to the fastest rate, the vowels were compressed by 50% and the consonants by 26%. In this change, the short vowels were compressed by 64% and the long vowels by 43%, while the plosive consonants were compressed by 36% and the fricative consonants by 26%.

3.6 Speech Perception

In this section, we first briefly describe the acoustic properties of speech sounds that are essential for phoneme discrimination by the auditory system, i.e., the acoustic aspects of the speech sound that act as perceptual cues. Our motivation is the goal of preserving such properties in speech signal processing.\footnote{We are also interested in acoustic properties that convey speaker recognizability, some of which were described throughout this chapter.} We then take a glimpse at perceptual models of how the discriminatory features may be measured and processed by the listener.

3.6.1 Acoustic Cues

We are interested in acoustic components of a speech sound used by the listener to correctly perceive the underlying phoneme. In previous sections, we touched upon a number of these acoustic cues. Here we review and elaborate on these cues and give some additional insights. We look first at vowels and then at consonants.

Vowels — Formant frequencies have been determined to be a primary factor in identifying a vowel. We have already seen that the listening experiments of Peterson and Barney approximately map the first two formants ($F_1$ and $F_2$) to vowel identification. Higher formants also have a role in vowel identity [29]. As we will see in Chapter 4, because formant frequencies scale with tract length, we would expect the listener to normalize formant location in doing phoneme recognition. In fact, there is evidence that such normalization is performed making relative formant spacings essential features in vowel identification [36]. Another factor in vowel perception is nasalization, which is cued primarily by the bandwidth increase of the first formant ($F_1$) and the introduction of zeros [16]. As we pointed out earlier, however, unlike other languages, vowel nasalization is not used in the American English language to aid in phoneme discrimination.

Consonants — Consonant identification depends on a number of factors including the formants of the consonant, formant transitions into the formants of the following vowel, the voicing (or unvoicing) of the vocal folds during or near the consonant production, and the relative timing...
of the consonant and the onset of the following vowel. Consider the plosive voiced consonants: /b/, /d/, and /g/ and their unvoiced counterparts /p/, /t/, and /k/. Each plosive consonant is characterized by a spectrum determined by the vocal tract configuration in front of the closure, referred to as the formant locus, and by formant transitions that represent the movement from the locus spectrum to the spectrum of the following vowel configuration [33]. Generally, perception of consonants depends upon the formant locus and formant transitions. Consider, for example, discrimination between /b/ and /d/ followed by the vowel /a/ as in “ba” and “da.” Two perceptual cues in this case are: (1) $F_1$ of the formant locus is lower in /b/ than in /d/ and (2) $F_2$ transitions are upward from /b/ to the following vowel and downward from /d/ to the following vowel [33]. As with plosive consonants, with fricative consonants both the speech spectrum during the frication noise and the transition into the following vowel are perceptual cues for identification.

We noted in a previous section that the time between the initial burst and the vowel onset, i.e., the voice onset time, is an important cue in discriminating voiced versus unvoiced plosive consonants. Consider, for example, a comparison between /t/ and /d/. If /d/ is extracted in the word “do” and the delay between the /d/ and /o/ continuously increased, then the “do” will be perceived as “to” when the voice onset time exceeds about 25 ms [33]. For both plosive and fricative consonants, we also noted previously that the presence (or lack) of voicing also serves as a perceptual cue in the discrimination of voiced versus unvoiced consonants, although its presence is not necessary in this discrimination.

In addition to the direction of the formant transition, another important perceptual cue is the rate of the transition. For example, it is possible to transform the perception of a plosive consonant into a semi-vowel by decreasing the formant transition rate between the plosive burst and the following vowel. Consider, for instance, the phoneme /b/ in the word “be.” The formant transition duration between the /b/ and the following vowel /i/ is about 10 ms. As the transition duration increases beyond about 30 ms the word “be” is transformed into the word “we,” i.e., the plosive consonant /b/ is perceived as the semi-vowel /w/, which is characterized by a slower formant transition rate (Exercise 3.19).

### 3.6.2 Models of Speech Perception

In the previous section, we described some acoustic attributes of speech sounds that allow their perceptual discrimination. We are now interested in the physiological correlates of these attributes, i.e., the mechanism in the brain that detects the acoustic features ultimately leading to the meaning of the message. One paradigm is the high-level motor theory model of speech perception which states that acoustic features map back to articulatory features. For example, formants of a vowel are detected (or estimated) and then unconsciously interpreted as the place of articulation used in the vowel production [3]. Other examples are spectral harmonicity, or lack thereof, corresponding to a voiced or unvoiced source, and the burst or frication noise corresponding to a complete or partial closure of the oral cavity. In essence, according to the motor theory, we create in our brains a picture of the basic (discrete) articulatory movements responsible for the acoustic feature.\(^\text{15}\) We should keep in mind, however, that there exist other

---

\(^\text{15}\) Stevens [34] has proposed that acoustic features tend to be insensitive to small changes in articulatory configurations so that we do not need precise adjustment of these configurations; e.g., a small change in the place of a constriction of a consonant produces little difference in the corresponding formant pattern. In this paradigm, even vowels formed in continuous movement of the vocal tract have this “quantal” nature.
developing theories of speech perception. For example, the observation of large articulatory variability in the phoneme /r/, using magnetic resonance imaging, suggests that speakers and listeners rely on an acoustic representation in production and perception of speech, rather than an articulatory representation [13]. Similar acoustic features for the phoneme /r/ can be generated with very different vocal tract shapes involving different tongue positions and constrictions, as we described in Section 4.6. Such large articulatory variability can depend on phonetic context and has been observed in other phonemes as well.

In the motor theory of perception, prior to making an articulatory mapping, we detect (or estimate) the acoustic features that serve as perceptual cues. There is evidence that feature detection is carried out by the auditory system in the ear and auditory centers of the brain [5],[29]. Experiments have been carried out to study the possibility of special auditory detectors for the perception of formant energy, bursts, voice onset times, formant transitions, and the presence (or lack) of voicing. Different features were found to correspond to different neural cell firing patterns in the early and higher level auditory centers. There are two primary neural cell firing characteristics used in feature detection: increase in the firing rate with a sudden change in energy (within specific frequency bands) and the synchronizing of a firing pattern to frequency. With plosive consonants, for example, some cells in the higher auditory levels, called onset cells, appear to fire at both the burst release and the vowel onset and thus may directly encode voice onset time in their response pattern. There is also evidence that in the detection of plosive consonants, detectors may exist that are tuned to sense the burst and the formant transition into the following vowel. With vowels, formant energy can be identified in two ways. First, there is an increase in the firing rate of nerve fibers tuned to specific frequencies and, second, nerve fibers synchronize with the formant frequency (firing at the peaks in each cycle of these frequencies), a characteristic known as phase synchrony. Similar neural firing mechanisms are hypothesized to exist for identifying the pitch associated with harmonic spectra.

In the design of speech analysis and synthesis systems, we will need to consider how changes in the temporal and spectral properties of a speech waveform affect its perception. For example, changes in the short-time Fourier transform phase characteristics of speech may affect the phase synchrony of frequency components in firing patterns of auditory nerves. We will address such issues in Chapter 8 and other places throughout the text.

3.7 Summary

This chapter described qualitatively the main functions of the speech production mechanism and the associated anatomy. Articulatory and acoustic descriptors of speech sounds were given and, based on these features, the study of phonemics and phonetics was introduced. Some implications of sound production and perception mechanisms for signal processing algorithms were discussed. For example, the source mechanism can result in a variety of voice types, such as the hoarse, breathy, or diplophonic voice, that can influence speech analysis and synthesis strategies. Glottal secondary pulses that occur in the diplophonic voice, for instance, have in part motivated multi-pulse speech analysis/synthesis, and aspiration that occurs in the breathy voice has in part motivated residual-based speech analysis/synthesis, both of which we will study in Chapter 12.
It was emphasized that fine structure and transitory characteristics of both the source and system, such as formant transitions, voice onset time, and degree and timing of consonantal aspiration, are aspects of a sound that are important for its perceptual identity, and thus must be considered in the development of signal processing algorithms. These features were shown to form the smallest elements of a language, being combined to form its phonemes, and hypotheses were presented that the auditory system is predisposed to these kinds of particular acoustic and articulatory features. The property of coarticulation, which is the influence of the articulation of one sound on the articulation of other sounds in the same utterance, and the importance of prosodics were also briefly discussed. In the next chapter, we develop a more quantitative description of the acoustics of speech production, showing how the heuristics of this chapter are approximately supported with mathematical models. We also will predict other acoustic effects not seen by a qualitative treatment.

**EXERCISES**

3.1 Consider the following two-pole model for the glottal pulse:

\[
G(z) = \frac{1}{(1 - \alpha z^{-1})(1 - \beta z^{-1})}
\]

with \( \alpha \) and \( \beta \) both real, positive, and less than one, and where the region of convergence includes the unit circle. In the time domain, \( g[n] \) can be expressed as the convolution of two decaying exponentials:

\[
g[n] = (\alpha^n u[n]) \ast (\beta^n u[n]).
\]

(a) Sketch \( g[n] \) and \( |G(\omega)| \). Assume \( \alpha \) and \( \beta \) are close to unity, say, about 0.95. Why are Equations (3.7) and (3.8) a reasonable model for the spectral magnitude, but not for the shape of the glottal pulse?

(b) Explain why an improved model for the glottal pulse is given by

\[
\tilde{g}[n] = g[-n].
\]

Derive the \( z \)-transform of \( \tilde{g}[n] \). Where are the poles of \( \tilde{g}[n] \) in relation to those of \( g[n] \)?

(c) Consider the periodic glottal flow waveform \( u[n] = \tilde{g}[n] \ast \sum_{k=-\infty}^{\infty} \delta[n - kP] \) where \( P \) is the pitch period. Assuming the glottal pulse is two convolved decaying exponentials as in part (b), sketch the Fourier transform magnitude of the windowed glottal flow waveform \( u[n] \) for a rectangular window with length equal to \( N \) and also \( 2N \). Which window length would be used in the calculation of a narrowband spectrogram of the glottal flow waveform?

3.2 Suppose we ask a speaker to hold a vowel steady. Ideally, the speaker would give a perfectly periodic waveform with constant pitch and loudness. In practice, however, these characteristics change in time for a “steady-state” vowel. Such small alternating deviations can result in hoarseness in a voice. In this exercise, you explore the consequences of small changes in pitch and loudness on the speech spectrum.
(a) Suppose that the pitch period of the speaker is steady except for a small deviation of $\epsilon$ that alternates in sign every pitch period (Figure 3.31a). Show that the glottal pulse train (assume no shaping) can be expressed as

$$p[n] = \sum_{k=-\infty}^{\infty} \delta[n - (2k)P] + \sum_{k=-\infty}^{\infty} \delta[n + \epsilon - (2k + 1)P].$$

Then derive the following expression:

$$|P(\omega)|^2 = 2(1 + \cos((\epsilon - P)\omega)) \left[ \sum_{k=-\infty}^{\infty} \frac{2\pi}{2P} \delta(\omega - \frac{2\pi}{2P}) \right]^2$$

where $P(\omega)$ is the Fourier transform of $p[n]$. Plot $|P(\omega)|^2$ for $\epsilon = 0$, for $0 < \epsilon \ll P$, and for $\epsilon = P$. What is the effect of pitch jitter on the short-time speech spectrum? Sketch the effect and assume a short-time rectangular window of length $N_w = 3P$. Argue why the presence of pitch jitter may lead to the appearance of “false resonances” in the speech spectrum. In determining the effect of pitch jitter on the short-time spectrum, use MATLAB, if helpful.

(b) Suppose now that the amplitude of the ideal glottal pulses is constant except for a small alternating deviation of $\Delta$, as illustrated in Figure 3.31b. Derive the following expression for $|P(\omega)|^2$:

$$|P(\omega)|^2 = 2(1 + \Delta^2) \left[ 1 + \frac{(1 - \Delta^2)}{(1 + \Delta^2)} \cos(P\omega) \right] \left[ \sum_{k=-\infty}^{\infty} \frac{2\pi}{2P} \delta(\omega - \frac{2\pi}{2P}) \right]^2.$$
Plot \( |P(\omega)|^2 \) for \( \Delta = 0, 0 < \Delta \ll 1, \) and \( \Delta = 1. \) Under the same short-time assumptions as in part (a), argue why the presence of amplitude jitter may lead to the appearance of “false resonances” in the speech spectrum.

3.3 Consider a vocal fold oscillation in a vocal fry or diplophonic state, where a secondary glottal flow pulse occurs within a glottal cycle. We model this condition over one pitch period as

\[
\tilde{g}[n] = g[n] + \alpha g[n - n_o]
\]

where \( n_o \) is the time delay between the primary and secondary pulses. The resulting periodic glottal flow waveform is given by \( u[n] = \sum_{k=-\infty}^{\infty} \tilde{g}[n - kP], \) where \( P \) is the pitch period.

(a) Determine the Fourier transform of \( \tilde{g}[n] \) in terms of the Fourier transform of \( g[n] \). With this result, write the Fourier transform of the periodic glottal flow waveform \( u[n] \), i.e., \( U(\omega) \).

(b) Suppose, in a diplophonic state, that \( n_o = P/2 \), where \( P \) is the glottal pitch period. Describe how the presence of \( g[n - n_o] \) affects, at the harmonic frequencies, the squared magnitude of the Fourier transform \( U(\omega) \) derived in part (a), i.e., of the Fourier transform of the glottal flow waveform \( u[n] \). Describe the effect as \( \alpha \) changes from 0 to 1.

(c) Repeat part (b) with \( n_o = 3P/4 \) corresponding to a vocal fry state where the secondary glottal pulse is close to the primary pulse.

3.4 Argue qualitatively why the first formant frequency of the singing voice is raised when a singer opens his/her jaw wide.

3.5 Referring to the phoneme symbols in Figure 3.17, do a complete phonemic transcription of the following sentences:

(a) May we all learn a yellow lion roar.

(b) She sells sea shells by the sea shore.

(c) The next man ate an azure estimate.

(d) Thieves who rob friends deserve jail.

Whenever you feel that more than one phoneme might occur, indicate which ones.

3.6 There are a variety of ways of classifying speech sounds into distinctive sounds (i.e., phonemes). These methods fall under the study of articulatory phonetics and acoustic phonetics.

(a) List the procedures used for distinguishing speech sounds in both areas of study.

(b) Describe the articulatory and acoustic differences and similarities between the voiced fricative /z/, as in “azure,” and the unvoiced fricative /s/, as in “she.”

3.7 In this exercise you work with the speech waveform given in Figure 3.13.

(a) Perform a phonetic transcription of the sentence illustrated in Figure 3.13. Place phonetic symbols next to each phoneme.

(b) Assume that the utterance is plotted for 2 seconds and the sampling rate is 10000 samples per second. Also assume that spoken English can be represented in terms of 40 phonemes, and sentences are formed by simple concatenation of these phonemes. Estimate the information rate (bit rate) of spoken English. Hint: Compute the number of bits represented by 40 phonemes.
3.8 Consider the spectrograms defined in Section 3.3.
(a) Derive the narrowband spectrogram approximation in Equation (3.5). Assume that the width of the window Fourier transform mainlobe is less than $\frac{2\pi}{P}$, where $P$ is the pitch period.
(b) Derive the wideband spectrogram approximation in Equation (3.6). Assume the vocal tract impulse response decays to roughly zero by the end of a glottal cycle. Hint: Consider the vocal tract impulse response as a set of decaying sinewaves. The short sliding window sees these sinewaves as having decreasing amplitudes.

3.9 Consider the variability of vowel characteristics among speakers. If humans can create a phoneme from similar vocal tract shapes, but with different speaker identifiability, then the source and prosody must play a large role in speaker characteristics, as well as the vocal tract. A ventriloquist, for example, can keep his/her vocal tract in roughly the same shape (i.e., for a particular phoneme), yet mimic different voices. Offer an explanation in terms of source and prosodic characteristics.

3.10 Determine the Fourier transform of the response $x[n]$ for the voiced fricative model in Example 3.4. Assuming you are given the vocal tract and front cavity responses $h[n]$ and $h_f[n]$, respectively, propose a method for determining the source functions $u[n]$ and $q[n]$.

3.11 Based on the spectrogram, propose a method for measuring a speaker’s rate of articulation. Hint: Consider the difference in spectral magnitudes in time, integrate this difference across frequency, and form a speaking rate metric.

3.12 Propose a simplified mathematical model of an unvoiced plosive, accounting for the burst after the opening of the oral tract closure and aspiration noise prior to the onset of voicing. Model the burst as an impulse $\delta[n]$ and aspiration as a noise sequence $q[n]$. Assume a linear time-varying oral tract.

3.13 Consider the two utterances of Example 3.6 shown in Figure 3.29. For each case, normal and stressed, make a rough estimate of the pitch contour during the vowel /e/ in the word “today.” Also, make a sketch of possible glottal airflow velocity, and the corresponding source spectrum, in each case.

3.14 Propose weaknesses in the motor theory of speech production and perception.

3.15 Using the simple pitch generation rules given at the beginning of Section 3.5, sketch roughly a pitch contour for the declarative statement, “I think, therefore I am,” and for the question, “I think, therefore I am?” It will help to listen to yourself recite the two phrases and feel the change in tension in your vocal folds.

3.16 Consider a continuous-time window $w(t)$ that is a rectangle over $|t| \leq T$ and zero elsewhere.
(a) Define the window’s duration as $D = 2T$ and the bandwidth $B$ as the distance between its two first zero crossings in frequency to the right and left of the frequency origin. State the relation constraining $D$ and $B$. It is important to recognize in this case that the product $DB$ has a lower limit that is not equal to that given in the uncertainty principle of Chapter 2 (see Chapter 11 for the continuous-time version). This is because the definition of signal duration and bandwidth in this problem is not the same as that in Chapter 2. The specific lower-limit constant will change with these definitions. The important property is the inverse relation between duration and bandwidth.
(b) Consider a very simple speech waveform consisting of only one sinewave and bandpass noise whose spectrum is shown in Figure 3.32a. Multiply the speech by the window of part (a) and find the shortest window length $D$ so that the Fourier transform of the windowed sinewave component does not interact with the Fourier transform of the windowed noise. Approximate the Fourier transform of the window by only its main lobe, i.e., up to the first positive- and negative-frequency zero crossing.
(c) Consider a more complex speech waveform consisting of five sinewaves harmonically related as in Figure 3.32b and the same noise spectrum as above; one can think of this as a simple model of a voiced fricative. The fundamental frequency or “pitch” is denoted by $\Omega_o$. The Fourier transform of the rectangular window is again approximated as one main lobe, but now the width of this main lobe is given by $2\Omega_o$. Multiply the speech by the window and find the largest pitch possible so that the Fourier transform of the windowed voiced speech component does not interact with the Fourier transform of its windowed unvoiced (noise) component.

3.17 (MATLAB) On the companion website, you will find in the directory Chap_exercises/chapter3 a workspace ex3M1.mat. Load this workspace and plot the speech waveform labeled speech1_10k. This speech segment was taken from a vowel sound that is approximately periodic, 25 ms in duration, and sampled at 10000 samples/s.

(a) Plot the log-magnitude of the Fourier transform of the signal over the interval $[0, \pi]$, using a 1024-point FFT. The signal should be windowed with a Hamming window of two different durations: 25 ms and 10 ms with the window placed at the signal’s center in each case. Show the log-magnitude plot for each duration.

(b) Which spectral estimate is more suitable to pitch estimation? Explain.

3.18 (MATLAB) In this MATLAB exercise, you design a number of glottal pulse trains with different shapes and pitch, and analyze their spectral and perceptual properties.

(a) Create in MATLAB the following glottal pulse, which is a time-reversed decaying exponential convolved with itself:

$$g[n] = (\alpha^{-n}u[-n]) \ast (\alpha^{-n}u[-n]).$$

Set the length of $g[n]$ to where it has effectively decayed to zero. Experiment with the different values of $\alpha = 0.9, 0.95, 0.99, 0.998$ and compute the resulting Fourier transform magnitude, using a FFT length sufficiently long to avoid significant truncation of the pulse.

(b) Convolve $g[n]$ from part (a) (with $\alpha = 0.99$) with a periodic impulse train of pitch 100 Hz and 200 Hz. Assume an underlying sampling rate of 10000 samples/s. Using the MATLAB sound.m function, listen to the two pulse trains and make a perceptual comparison. Window a 20-ms piece (with a Hamming window) of the waveform and compare the spectral envelope and harmonic structure for the two pitch selections.

(c) We can consider the glottal pulse in part (a) as arising from vocal folds with an abrupt closure and thus corresponding to a “hard” voice. Modify the pulse shape ($\alpha = 0.99$) so that the
sharp edge at closure is rounded off, corresponding to vocal folds that close more gradually. Compute the resulting Fourier transform magnitude. Repeat part (b) with your new glottal pulse. How does your result compare perceptually with the pulse train of part (b)?

3.19 (MATLAB) On the companion website, you will find in the directory Chap_exercises/chapter3 a workspace `ex3M2.mat`. In this workspace, you will find waveforms of the words “we” and “be” (at 10000 samples/s) in the array `we_be_10k`. The waveforms and a spectrogram of the waveforms are shown in Figure 3.33. To display the spectrogram and waveform, use the function `specgram_ex3p19.m` also in Chap_exercises/chapter3. To listen to the waveform, use the MATLAB function `sound.m`.

(a) The spectrogram in Figure 3.33b was obtained with a window duration of 20 ms (200 samples for a 10000-samples/s sampling rate) and a frame interval of 1 ms (10 samples for a 10000-samples/s sampling rate). Display with MATLAB the spectrogram in Figure 3.33b using the routine `specgram_ex3p19.m`. Is this a narrowband or wideband spectrogram? Now display the spectrogram using a 5-ms window and a 1-ms frame interval. Is the result a narrowband or wideband spectrogram? For this 5-ms window, describe how the time and frequency resolution of speech events differs from that of the spectrogram in Figure 3.33b. Then display a variety of spectrograms with window durations longer than 20 ms and shorter than 5 ms, together with frame intervals both equal to and greater than 1 ms, and describe the resulting time-frequency tradeoffs.

(b) Splice out the voiced plosive “b” in the word “be” and synthesize a word with an increase in the voice onset time of the “b.” What is the perceptual effect? Is it possible to transform the voiced plosive “b” to the unvoiced plosive “p” with this change in the voiced onset time? Consider attributes such as vocal cord vibration and aspiration. (In splicing, display the waveform and spectrogram of the word “be” and use the MATLAB function `ginput.m`.) Optional: If your transformed sound does not sound like a “p” then, again considering vocal cord vibration and aspiration, make a more accurate transformation.

![Figure 3.33](image_url) (a) Waveform and (b) narrowband spectrogram of sequence of words “we” and “be.”
(c) Now interchange the “b” and “w” in the two words “be” and “we.” In splicing, be careful that you capture all important perceptual cues of the two phonemes. What is the effect? Can you interchange the two words by interchanging the two initial phones?

BIBLIOGRAPHY


