7
Speech (Sound) Processing

Acoustic

Human communication is achieved when thought is transformed through language into speech. The sounds of speech are initiated by activity in the central nervous system, and this stimulates the respiratory and vocal tracts. Speech sounds are usually produced through the expulsion of air through the larynx. Voiced sounds are due to the vibration of the vocal cords, allowing puffs of air to be transmitted to the vocal tract. With unvoiced speech the larynx is held open, and a turbulent flow of air at a point of constriction creates the basic sound. The vocal tract acts as a multiresonant filter for the transmitted sound. Resonances that result are referred to as formants. Communication is completed when the speech sounds are received by the auditory system of another person and translated back into thought. Speech sounds are important, but not essential for communication. Visual signals such as sign language of the deaf or signed English can also be used, as well as the written word.

Speech is a complex acoustic signal, and information is transmitted to the brain at a rapid rate. For example, three to seven syllables are uttered per second during a conversation (Pickett 1980). The phonemes are discrete perceptual units consisting of complex sound elements and are coded into patterns of neural discharges in the lower auditory brain centers for decoding into speech by the higher auditory centers. Speech understanding by cochlear implant patients requires rapid processing of the speech signals into patterns of electrical stimulation in the auditory nerve in the cochlea that partially reproduce those for normal hearing.

A knowledge of speech science is important for developing cochlear implant speech processing strategies. There are many excellent references, such as Fant (1973), Flanagan and Rabiner (1973), Ainsworth (1976, 1992), O'Shaughnessy (1987), Ainsworth (1992), and Ladefoged (1993). This chapter discusses issues that are especially relevant to cochlear implants.

Articulators and Vocal Tract Shape

The articulators that shape the cavities to create resonances (formants) in the speech signal are part of the oropharynx. They can be classified for consonants into those for place and manner of articulation. The articulators for place of articulation in the following words can be classified as (1) bilabial, with the two lips together for "bee"; (2) labiodental, with the lower lip behind the upper front teeth for "fee"; (3) dental, with the tip of the tongue or blade next to the upper front teeth for "thee"; (4) alveolar, with the tongue tip or blade close to or against the upper alveolar ridge for "dee"; (5) retroflex, with the tip of the tongue close to the back of the alveolar ridge for "re"; (6) palato-alveolar, with the tongue blade at the back of the alveolar ridge for "she"; and (7) velar, with the back of the tongue against the soft palate for "key."

The manner of articulation may be (1) voiced or unvoiced, providing discrimination between "do" and "to"; (2) nasal, where the soft palate is down and the air passes out through the nose for the sounds "me," "knee," and "ing"; (3) oral stop (plosive), where the air stream is completely obstructed at a point in the vocal tract as in "bet," "debt," and "get"; (4) fricative, due to the close approximation of two articulators, so that the air stream is partially obstructed but turbulent (the higher pitch sounds with more acoustic energy as in "see" are sibilants, and others are fricative nonsibilants); (5) lateral sounds, where obstruction of the air stream occurs at a point along the center of the oral tract with incomplete closure between one or both sides of the tongue and the roof of the mouth as in "lee"; and (6) affricate, where the sound is a combination of plosive and fricative as in "cheer."

The complex speech sounds produced by the articulators referred to above have specific acoustic features, and can be classified accordingly. Their features are processed by the central auditory pathways, and are discussed below.

Speech Analysis

Speech is a complex signal, and its analysis provides an understanding of the cues of importance in perception. Representing these cues with electrical stimulation helps provide optimal speech processing for profoundly deaf people with cochlear implants.

Band-Pass Filtering

Speech consists of the fundamental or voicing frequency and the harmonic frequencies produced by the resonant cavities. An effective method of analyzing speech is to pass the signal through a bank of band-pass filters. The pattern of energy in each filter or the relative strength of the frequency components is the speech spectrum. The speech spectrum over a period of time can be represented graphically as a spectrogram, as illustrated in Figure 7.1.

Each filter represents the energy of speech across a certain frequency band.

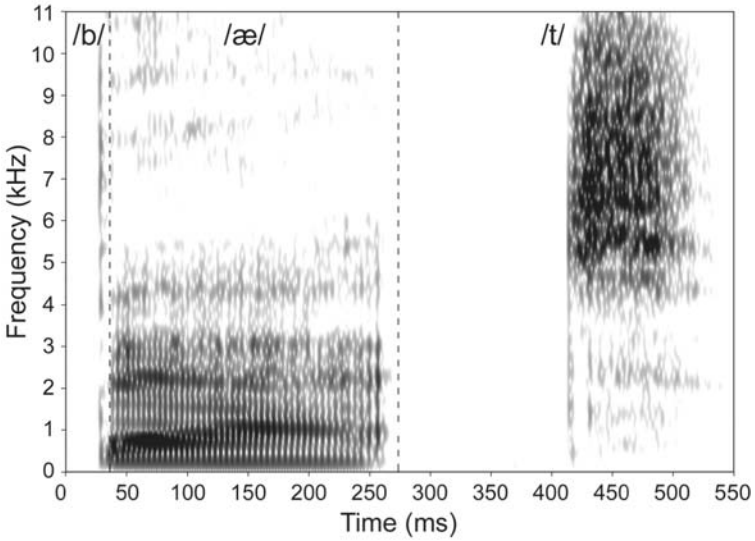

FIGURE 7.1. A speech spectrograph of the word "bat" showing frequency versus time with the intensity of the sound increasing from light to dark. The locations of the three phonemes /b/, /æ/, and /t/, are shown by the dashed lines.

The center frequency and the slope of the filter on the high- and low-frequency side can be varied, and thus change the extent of the filtering over the frequency range. The ideal filter should transmit only the frequencies within the pass band. Their amplification should be the same, and there should be no change in the phase relations of the frequencies within the pass band. It is usual to consider the upper and lower ends of the band pass separately. A low-pass (high-cut) filter determines the upper limit of the band pass. A high-pass (low-cut) filter determines the lower limit of the pass band. The simplest and most fundamental type of filter is the RC circuit that has a resistor (R) and capacitor (C) in parallel. The gain and phase shift of this filter falls short of the ideal. For this reason digital filters or active filters incorporating operational amplifiers are used. Filters are discussed further in Chapter 8, and reference can be made to an appropriate text such as Dewhurst (1976), Brophy (1977), and O'Shaughnessy (1987).

Spectrogram

The sound spectrogram, referred to above, is produced by a spectrograph that was first developed at the Bell Telephone Laboratories to analyze and synthesize speech. It plots the amplitude of the signal from variable filters over time. The electromagnetic data were originally displayed as a paper readout, and now digitally on a computer. It is an approximation to the continuous spectrum at a number of instances in time. An example is shown in Figure 7.1 for the word "bat." The intensity of the signal at each frequency is indicated by the depth of the shading. It is possible to see the concentrations of energy that indicate formant

frequencies. The voicing frequency is represented by the periodic striations resulting from the repetitive puffs of air produced by the larynx. The spectrogram is helpful in understanding how speech can be analyzed by cochlear implant speech processors.

Formants

When speech is filtered into its spectral components, there are peaks of energy at certain frequencies called formants. They are, as referred to above, resonances produced by variations in the dimensions of the pharynx during articulation. Speech can have many well-defined formants, but the first (F1) and second (F2) formants are the most important for the identification of vowels. The F1 frequency range for American English and Swedish is from approximately 250 to 700 Hz, and the F2 range is from 700 to 2300 Hz (Peterson and Barney 1952; Fant 1959).

Fourier Analysis—Fast Fourier Transform

The most obvious way to view sound is in the time domain (Fig. 7.2). The amplitude of the waves can be inspected over instants in time. However, with a complex wave it is difficult to determine the constituent frequencies. This can be done using a Fourier analysis. First, the oscillations of sound can be expressed mathematically as both sine and cosine terms. Second, the constituent frequencies are those that are determined by the analysis to correlate with the sine and cosine waves. The waves can be defined by either sine or cosine functions, and the difference between the two provide the phase information.

An example of the Fourier transforms of two composite waves from two sine waves of the same frequency and amplitude, but with a 45-degree phase shift, are shown in Figure 7.3. From this it can be seen that the composite waves are different even though the components have the same frequencies and amplitudes. This difference is reflected in the pattern of interspike intervals in the neural

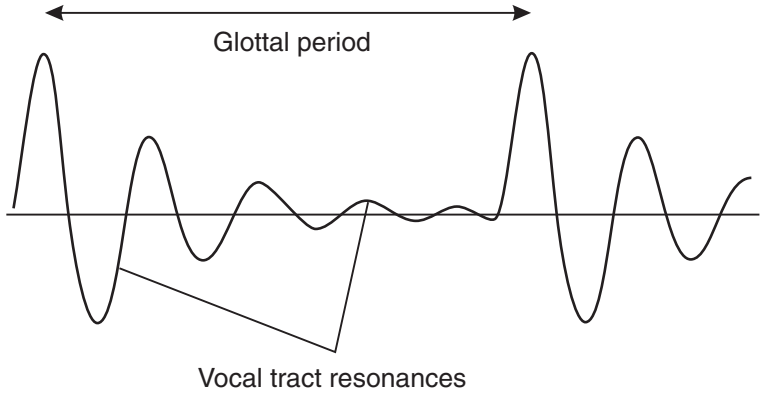

FIGURE 7.2. A diagram of the speech waveform for a vowel showing the glottal period and the vocal tract resonance.

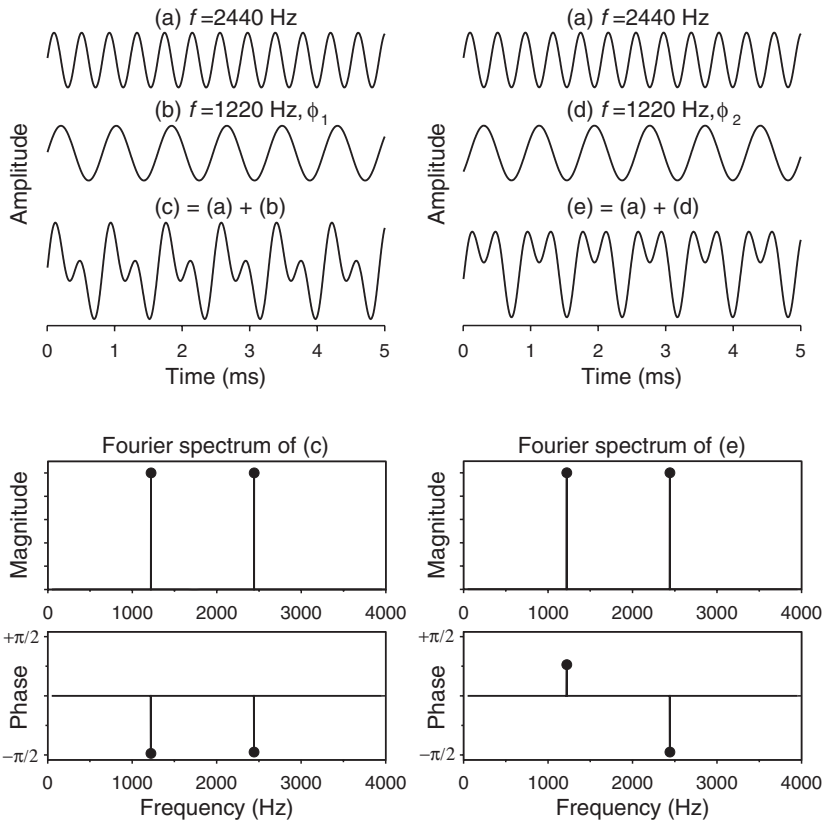

FIGURE 7.3. Top: The superposition of sine waves of frequency (a) 2440 Hz, and (b and d) 1220 Hz, resulting in waveforms (c) and (e). The phase difference between waves (b) and (d) is $\phi_1 - \phi_2 = 45$ degrees. Bottom: The Fourier spectra of waves (c) and (e) showing the magnitude and phase of each frequency component.

activity, as shown in the study by Rose et al (1969). If a Fourier analysis of the speech signal is carried out, the frequency spectrum is well preserved but the phase information is lost. The fine time information is probably important for the encoding of frequency and should be preserved in new speech-processing strategies.

There are other methods of analyzing speech such as models of cochlear function and linear predictive coding, but they will not be discussed as they are not yet used routinely in speech-processing strategies for cochlear implants.

Speech Perception and Production

Speech conveys meaning (semantics) through words connected as sentences based on grammatical rules (syntax). The words and grammatical rules are the same for spoken and written speech, and constitute language. The perception of

Springer

http://www.springer.com/978-0-387-95583-4

Cochlear Implants
Fundamentals and Applications
Clark, G.
2003, XXXVIII, 831 p., Hardcover
ISBN: 978-0-387-95583-4