14 Speech and Its Perception

Pure tones, clicks, and the like enable us to study specific aspects of audition in a precise and controllable manner. On the other hand, we communicate with each other by speech, which is composed of particularly complex and variable waveforms. A knowledge of how we perceive simpler sounds is the foundation upon which an understanding of speech perception must be built. As one might suppose, speech perception and intimately related areas constitute a voluminous subject encompassing far more than hearing science, per se, and the interested student will find numerous sources addressing the topic at various levels (e.g., Miller, 1951; Fletcher, 1953; Fant, 1970; Flanagan, 1972; Massaro, 1987, 1998; Pickett, 1999; Miller et al., 1991; Kent and Read, 2002; Liberman, 1996; Ryalls, 1996; Jusczyk and Luce, 2002; Diehl et al., 2004; Galantucci et al., 2006; Ferrand, 2007; Raphael et al., 2007; Massaro and Chen, 2008).

Speech perception and speech production are inherently interrelated. We must be able to speak what we can perceive, and we must have the ability to perceive the sounds that our speech mechanisms produce. Traditionally, the sounds of speech have been described in terms of the vocal and articulatory manipulations that produce them. We too shall begin with production. For the most part, our discussion will focus upon phonemes.

By a **phoneme** we mean a group of sounds that are classified as being the same by native speakers of a given language. Let us see what the "sameness" refers to. Consider the phoneme /pi/ as it appears at the beginning and end of the word "pipe." There are actually several differences between the two productions of /p/ in this word. For example, the initial /p/ is accompanied by a release of a puff of air (aspiration), whereas the final /p/ is not. In other words, the actual sounds are different, or distinct **phonetic elements**. (By convention, phonemes are enclosed between slashes and phonetic elements between brackets.) In spite of this, native speakers of English will classify both as belonging to the family designated as the /p/ phoneme. Such phonetically dissimilar members of the same phonemic class are called allophones of that phoneme. Consider a second example. The words "beet" and "bit" (/bit/ and /bIt/, respectively) sound different to speakers of English but the same to speakers of French. This happens because the phonetic elements [i] and [I] are different phonemes in English, but are allophones of the same phoneme in French. Since the French person classifies [i] and [I] as members of the same phonemic family, he hears them as being the same, just as English speakers hear the aspirated and unaspirated productions of /p/ to be the same.

This last example also demonstrates the second important characteristic of phonemes. Changing a phoneme changes the meaning of a word. Thus, /i/ and /I/ are different phonemes in English, because replacing one for the other changes the meaning of at least some words. However, [i] and [I] are not different phonemes in French; that is, they are **allophones**, because replacing one for the other does not change the meaning of words. Implicit in the distinction of phonetic and phonemic elements is that even elementary speech sound classes are to some extent learned. All babies the world over produce the same wide range of sounds phonetically; it is through a process of learning that these phonetic elements become classified and grouped into families of phonemes that are used in the language of the community.

SPEECH SOUNDS: PRODUCTION AND PERCEPTION

Our discussion of speech sounds will be facilitated by reference to the simplified schematic diagram of the vocal tract in Fig. 14.1 The power source is the air in the lungs, which is directed up and out under the control of the respiratory musculature. Voiced sounds are produced when the vocal folds (vocal cords) are vibrated. The result of this vibration is a periodic complex waveform made up of a fundamental frequency on the order of 100 Hz in males and 200 Hz in females, with as many as 40 harmonics of the fundamental represented in the waveform (Flanagan, 1958) (Fig. 14.2a). Voiceless (unvoiced) sounds are produced by opening the airway between the vocal folds so that they do not vibrate. Voiceless sounds are aperiodic and noise-like, being produced by turbulences due to partial or complete obstruction of the vocal tract. Regardless of the source, the sound is then modified by the resonance characteristics of the vocal tract. In other words, the vocal tract constitutes a group of filters that are added together, and whose effect is to shape the spectrum of the waveform from the larynx. The resonance characteristics of the vocal tract (Fig. 14.2b) are thus reflected in the speech spectrum (Fig. 14.2c). The vocal tract resonances are called **formants** and are generally labeled starting from the lowest as the first formant (F1), second formant (F2), third formant (F3), etc. This is the essence of the **source-filter theory**, or the **acoustic theory of speech production** (Fant, 1970).

Vowels

Speech sounds are generally classified broadly as vowels and consonants. Vowels are voiced sounds whose spectral characteristics are determined by the size and shape of the vocal tract. (Certain exceptions are notable. For example, whispered speech is all voiceless, and vowels may also be voiceless in some contexts of voiceless consonants in connected discourse. Also, the nasal cavity is generally excluded by action of the velum unless the vowel is in the environment of a nasal sound.) Changing the shape of the vocal tract changes its filtering characteristics, which in turn change the formant structure, that is, the frequencies at which the speech signal is enhanced or de-emphasized (Fig. 14.2). Diphthongs such as /aI/ in "buy" and /oU/ in "toe" are heard when one vowel glides into another.

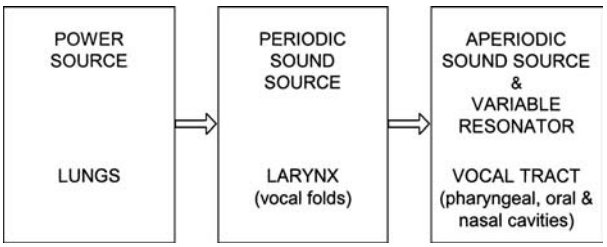

Figure 14.1 Schematic representation of speech production.

In general, the formant frequencies depend upon where and to what extent the vocal tract is constricted (Peterson and Barney, 1952; Stevens and House, 1955, 1961; Flanagan, 1972). The locations and degrees of these constrictions control the sizes and locations of the volumes in the vocal tract. For example, elevation of the back of the tongue results in a larger volume between this point of constriction and the lips than does elevation of the tongue tip. We may thus describe a vowel from front to back in terms of the amount of tongue elevation. Lip rounding is another important factor. In English, front vowels (/i, I, e, ɛ, æ/) are produced with retraction of the lips, while the lips are rounded when the back vowels (/u, U, o, ɔ, a/) are formed. Rounding the front vowel /i/ as in "tea" while keeping the high-front tongue placement results in the French vowel /y/, as in "*tu*." The degree of tenseness associated with the muscle contractions is also a factor in vowel production and perception, as in the differentiation of the tense /i/ ("peat") from the lax /I/ ("pit"). Tense vowels are generally more intense and longer in duration than their lax counterparts.

The middle vowels (/ʌ,ə,ɜ,ɝ,ɚ/) are produced when tongue elevation is in the vicinity of the hard palate. These include the *neutral vowel* or *schwa*, /ə/, associated mainly with unstressed syllables (e.g., "*a*bout" and "*su*pport").

Without going into great detail, the frequency of the **first format (F1)** is largely dependent upon the size of the volume behind the tongue elevation, that is, upon the larger of the vocal tract volumes. This volume must, of course, increase as the elevated part of the tongue moves forward. Thus, front tongue elevation produces a larger volume behind the point of constriction, which in turn is associated with lower F1 frequencies,

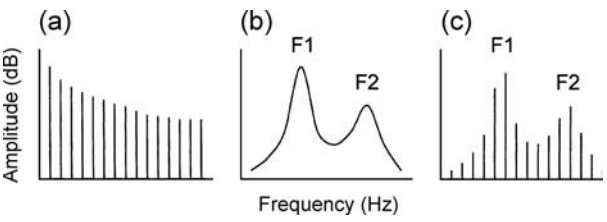

Figure 14.2 The source-filter theory (acoustic theory) of speech production: Idealized spectra showing that when the glottal source spectrum (a) is passed through the vocal tract filters (b) the resulting (output) spectrum (c) represents characteristics of the vocal tract. F1 and F2 indicate the first two formants.

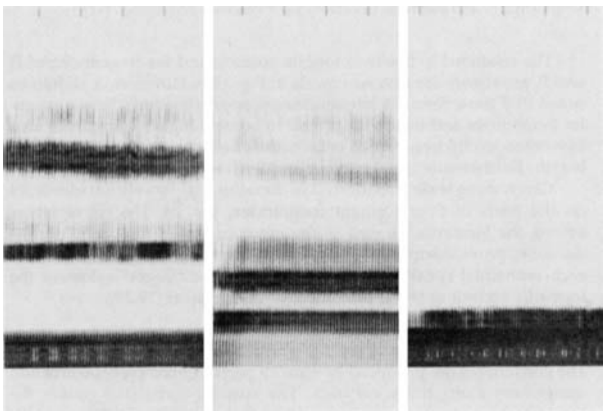

Figure 14.3 Spectrograms showing sustained production of the vowels /i/, /æ/, and /u/ (left to right). Timing marks along the top are 100 ms apart.

whereas back tongue elevations decrease the size of this volume, thereby raising the frequency of F1. The frequency of the **second formant (F2)** depends largely upon the size of the volume in front of the point of tongue elevation, becoming higher when the cavity is made smaller (when the tongue is elevated closer to the front of the mouth). Lip rounding lowers the first two formants by reducing the size of the mouth opening.

Figure 14.3 shows sustained productions of several vowels in the form of sound **spectrograms** (Koenig et al., 1946; Potter et al., 1947). Frequency is shown up the ordinate, time along the abscissa, and intensity as relative blackness or gray-scale. Thus, blacker areas represent frequencies with higher energy concentrations and lighter areas indicate frequency regions with less energy. The formants are indicated by frequency bands much darker than the rest of the spectrogram. These horizontal bands represent frequency regions containing concentrations of energy and thus reflect the resonance characteristics of the vocal tract. The vertical striations correspond to the period of the speaker's fundamental frequency.

The relationship between tongue position and the frequencies of F1 and F2 are shown for several vowels in Fig. 14.4. However, it should be noted that these formant frequencies are approximations based on average male values from just one study. Formant center frequencies and bandwidths tend to become higher going from men to women to children, which reflects the effect of decreasing vocal tract length (Fig. 14.5). Formant parameters vary appreciably among talkers and even between studies (e.g., Peterson and Barney, 1952; Hillenbrand et al., 1995); they are affected by such factors as neighboring phonemes, by whether the syllable is stressed or unstressed, etc.

The lower formants (especially F1 and F2, as well as F3) are primarily responsible for vowel recognition (Peterson, 1952; Peterson and Barney, 1952; Delattre et al., 1952; Hillenbrand et al., 1995). However, given the wide variations alluded to above, it is doubtful that vowels are identified on the basis of their formant frequencies, per se. The relationships among the

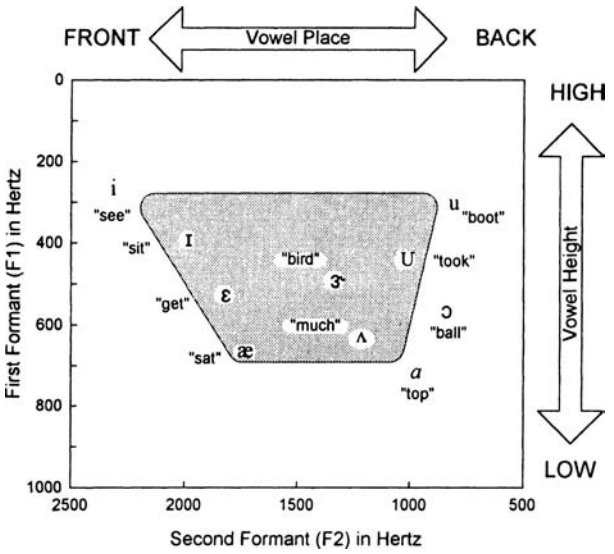

Figure 14.4 Vowel quadrilateral showing the approximate values of the first (F1) and second (F2) formants of several vowels as they relate to tongue height and place based on data for male talkers by Peterson and Barney (1952).

formants, as well as the environment of the vowel and its duration, provide important cues (Tiffany, 1953; Stevens and House, 1963; Lindblom and Studdert-Kennedy, 1967). It has been suggested that, for each individual speaker, the listener adjusts the "target" values of the formants according to the utterances of that speaker (e.g., Ladefoged and Broadbent, 1957; Lieberman, 1973); however, this is certainly not the only explanation for vowel perception. A lucid review of vowel perception issues and theories may be found in Kent and Read (2002), and more advanced students will find informative discussions in a series of papers by Miller (1989), Nearey (1989), and Strange (1989a, 1989b).

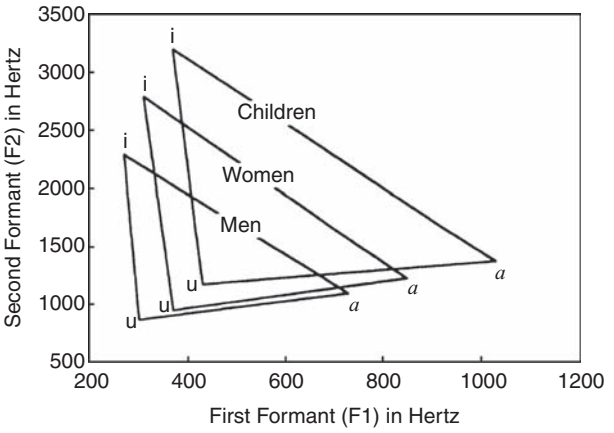

Figure 14.5 Average formant frequencies increase going from men to women to children, as illustrated here by the shifts in the average values of F1 and F2 for the vowels /i/, /a/, and /u/. *Source*: Based on data by Peterson and Barney (1952).

Consonants

The consonants are produced by either a partial or complete obstruction somewhere along the vocal tract. The ensuing turbulence causes the sound to be quasi-periodic or aperiodic and noise-like. The consonants are differentiated on the basis of manner of articulation, place of articulation, and voicing; that is, on how and where the obstruction of the vocal tract occurs and on whether there is vocal cord vibration. Table 14.1 shows the English consonants, arranged horizontally according to place of articulation and vertically by manner of articulation and voicing. Examples are given where the phonetic and orthographic symbols differ.

The stops, fricatives, and affricates may be either voiced or voiceless, whereas the nasals and semivowels are virtually always voiced. The nasal cavities are excluded from the production of all consonants except the nasals by elevation of the velum. We shall briefly discuss the consonants in order of the manner of their articulation.

The **stops** are produced by a transition (see below) from the preceding vowel, a silent period on the order of roughly 30 ms during which air pressure is impounded behind a complete obstruction somewhere along the vocal tract, a release (burst) of the built-up pressure, and finally a transition into the following vowel (Cooper et al., 1952; Fischer-Jorgensen, 1954; Liberman et al., 1956; Halle et al., 1957). Of course, whether there is a transition from the preceding vowel and/or into the following vowel depends upon the environment of the stop consonant. Voiceless stops in the initial position are generally aspirated, or released with a puff of air. Initial voiced stops and all final stops tend not to be aspirated, although this does not apply always or to all speakers. The voiceless stops (/p,t,k/) and their voiced cognates (/b,d,g/) are articulated in the same way except for the presence or absence of voicing and/or aspiration. As a rule, the voiceless stops tend to have longer and stronger pressure buildups than do their voiced counterparts (Sharf, 1962; Arkebauer et al., 1967).

The six stops are produced at three locations. The **bilabials** (/p,b/) are produced by an obstruction at the lips, the **alveolars** (/t,d/) by the tongue tip against the upper gum ridge, and the **velars** (/k,g/) by the tongue dorsum against the soft palate. Whether the sound is heard as voiced or voiceless is, of course, ultimately due to whether there is vocal cord vibration. However, cues differ according to the location of the stop in an utterance. The essential voicing cue for initial stops is **voice onset time (VOT)**, which is simply the time delay between the onset of the stop burst and commencement of vocal cord vibration (Lisker and Abramson, 1964, 1967). In general, voicing onset precedes or accompanies stop burst onset for voiced stops but lags behind the stop burst for voiceless stops. For final stops and those that occur medially within an utterance, the essential voicing cue appears to be the duration of the preceding

Table 14.1 Consonants of English.

	Bilabial	Labiodental	Linguadental	Alveolar	Palatal	Velar	Glottal
Stops							
Voiceless	p			t		k	
Voiced	b			d		g	
Fricatives							
Voiceless	ʍ(**wh**ich)	f	θ (**th**ing)	s	ʃ(**sh**oe)		
Voiced		v	ð (**th**is)	z	ʒ(bei**g**e)		
Affricates							
Voiceless					tʃ(ca**tch**)		h
Voiced					ʤ (do**dg**e)		
Nasals[a]	m			n		ŋ (si**ng**)	
Liquids[a]					r, l		
Glides[a]	w					j (**y**es)	

[a]The nasals, liquids, and glides are voiced.

vowel (Raphael, 1972). Longer vowel durations are associated with the perception that the following stop is voiced. Voiceless stops are also associated with longer closure durations (Lisker, 1957a, 1957b), faster formant transitions (Slis, 1970), greater burst intensities (Halle et al., 1957), and somewhat higher fundamental frequencies (Haggard et al., 1970) than voiced stops.

Place of articulation (bilabial vs. alveolar vs. velar) for the stops has been related to the **second formant (F2) transition** of the associated vowel (Liberman et al., 1954; Delattre et al., 1955), along with some contribution from the F3 transitions (Harris et al., 1958). By a formant transition we simply mean a change with time of the formant frequency in going from the steady-state frequency of the vowel into the consonant (or vice versa). Formant transitions may be seen for several initial voiced stops in Fig. 14.6. The F2 transitions point in the direction of approximately 700 Hz for bilabial stops, 1800 Hz for the alveolars, and 3000 Hz for the velars (Liberman et al., 1954). The second formation transition **locus principle** is illustrated in Fig. 14.7 These directions relate to the location of vocal tract obstruction. That is, a larger volume is enclosed behind an obstruction at the lips (/p,b/) than at the alveolus (/t,d/) or the velum (/k,g/) so that the resonant frequency associated with that volume is lower for more frontal obstructions. Moving the point of obstruction backward reduces the cavity volume and thus increases the resonant frequency. Additional place information is provided by the frequency spectrum of the stop burst. Stop bursts tend to have concentrations of energy at relatively low frequencies (500–1500 Hz) for the bilabials, at high frequencies (about 4000 Hz and higher) for the alveolars, and at intermediate frequencies (between around 1500 and 4000 Hz) for the velars (Liberman et al., 1956).

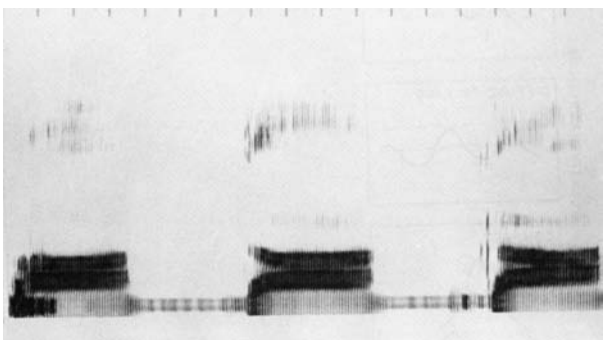

Figure 14.6 Spectrograms of /ba/, /da/, and / ga/ (left to right). Note second formant transitions.

There tends to be a considerable amount of *variability* (often described as a *lack of invariance*) in the formant cues because the configurations of the formant transitions change according to the associated vowels. This variability is readily observed in Fig. 14.7 and is especially apparent for the alveolars. For example, the second formant transition from /d/ into the following vowel is different for /di/ and /du/. An invariant place of articulation cue has been proposed on the basis of several acoustical and perceptual studies (e.g., Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980; Blumstein, Isaacs, and Mertus, 1982; Kewley-Port, 1983; Kewley-Port and Luce, 1984; Furui, 1986). For example, Stevens, Blumstein, and colleagues demonstrated invariant patterns in the gross configurations of the spectra integrated over a period of roughly 20 ms in the vicinity of the consonant release. Figure 14.8 shows the general configurations of these **onset spectra**, which are (a) *diffuse and falling* for bilabials, /p,b/; (b) *diffuse and rising* for the alveolars, /t,d/; and (c) *compact* for the velars, /k,g/.

In addition to the perceptual findings, further support for this concept comes from experiments using computer models (Searle, Jacobson, and Rayment, 1979) and auditory nerve discharge patterns (Miller and Sachs, 1983). There is theoretical and experimental support for the notion that children use onset spectra as the primary cues for place of articulation for stops before they learn to use formant transitions as secondary perceptual cues (e.g., Blumstein and Stevens, 1979, 1980; Ohde et al., 1995), although there are contrary models and results, as

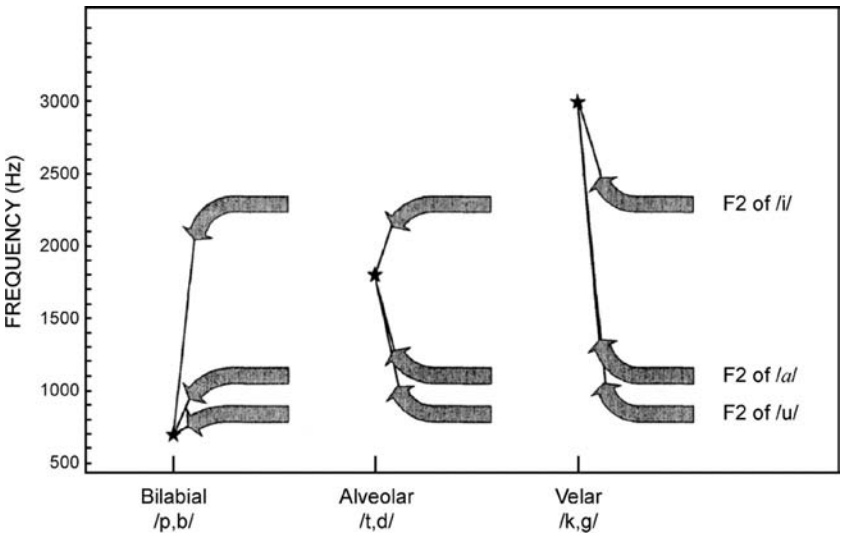

Figure 14.7 Artist's conceptualization of the second formant transition locus principle for stop consonant place of articulation. Stars indicate the locus (target frequency) toward which the second formant transitions point for bilabials, alveolars, and velars.

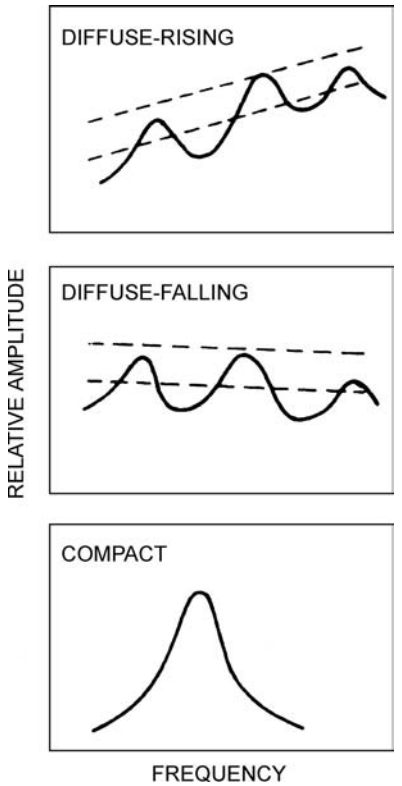

Figure 14.8 Invariant onset spectrum configurations associated with stop-consonant place of articulation: alveolars, diffuse, and rising; labials, diffuse, and falling; velars, compact. *Source*: Adapted from Blumstein and Stevens (1979), with permission of *J. Acoust. Soc. Am.*

well (e.g., Walley and Carrell, 1983; Nittrouer and Studdert-Kennedy, 1987; Nittrouer, 1992). Repp and Lin (1989) have suggested that the onset spectrum may provide the listener with a very brief "acoustic snapshot of the vocal tract" that might be supplemented by the dynamic cues of formant transitions if more information is needed to accurately identify the sound.

The **fricatives** are produced by a partial obstruction of the vocal tract so that the air coming through becomes turbulent. The nature of the fricatives has been well described (House and Fairbanks, 1953; Huges and Halle, 1956; Harris, 1958; Strevens, 1960; Heinz and Stevens, 1961; Jassem, 1965; Guerlekian, 1981; Jongman, 1985; Behrens and Blumstein, 1988). Several examples of the fricatives are shown in the spectrograms in Fig. 14.9 Fricatives are distinguished from other manners of articulation by the continuing nature of their turbulent energy (generally lasting 100 ms or more); vowels preceding fricatives tend to have greater power and duration, and somewhat longer

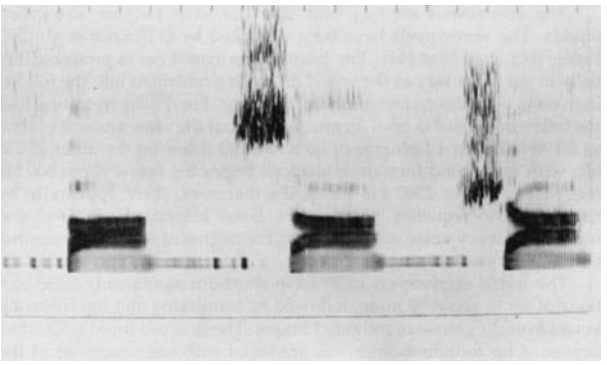

Figure 14.9 Spectrograms of /fa/, /sa/, and /ʃa/ (left to right).

fundamental frequencies, than vowels preceding stops. As is true for the stops, fricatives may be either voiced or voiceless. However, unlike the stops, the duration of the fricatives makes it possible for voicing to be cued by the presence versus absence of periodic energy during frication. In addition, fricative voicing is also cued by VOT and by the nature of the preceding vowel. Moreover, voiceless fricatives tend to have longer durations than their voiced counterparts.

Spectral differences largely account for place distinctions between the alveolar (/s,z/) and the palatal (/ʃ,ʒ/) sibilant fricatives. The palatals have energy concentrations extending to lower frequencies (about 1500–7000 Hz) than do the alveolars (roughly 4000–7000 Hz), most likely because of the larger volume in front of the point of vocal tract obstruction for /ʃ,ʒ/ than for /s,z/. On the other hand, /θ,ð/, and /f, v/ are differentiated largely on the basis of formant transitions. Because of resonation of the entire vocal tract above the glottis, /h/ possesses more low frequencies than the more anterior fricatives. The amplitudes of /s/ and /ʃ/ are considerably greater than those of /f/ and /θ/. However, perceptual experiments by Behrens and Blumstein (1988) have demonstrated that these amplitudes are less important than spectral properties in differentiating between these two groups of fricatives. The affricates are produced by the rapid release of their stop components into their fricative components.

The **nasals** are produced by opening the port to the nasal cavities at the velum. They are semivowels in that they are voiced and have some of the formant characteristics of vowels. However, they differ from other sounds by the coupling of the nasal and oral cavities. The characteristics of nasals have been described by Fujimura (1962) and others (Cooper et al., 1952; Malécot, 1956; House, 1957; Kurowski and Blumstein, 1987a, 1987b). The coupling of the nasal cavities to the volume of oral cavity behind the point of obstruction (the velum for /ŋ/, alveolus for /n/, and lips for /m/) constitutes a side-branch resonator. This results in antiresonances at frequencies that become lower as the volume of the side branch (oral cavity) becomes larger. Thus, we find that antiresonances appear in the frequency regions of roughly 1000 Hz for /m/ (where the side branch is the largest), 1700 Hz for /n/, and 3000 Hz for /ŋ/ (where the side branch is the shortest). Furthermore, overall intensity is reduced, and for the first formant is lower than for the vowels, constituting a characteristic low-frequency nasal murmur. Place of articulation is cued by differences in spectrum of the nasal murmur (e.g., Kurowski and Blumstein, 1993), and spectral and amplitude changes at the juncture between the nasal and the adjoining vowel (e.g., Kurowski and Blumstein, 1987a, 1987b; Ohde, Haley, and Barnes, 2006).

The **semivowels** are /w,j/ and /r,l/. The former two are known as the **glides** and the latter ones are **liquids**. The semivowels have been described by O'Connor et al. (1957), Lisker (1957a, 1957b), and Fant (1970). The bilabial glide /w/ is produced initially in the same way as the vowel /u/, with a transition into the following vowel over the course of about 100 ms. For the glide /j/, /j/, the transition into the following vowel is from /i/ and takes about the same amount of time as for /w/. The first formants of both /w/ and /j/ are on the order of 240 Hz, with the second formant transitions beginning below about 600 Hz for /w/ and above 2300 Hz for /j /. Furthermore, there appears to be relatively low-frequency frication-like noise associated with /w/ and higher-frequency noise with /j /, due to the degree of vocal tract constriction in their production.

The liquids (/r,l/) have short-duration steady-state portions of up to about 50 ms, followed by transitions into the following vowel over the course of roughly 75 ms. The /r/ is produced with some degree of lip rounding, and /l/ is produced with the tongue tip at the upper gum ridge so that air is deflected laterally. The first formants of the liquids are on the order of 500 Hz for /r/ and 350 Hz for /l/, which are relatively high, and the first and second formant transitions are roughly similar for both liquids. The major difference appears to be associated with the presence of lip rounding for /r/ but not for /l /. Since lip rounding causes the third formant to be lower in frequency, /r/ is associated with a rather dramatic third formant transition upward in frequency, which is not seen for /l /, at least not for transitions into unrounded vowels. The opposite would occur for /r/ and /l / before a rounded vowel.

DICHOTIC LISTENING AND CEREBRAL LATERALIZATION

Dichotic listening studies involve asking listeners to respond to two different signals presented at the same time, one to the right ear and another to the left ear (see, e.g., Berlin and McNeil (1976). This approach was introduced to the study of speech perception by Broadbent (1954, 1956) as a vehicle for the study of memory, and was extended to the study of hemispheric lateralization for speech by Kimura (1961, 1967). Kimura asked her subjects to respond to different digits presented to the two ears. The result was a small but significant advantage in the perception of the digits at the right ear—the **right ear advantage (REA)**. On the other hand, there was a left ear advantage when musical material was presented dichotically (Kimura, 1964).

The study of dichotic listening was enhanced with the use of CV syllables by Shankweiler and Studdert-Kennedy (1967) and others (e.g., Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy et al., 1970; Berlin et al., 1973; Cullen et al., 1974). The basic experiment is similar to Kimura's, except that the dichotic digits are replaced by a pair of dichotic CV syllables. Most often, the syllables /pa, ka, ta, ba, da, ga/ are used. The CV studies confirmed and expanded the earlier digit observations. For example, Studdert-Kennedy and Shankweiler (1970) found a significant REA for the CVs but not for vowels. They further found that the REA was larger when the consonants differed in both place of articulation and voicing (e.g., /pa/ vs. /ga/) than when the contrast was one of place (e.g., /pa/ vs. /ta/) or voicing (e.g., /ta/ vs. /da/) alone. Similarly, Studdert-Kennedy and

Shankweiler (1970) found that errors were less common when the dichotic pair had one feature (place or voicing) in common than when both place and voicing were different.

Since the primary and most efficient pathways are from the right ear to the left cerebral hemisphere and from the left ear to the right hemisphere, these have been interpreted as revealing right-eared (left hemisphere) dominance for speech and left-eared (right hemisphere) dominance for melodic material. That the left hemisphere is principally responsible for the processing of speech material is also supported by physiological findings using a variety of approaches (Wood et al., 1971; Wood, 1975; Mäkelä et al., 2003, 2005; Josse et al., 2003; Tervaniemi and Hugdahl, 2003; Price et al., 2005; Shtyrov et al., 2005).

The robustness of the REA was demonstrated by Cullen et al. (1974), who showed that the REA is maintained until the signal to the right ear is at quite a disadvantage relative to the one presented to the left. Specifically, the REA was maintained until (1) the stimuli presented to the left ear were 20 dB stronger than those to the right, (2) the signal-to-noise ratio (SNR) in the right ear was 12 dB poorer than in the left, and (3) the CVs presented to the right ear were filtered above 3000 Hz while the left ear received an unfiltered signal. Interestingly, Cullen et al. also demonstrated that when the right-ear score decreased, the left ear score actually became proportionally better so that the total percent correct (right plus left) was essentially constant. This suggests that there is a finite amount of information that can be handled at one time by the speech-handling mechanism in the left hemisphere.

When the CV delivered to one ear is delayed relative to the presentation of the CV to the other ear, then there is an advantage for the ear receiving the *lagging* stimulus, particularly for delays on the order of 30 to 60 ms (Studdert-Kennedy et al., 1970; Berlin et al., 1973). This phenomenon is the **dichotic lag effect**. Since it also occurs for nonspeech (though speech-like) sounds, there is controversy over whether the lag effect is a speech-specific event or a more general phenomenon such as backward masking (Darwin, 1971; Pisoni and McNabb, 1974; Mirabile and Porter, 1975; Porter, 1975).

Since the primary and most efficient pathways are from the right ear to the left cerebral hemisphere and from the left ear to the right hemisphere, the these have been interpreted as revealing right-eared (left hemisphere) dominance for speech and left-eared (right hemisphere) dominance for melodic material. That the left hemisphere is principally responsible for the processing of speech material is also supported by physiological findings using a variety of approaches (Wood et al., 1971; Wood et al., 1971; Wood, 1975; Mäkelä et al., 2003, 2005; Josse et al., 2003; Tervaniemi and Hugdahl, 2003;Price et al., 2005; Shtyrov et al., 2005).

CATEGORICAL PERCEPTION

Liberman et al. (1961) prepared synthetic consonant–vowel (CV) monosyllables composed of two formants each. They asked their subjects to discriminate between these pairs of synthetic CVs, as the second formant transition was varied, and obtained a finding that has had a profound effect upon the study of speech perception. Subjects' ability to *discriminate* between the two CVs in a pair was excellent when the consonants were *identifiable* as different phonemes, whereas discrimination was poor when the consonants were identified as belonging to the same *phonemic category*.

This phenomenon of **categorical perception** is illustrated in Fig. 14.10 which shows idealized results from a hypothetical study of how VOT affects the perception of initial alveolar stops. Recall that VOT is the voicing cue for initial stops, so we are dealing with the perception of /t/ versus /d/. The stimuli are CV syllables differing in VOT. For simplicity, we will identify VOTs by letters instead of actual durations in milliseconds. Two types of perceptual tasks are involved. In the first test, the amount of VOT is varied in 10 equal increments from A (the shortest) to J (the longest), and the subjects must identify the CVs as /ta/ or /da/. The upper frame of the figure shows that just about all of the shorter VOTs (A to E) were heard as /d/, whereas virtually all of the longer VOTs (F to J) were heard as /t/. In other words, there was an abrupt change in the categorization of the stimuli

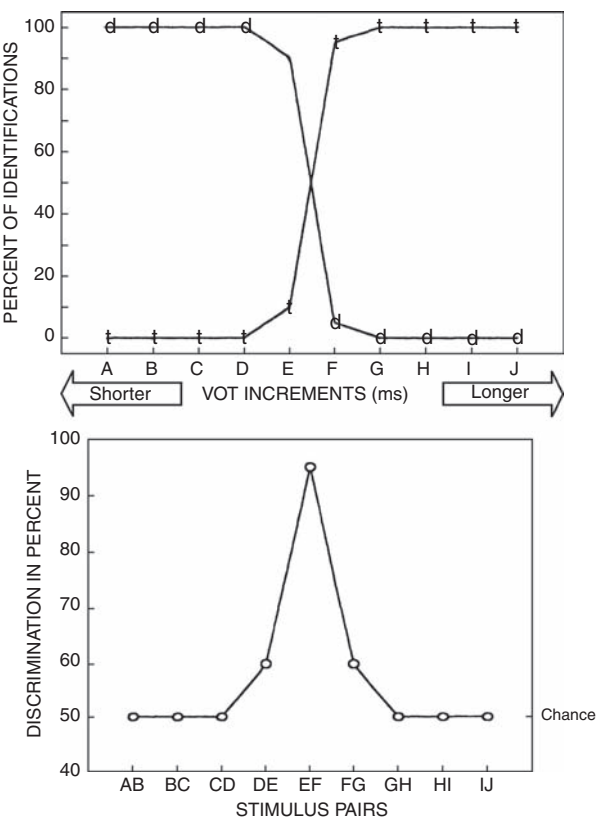

Figure 14.10 Idealized results for in a categorical perception experiment (see text).

as either voiced or voiceless between VOT increments E and F, constituting a *category boundary*. The second task is to discriminate between pairs of these stimuli; for example, between A–B, B–C, C–D, etc. Notice that the VOT difference (in milliseconds) is the same between the members of each pair. These results are shown in the lower frame of the figure, where 50% is random guessing (chance). Notice that the subjects can easily discriminate between E and F, where one member of the pair (E) is identified as /da/ and the other member of the pair (F) is identified as /ta/. On the other hand, they cannot discriminate between any of the other pairs, where both members were identified as /da/ (e.g., C vs. D) or as /ta/ (e.g., G vs. H). Thus, the CVs identified as belonging to the *same* category are poorly discriminated, whereas those identified as belonging to *different* categories are easily discriminated.

Categorical perception has been explored in many studies using a variety of speech and nonspeech (e.g., Fry et al., 1962; Lisker and Abramson, 1964, 1970; Liberman et al., 1961, 1967; Abramson and Lisker, 1970; Miller et al., 1976; Pisoni, 1977; Cutting and Rosner, 1974; Mitler et al., 1976; Stevens and Klatt, 1974; Repp, 1984; Schouten and vanHessen, 1992; for reviews, see Repp, 1984; Diehl, Lotto, and Holt, 2004). Categorical perception was originally observed for speech sounds (especially for consonants, and to a lesser extend for vowels), but not for nonspeech stimuli. Categorical perception has also been shown to occur in infants (Eimas et al., 1971; Eimas, 1974; Bertoncini et al., 1988). In addition, categorical perception is subject to **selective adaptation**, seen as changes in the boundary between categories (voiced–voiceless or different places of articulation) after hearing many repeated presentations of a prototype of just one of two opposing stimuli (e.g., Eimas and Corbit, 1973; Cooper and Blumstein, 1974). For example, hearing many repetitions of a /ba/ prototype (i.e., from the voiced end of the VOT continuum) will cause the listener's voiced–voiceless boundary between /pa/ and /ba/ to shift toward /ba/, thus favoring the perception of /pa/.

Findings like the ones just described were originally interpreted as suggesting that categorical perception reflects an innate **speech or phonetic module**, that is, an underlying process that is phonetic or speech-specific in nature rather than involving more general auditory mechanisms (e.g., Liberman et al., 1967; see below). However, the accumulated evidence has shown that categorical perception involves more general auditory mechanisms rather than phonetic or speech-specific processes. In particular, categorical perception and selective adaptation have been shown to occur for a variety of nonspeech materials, and an association of categorical perception with psychoacoustic phenomena such as perceived temporal order and across-channel gap detection (Miller et al., 1976; Pisoni, 1977; Cutting and Rosner, 1974; Tartter and Eimas, 1975; Mitler et al., 1976; Sawusch and Jusczyk, 1981; Formby et al., 1993; Nelson et al., 1995; Phillips et al., 1997; Phillips, 1999; Elangovan and Stuart, 2008). Experiments revealing categorical perception in a variety of animals provide even more impressive evidence for an underlying auditory rather than phonetic mechanism (Kuhl and Miller, 1975, 1978; Kuhl and Padden, 1982, 1983; Nelson and Marler, 1989; Dooling et al., 1995).

THE SPEECH MODULE

Whether speech perception actually involves a specialized phonetic module as opposed to general auditory capabilities is an unresolved issue. We just considered this controversy while addressing categorical perception, which was originally considered to be evidence of a specialized speech mode, but is now understood to reflect underlying auditory capabilities. Other phenomena have also been implicated in the fundamental issue of whether speech perception involves a specialized speech module or general auditory processes, such as duplex perception, the perception of sine wave speech, and the McGurk effect.

Duplex perception (e.g., Whalen and Liberman, 1987) refers to hearing *separate speech and nonspeech sounds* when the listener is presented with certain kinds of stimuli. It can be demonstrated by splitting the acoustical characteristics of a synthetic consonant–vowel syllable like /da/, and presenting them separately to the two ears as shown in the lower part of Fig. 14.11 In this version of the duplex perception experiment, one ear receives only the syllable base, composed of F1 with its transition and F2 *without* its transition. The other ear receives only the second formant transition. Neither of these sounds is heard as speech when presented alone. However, when they are presented simultaneously, the listener hears *both* a *speech* sound (/da/) in one ear and a *nonspeech* sound (a chirp) in the other ear, as illustrated in the upper part of the figure. The ability of the same stimuli to evoke separate speech and nonspeech perceptions implies the existence of separate auditory and phonetic perceptual mechanisms. One should note, however, that Fowler and Rosenblum (1991) found duplex perception for a door slamming sound, a non-speech stimulus that would not involve a specialized speech module.

The perception of **sine wave speech** has often been associated with a specialized speech mode because it reveals how the same signal can be experienced as either speech or nonspeech (e.g., Remez, Rubin, Pisoni, and Carrel, 1981; Remez, Rubin, Berns, et al., 1994).[1] Sine wave speech is a synthetic signal composed of three or more pure tones that increase and decrease in frequency over time to mimic the changing formants of naturally spoken speech, but with all other aspects of the speech signal omitted. Naïve listeners experience these signals as peculiar complex tonal patterns, but subjects will perceive them as speech when told that they are listening to intelligible computer-generated speech.

[1] On-line demonstrations of sine wave speech by Robert Remez may be found at http://www.columbia.edu/~remez/Site/Musical%20Sinewave%20Speech.html, and by Christopher Darwin at http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS/.

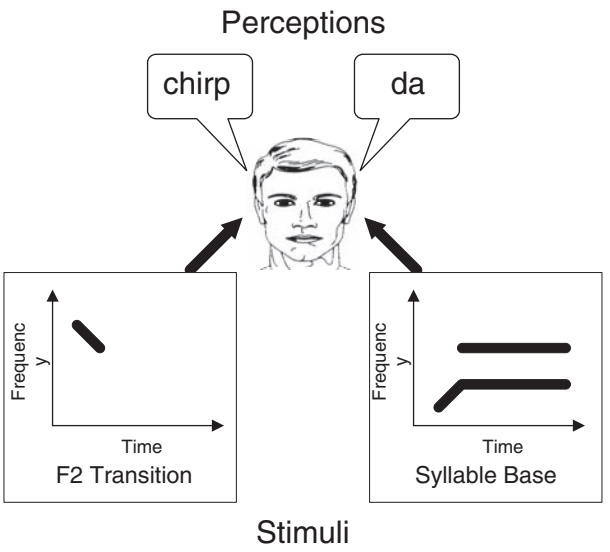

Figure 14.11 Idealized illustration of duplex perception. Presenting a syllable base without the second formant transition to one ear and just the second formant transition to other ear, causes the perception of both a speech sound (syllable /da/) in one ear and a nonspeech chirp-like sound in the other ear.

The **McGurk (McGurk–MacDonald) effect** illustrates the interaction of auditory and visual information in speech perception (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978), and has been proposed as evidence for a specialized speech module and/or the perception of speech gestures for some time (e.g., Liberman and Mattingly, 1985). To understand this effect, remember that real face-to-face conversations involve auditory and visual signals that agree: the listener *sees* lip and facial manipulations on the talker's face that correspond with what he *hears* the talker saying. For example, when a talker says /ba/, the listener hears /ba/ and also sees /ba/ on the talker's face, and, as expected, perceives /ba/. In contrast, the McGurk effect occurs when the listener is presented with *competing* auditory and visual representations of the speech. For example, the listener might be presented with an audio recording of /ba/ from earphones along with a video recording of a /ga/ on a screen. In this case, the listener perceives another syllable (e.g., /da/) or just one of the two originals, constituting the McGurk effect. This illusion is so strong that the listener even experiences it if he knows that two different stimuli were presented. The McGurk effect has been taken as evidence for a specialized speech module and/or the perception of speech gestures for some time (e.g., Liberman and Mattingly, 1985), but other interpretations also have been offered (e.g., Massaro, 1987, 1998). The accumulating physiological findings suggest that the McGurk effect involves mechanisms dealing with phonetic information, which are located in cortical auditory areas and have left hemispheric dominance (Sams et al., 1991; Näätänen, 2001; Colin et al., 2002, 2004; Möttönen et al., 2002; Saint-Amour et al., 2007).

Additional physiological evidence for a speech module has been provided by functional MRI (fMRI) studies (Benson, Whalen, Richardson, et al., 2001; Whalen, Benson, Richardson, et al., 2006), which found that changes in the complexity of speech versus nonspeech signals during passive listening resulted in different patterns of activation in the primary auditory and auditory association cortices.

POWER OF SPEECH SOUNDS

Several points about the power of speech sounds are noteworthy prior to a discussion of speech intelligibility. From the foregoing, we would expect to find most of the power of speech in the vowels, and since the vowels have a preponderance of low-frequency energy, we would expect the long-term average spectrum of speech to reflect this as well. This expectation is borne out by the literature (Fletcher, 1953). The weakest sound in English is the voiceless fricative /θ/ and the strongest is the vowel /ɔ/ (Sacia and Beck, 1926; Fletcher, 1953). If /θ/ is assigned a power of one, then the relative power of /ɔ/ becomes 680 (Fletcher, 1953). The relative power of the consonants range up to 80 for /ʃ/, are between 36 (/n/) and 73 (/ŋ/) for the nasals, are on the order of 100 for the semivowels, and range upward from 220 (/i/) for the vowels (Fletcher, 1953). As one would expect, the more powerful sounds are detected and are more intelligible at lower intensities than are the weaker ones.

The spectrograms shown earlier in this chapter show how the speech spectrum changes from moment to moment. In contrast, the spectrum of the speech signal over the long run is shown by the **long-term average speech spectrum (LTASS)**. Two sets of LTASS values are illustrated in Fig. 14.12 Here,

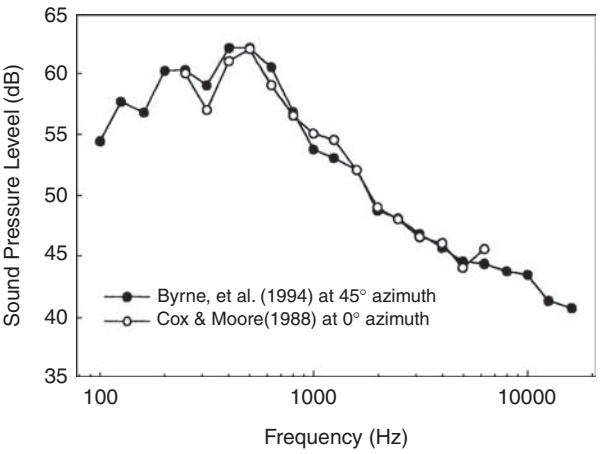

Figure 14.12 Long-term average speech spectra for speech presented at an overall level of 70 dB SPL. *Filled symbols*: composite of male and female speech samples from 12 languages at 45° azimuth (based on Byrne et al. (1994)). *Open symbols*: composite of male and female speech in English at 0° azimuth (based on Cox and Moore, 1988).

we see the sound pressure levels at each frequency (actually in third-octave bands) when the overall speech level is 70 dB SPL. (If the overall speech level is higher or lower, then the values shown on the y-axis would simply be scaled up or down proportionately.) The LTASS developed by Byrne et al. (1994) is identified by the closed symbols. This is an overall composite for male and female speech across 12 languages and was measured at an azimuth of 45°. Notice that these values are very similar to the LTASS described by Cox and Moore (1988) for combined male and female speech by using English materials presented from straight ahead of the listener (at 0° azimuth), which are identified by the open symbols.

The essential and expected implication of the material in Fig. 14.12 is that most of the energy in speech is found in the lower frequencies, particularly below about 1000 Hz, whereas intensity falls off as frequency increases above this range. It should be kept in mind that these curves show the relative speech spectrum averaged over time for male and female speakers combined. Although the LTASS tends to be similar for male and female speech in the 250 to 5000 Hz range, male levels are considerably higher at frequencies ≤160 Hz and female levels are slightly higher ≥6300 Hz (Byrne et al., 1994). The average overall sound pressure level of male speech tends to be on the order of 3 dB higher than that for females (e.g., Pearsons et al., 1977).

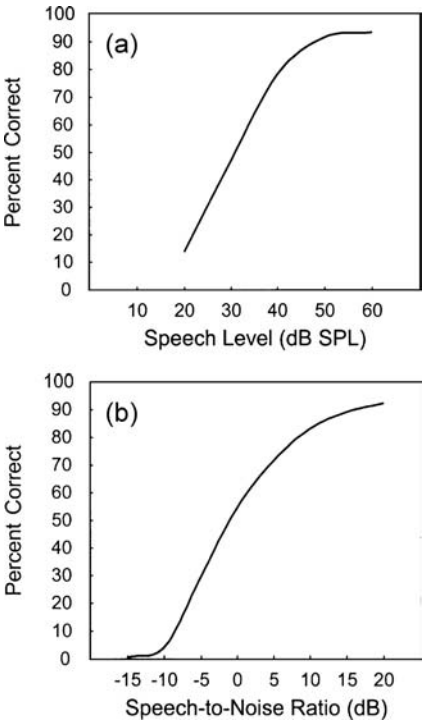

Figure 14.13 Speech recognition performance for single-syllable words improves with increasing (a) speech level and (b) speech-to-noise ratio. *Source*: Based on Gelfand (1998), used with permission.

SPEECH INTELLIGIBILITY

In general, **speech intelligibility** refers to how well the listener receives and comprehends the speech signal. The basic approach to studying speech intelligibility is quite simple and direct. The subject is presented with a series of stimuli (syllables, words, phrases, etc.) and is asked to identify what he has heard. The results are typically reported as the percent correct, which is called the **speech recognition**, **discrimination**, or **articulation score** (Campbell, 1910; Fletcher and Steinberg, 1929; Egan, 1948). The approach may be further broken down into **open set** methods requiring the subject to repeat (or write) what was heard without prior knowledge of the corpus of test items (Egan, 1948; Hirsh et al., 1952; Peterson and Lehiste, 1962), and **closed set** methods that provide a choice of response alternatives from which the subject must choose (Fairbanks, 1958; House et al., 1965). These tests were originally devised in the development of telephone communication systems. The factors that contribute to speech intelligibility (or interfere with it) may be examined by obtaining articulation scores under various stimulus conditions and in the face of different kinds of distortions.

Audibility: Speech Level and Signal-to-Noise Ratio

It is well established that speech intelligibility improves as the speech signal becomes progressively more audible (Fletcher and Steinberg, 1929; French and Steinberg, 1947; Fletcher, 1953). The dependence of speech intelligibility on the audibility is seen as an increase in speech recognition performance with increasing speech level (Fig. 14.13a) or signal-to-noise ratio (Fig. 14.13b). In other words, there are **psychometric functions** for speech intelligibility as well as for the other psychoacoustic phenomena we have discussed. As a rule, recognition performance generally becomes asymptotic when maximum intelligibility is reached for a given type of speech material; however, speech intelligibility may actually decrease if the level is raised to excessive levels.

Frequency

How much information about the speech signal is contained in various frequency ranges? The answer to this question is not only important in describing the frequencies necessary to carry the speech signal (an important concept if communication channels are to be used with maximal efficiency), but also may enable us to predict intelligibility. Egan and Wiener (1946) studied the effects upon syllable intelligibility of varying the bandwidth around 1500 Hz. They found that widening the bandwidth improved intelligibility, which reached 85% when a 3000-Hz bandwidth was available to the listener. Narrowing the passband resulted in progressively lower intelligibility; conversely, discrimination was improved by raising the level of the stimuli. That is, the narrower the band of frequencies, the higher the speech level must be in order to maintain the same degree of intelligibility.

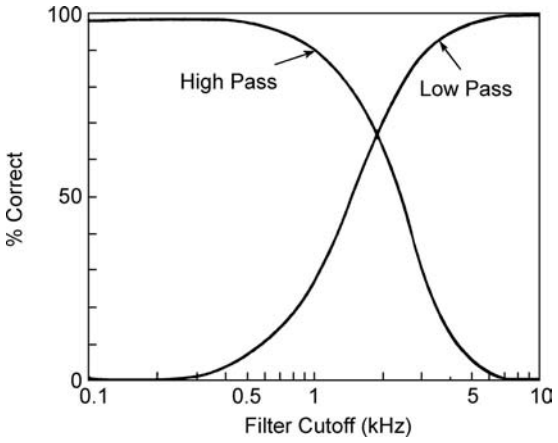

Figure 14.14 Syllable recognition as a function of high-pass and low-pass filtering. *Source*: Adapted from French and Steinberg (1947), with permission of *J. Acoust. Soc. Am.*

French and Steinberg (1947) determined the intelligibility of male and female speakers under varying conditions of low- and high-pass filtering. Discrimination was measured while filtering out the high frequencies above certain cutoff points (low-pass), and while filtering out the lows below various cutoffs (high-pass). Increasing amounts of either high- or low-pass filtering reduced intelligibility, and performance fell to nil when the available frequencies (the passband) were limited to those below about 200 Hz or above roughly 6000 Hz. As illustrated in Fig. 14.14 the high- and low-pass curves intersected at approximately 1900 Hz, where discrimination was about 68%. In other words, roughly equivalent contributions accounting for 68% intelligibility each were found for the frequencies above and below 1900 Hz. (That the frequency ranges above and below 1900 Hz each accounted for 68% intelligibility is but one of many demonstrations of the redundancy of the speech signal.) One must be careful, however, not to attach any magical significance to this frequency or percentage. For example, the crossover point dropped to about 1660 Hz only for male talkers. Furthermore, Miller and Nicely (1955) showed that the crossover point depends upon what aspect of speech (feature) is examined. Their high- and low-pass curves intersected at 450 Hz for the identification of nasality, 500 Hz for voicing, 750 Hz for frication, and 1900 Hz for place of articulation.

Amplitude Distortion

If the dynamic range of a system is exceeded, then there will be **peak-clipping** of the waveform. In other words, the peaks of the wave will be "cut off," as shown in Fig. 14.15 The resulting waveform approaches the appearance of a square wave, as the figure clearly demonstrates. The effects of clipping were studied by Licklider and colleagues (Licklider, 1946; Licklider et al., 1948; Licklider and Pollack, 1948), and the essential though surprising finding is that peak-clipping does not result in any appreciable decrease in speech intelligibility even though the waveform is quite distorted. On the other hand, if the peaks are maintained but the center portion of the wave is removed (*center-clipping*), then speech intelligibility quickly drops to nil.

Interruptions and Temporal Distortion

The effect of rapid interruptions upon word intelligibility was examined in a classical study by Miller and Licklider (1950). They electronically interrupted the speech waveform at rates from 0.1 to 10,000 times per second, and with speech–time fractions between 6.25 and 75%. The **speech–time fraction** is simply the proportion of the time that the speech signal is actually on. Thus, a 50% speech–time fraction means that the

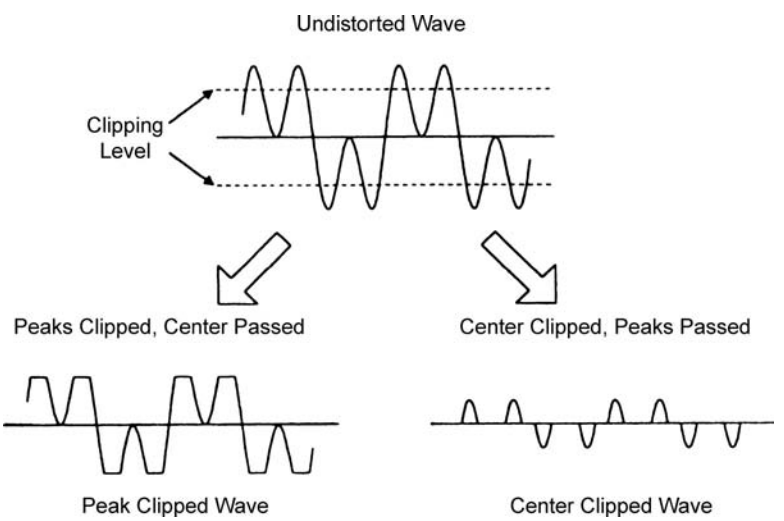

Figure 14.15 Effects of peak-clipping and center-clipping on the waveform. "Clipping level" indicates the amplitude above (or below) which clipping occurs.

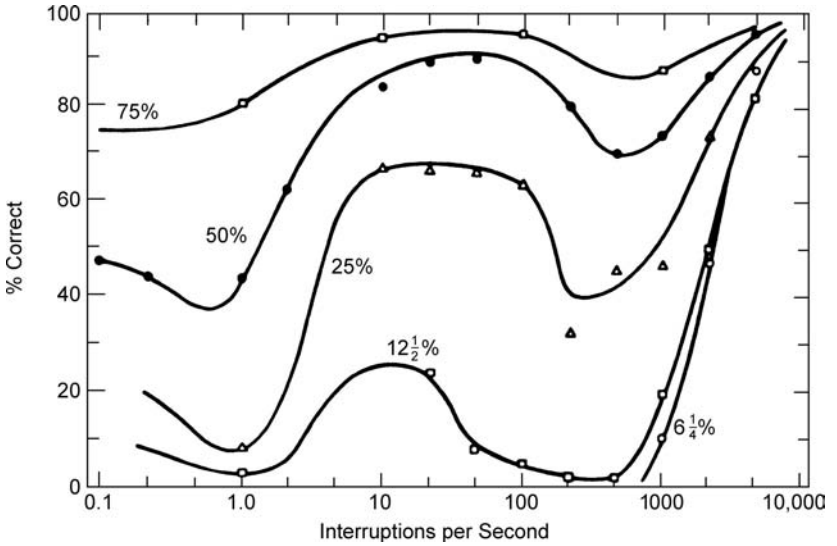

Figure 14.16 Discrimination as a function of interruption rate with speech–time fraction as the parameter. *Source*: Adapted from Miller and Licklider (1950), with permission of *J. Acoust. Soc. Am.*

speech signal was on and off for equal amounts of time, while 12.5% indicates that the speech signal was actually presented 12.5% of the time.

For the lowest **interruption rate**, the signal was alternately on and off for several seconds at a time. Thus, the discrimination score was roughly equal to the percent of the time that the signal was actually presented. The results for faster interruption rates are also shown in Fig. 14.16 When the speech–time fraction was 50%, performance was poorest when the signal was interrupted about one time per second. This is an expected finding, because we would expect intelligibility to be minimal when the interruption is long enough to overlap roughly the whole test word. At much faster interruption rates, the subjects get many "glimpses" of the test word. That is, assuming a word duration of 0.6 s and five interruptions per second, there would be about three glimpses of each word, and so forth for higher interruption rates. (The dip in the function between 200 and 2000 interruptions per second may be due to an interaction between the speech signal and the square wave that was used to modulate the speech signal to produce the interruptions.) Looking now at the remaining curves in Fig. 14.16 we find that the more the speech signal was actually on, the better the discrimination performance; whereas when the speech–time fraction fell well below 50%, intelligibility dropped substantially at all interruption rates. Essentially similar findings were reported by Powers and Speaks (1973). We see from such observations the remarkable facility with which we can "piece together" the speech signal, as well as the considerable redundancy contained within the speech waveform.

Other forms of temporal distortion also substantially decrease speech intelligibility, and this is particularly so for speeding or time compression of the speech signal (e.g., Calearo and Lazzaroni, 1957; Fairbanks and Kodman, 1957; Beasley et al., 1972). Space and scope preclude any detailed discussion except to note that intelligibility decreases progressively with speeding (or time compression) of the speech signal. An excellent review may be found in Beasley and Maki (1976).

Masking and Reverberation

The presentation of a noise has the effect of **masking** all or part of a speech signal. The general relationship between the effective level of the masker and the amount of masking for tones (Chap. 10) also holds true for the masking of speech by a broad-band noise (Hawkins and Stevens, 1950). That is, once the noise reaches an effective level, a given increment in noise level will result in an equivalent increase in speech threshold. Furthermore, this linear relationship between masker level and speech masking holds true for both the detection of the speech signal and intelligibility.

Recall from Chapter 10 that masking spreads upward in frequency, so that we would expect an intense low-frequency masker to be more effective in masking the speech signal than one whose energy is concentrated in the higher frequencies. This was confirmed by Stevens et al. (1946) and by Miller (1947). Miller also found that when noise bands were presented at lower intensities, the higher-frequency noise bands also reduced speech discrimination. This effect reflects the masking of consonant information concentrated in the higher frequencies.

Miller and Nicely (1955) demonstrated that the effect of a wide-band noise upon speech intelligibility is similar to that of low-pass filtering. This is expected, since a large proportion of the energy in the noise is concentrated in the higher frequencies. Both the noise and low-pass filtering resulted in rather

systematic confusions among consonants, primarily affecting the correct identification of place of articulation. On the other hand, voicing and nasality, which rely heavily upon the lower frequencies, were minimally affected.

Reverberation is the persistence of acoustic energy in an enclosed space after the sound source has stopped; it is due to multiple reflections from the walls, ceiling, and floor of the enclosure (normally a room). The amount of reverberation is expressed in terms of **reverberation time**, which is simply how long it takes for the reflections to decrease by 60 dB after the sound source has been turned off.

It is a common experience that intelligibility decreases in a reverberant room, and this has been demonstrated in numerous studies, as well (e.g., Knudsen, 1929; Bolt and MacDonald, 1949; Nabelek and Pickett, 1974; Gelfand and Hochberg, 1976; Nabelek, 1976; Nabelek and Robinette, 1978; Gelfand and Silman, 1979; Helfer, 1994). The amount of discrimination impairment becomes greater as the reverberation time increases, particularly in small rooms where the reflections are "tightly packed" in time.

In one sense, reverberation appears to act as a masking noise in reducing speech intelligibility; however, this is an oversimplification. The reflected energy of reverberation overlaps the direct (original) speech signal, so that perceptual cues are masked, but there are at least two distinguishable masking effects: In **overlap masking**, a subsequent phoneme is masked by energy derived from a preceding speech sound, whereas **self-masking** occurs when cues are masked within the same phoneme. In addition, reverberation distorts phoneme cues by causing a smearing of the speech signal over time, thereby also causing confusions that are not typical of masking. As a result, we are not surprised to find that, for example, stops are especially susceptible to the effects of reverberation, and final consonants are affected to a greater extent than are initial ones (e.g., Knudsen, 1929; Gelfand and Silman, 1979).

Different speech intelligibility outcomes have been associated with reverberation, masking, and the two combined: Lower percent-correct scores are obtained with noise-plus-reverberation than what would have been predicted from the scores obtained with masking alone and reverberation alone (e.g., Nabelek and Mason, 1981; Harris and Reitz, 1985; Helfer, 1992); and differences have also been found between the patterns of the perceptual errors obtained with reverberation, masking, and the two combined (e.g., Nabelek et al., 1989; Tanaka and Nabelek, 1990; Helfer, 1994). In addition, reverberation and noise have been found to produce different errors for vowels, which were often associated with the time course of the signal (Nabelek and Letowski, 1985; Nabelek and Dagenais, 1986). The student will find several informative reviews of reverberation effects in the literature (e.g., Nabelek, 1976; Helfer, 1994; Nabelek and Nabelek, 1994). With these and the preceding points in mind, one should be aware that contemporary standards for classroom acoustics call for unoccupied noise levels of 35 dBA and reverberation times of 0.4 s (40 dBA and 0.7 s for large rooms) and a SNR of +15 dB (ANSI S12.60, 2002; ASHA,2004).

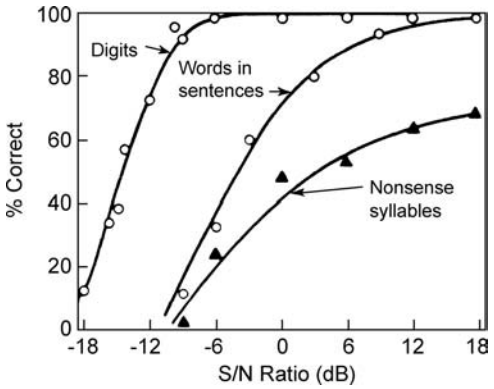

Figure 14.17 Psychometric functions showing the effects of test materials. *Source*: From Miller, Heise, and Lichten (1951), with permission of *J. Exp. Psychol.*

Nonacoustic Considerations

Speech perception depends on more than just the acoustical parameters of the speech signal. In their classical study, Miller, Heise, and Lichten (1951) asked subjects to discriminate (1) words in sentences, (2) digits from 0 to 9, and (3) nonsense (meaningless) syllables. A psychometric function was generated for each type of test material, showing percent correct performance as a function of SNR (Fig. 14.17). Note that the digits were audible at softer levels (lower SNRs) than were the words in sentences, which were in turn more accurately perceived than the nonsense syllables. We observe this result in several ways. First, at each SNR in the figure, percent-correct performance is best for the digits, less good for the words in sentences, and poorest for the syllables. Conversely, the subjects were able to repeat 50% of the digits at a level 17 dB softer than that needed to reach 50% correct for the monosyllables. Second, a small increase in SNR resulted in a substantial increase in digit discrimination but a much smaller improvement for the syllables, with improvement for word identification lying between these. Finally, notice that digit discrimination becomes asymptotic at 100% correct at low levels. On the other hand, words in sentences do not approach 100% intelligibility until the SNR reaches 18 dB, and monosyllable discrimination fails to attain even 70% at the highest SNR tested. Thus, we find that more redundant materials are more intelligible than the less redundant ones. In other words, we need only get a small part of the signal to tell one digit from another, whereas there is little other information for the subject to call upon if part of a nonsense syllable is unclear. For example, if one hears "-en" and knows in advance that the test item is a digit, the test is inordinately easier than when the initial sound might be *any* of the 25 or so consonants of English. Word redundancy falls between the exceptionally high redundancy of the digits and

the rather minimal redundancy of the nonsense syllables. As expected, Miller and associates found that words were more intelligible in a sentence than when presented alone, which also reflects redundancy afforded by the context of the test item.

In a related experiment, Miller et al. (1951) obtained intelligibility measures for test vocabularies made up of 2, 4, 8, 16, 32, or 256 monosyllabic words, as well as for an "unrestricted" vocabulary of approximately 1000 monosyllables. The results were similar to those just described. The fewer the alternatives (i.e., the more redundant or predictable the message), the better the discrimination performance. As the number of alternatives increases, greater intensity (a higher SNR) was needed in order to obtain the same degree of intelligibility.

It should be clear from these classical examples that perception of speech involves *top-down* processes as well as *bottom-up* processes are involved in the.

SPEECH PERCEPTION THEORIES AND APPROACHES

Many theories and models explaining the nature of speech perception have been proposed over the years, and we will give an overview of some of the key features of several of these theories here. Students interested in pursuing these issues will find a number of contemporary reviews which approach the topic and its issues from a variety of perspectives (e.g., Jusczyk and Luce, 2002; Diehl et al., 2004; Remez, 2005; Galantucci et al., 2006; Pardo and Remez, 2006; Massaro and Chen, 2008). Notice while reading this material that most approaches to speech perception involve the interplay of both the incoming signal (*bottom-up processing*) as well as higher level cognitive influences (*top-down processing*).

Models Implicating the Production System

Motor Theory

Perhaps the most widely known speech perception theory is Liberman's motor theory, the details of which have evolved over the years (Liberman, 1996; Liberman et al., 1957, 1967; Liberman and Mattingly, 1985, 1989; Mattingly and Liberman, 1988; Liberman and Whalen, 2000). Recall that coarticulation causes a particular phonetic element to have different acoustical characteristics depending on its context (e.g., different formant transitions for /d/ in /di/ vs. /du/). Motor theory proposes that speech perception involves identifying the *intended speech gestures* (effectively the neuromotor instructions to the articulators) that resulted in the acoustical signal produced by the speaker and heard by the listener. In other words, we perceive the invariant intended phonetic gestures (e.g., release of the alveolar closure in /di/ and /du/) that are encoded in the variable acoustical signals. This perceptual process involves biologically evolved interactions between the speech *perception and production* systems, and is accomplished by a *specialized speech* or *phonetic module (mode)* in the central nervous system.

Direct Realist Theory

The **direct realist theory** developed by Fowler and colleagues (e.g., Fowler, 1986, 1991, 1996; Galantucci et al., 2006) is related to motor theory but certainly distinct from it. It is similar to motor theory in the sense that direct realist theory involves the perception of speech gestures and incorporates interactions with the motor system in speech perception. However, it differs from motor theory by making use of the sound signal reaching the listener to recover the *actual* articulatory gestures that produced them (as opposed to intended gestures), and does not involve a biologically specialized phonetic module.

Analysis-by-Synthesis

The speech production system is also involved in the **analysis-by-synthesis theory** (e.g., Stevens and Halle, 1967; Stevens, 1972), although the process is somewhat different from those of the motor and direct realist theories. Here, the perceptual decision about the speech signal is influenced by considering the articulatory manipulations that the listener might use to produce them, with the decision based on which of these gives the best match to the signal.

General Auditory Approaches

The **general auditory approaches** (Diehl et al., 2004) include a variety of descriptions and theories that address speech perception in terms of *auditory capabilities* and *perceptual learning* rather than the gesture perception and related mechanisms involved in the motor and direct realist theories. Having already addressed some of the findings supporting the notion that speech perception makes use of general auditory capabilities rather than a speech-specific phonetic module, let us turn our attention to the contribution of perceptual learning (for concise reviews see, e.g., Jusczyk and Luce, 2002; Diehl et al., 2004).

The effects of perceptual learning on speech perception is illustrated by comparing the speech sound discriminations of younger versus older infants. Recall that adults typically discriminate speech sound differences across the phoneme categories of their language, but not within these categories. Infants less than about six months of age can discriminate speech sounds within or across the phoneme categories of their language environment, but they become less responsive to differences falling within the phoneme categories of their language over the course of the next year or so (Werker, Gilbert, Humphrey, and Tees, 1981; Werker and Tees, 1984; Best, McRoberts, and Sithole, 1988; Pegg and Werker, 1997). As conceptualized by Kuhl and colleagues (e.g., Kuhl, 1991; Kuhl, Williams, Lacerda, et al., 1992), infants over about six months of age begin to employ phoneme category prototypes, which act as **perceptual magnets**. Here, speech patterns relatively close to the prototype are perceptually drawn to it and thus perceived as the same, whereas speech patterns sufficiently far from the prototype are perceived as different.

Fuzzy Logical Model of Perception

The **fuzzy logical model of perception** (**FLMP**; e.g., Ogden and Massaro, 1978; Massaro, 1987, 1998. Massaro and Chen, 2008) may be viewed as an example of a general auditory approach. Unlike the gesture perception approach of the motor and direct realist theories, the FLMP involves evaluating the auditory (and visual) features [2] of the stimulus and comparing them to prototypes of alternative speech categories in the listener's long term memory. Speech perception in this model involves three processes. (1) The evaluation process involves analyzing the features of the signal. The term "fuzzy" is used because the features are valued along a continuum from 0 to 1.0 instead being assessed an all-or-none (present/absent) basis. (2) The integration process involves comparing the features with possible prototypes in the listener's long-term memory. (3) A decision is then made, which involves choosing the prototype with the best match to the features.

Word Recognition Models

Let us now briefly consider a number of speech perception models that concentrate on word recognition.

Prototype and Exemplar Models

Some word recognition approaches involve comparing the acoustical characteristics of the speech signal (as opposed abstract representations like phonetic features) to internalized perceptual references in the listener's long-term memory. In the **Lexical Access from Spectra (LAFS) model** (Klatt, 1989), the incoming speech spectra are compared to learned prototypes or templates. On the other hand, **exemplar models** involve comparing the incoming signal to all existing instances of the category in the listener's long-term memory. (Johnson, 1997).

Logogen Model

The **Logogen model** (Morton, 1969) envisions the existence of recognition units called **logogens**.[3] The activation levels of the logogens increase as they accumulate acoustic, visual, and semantic information from the incoming signal, and a given logogen is triggered once its activation level reaches a certain threshold, at which point the corresponding word is recognized by the listener. Logogens associated with more commonly encountered words have lower thresholds so that higher frequencies words are more likely to be recognized than lower thresholds words.

Cohort Model

Word recognition in the **Cohort model** (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980) involves progressively reducing the viable alternatives in the listener's lexicon until a decision can be reached and is based on both a bottom-up analysis of the sound pattern and top-down considerations such as syntactic and semantic constraints on the possible alternatives. Consider the word *remarkable*. The /r/ activates a *cohort* of all words in the listener's lexicon beginning with that sound. Then, the cohort is progressively narrowed with each successive aspect of the word over time. For example, /ri/ limits the cohort to words beginning with /ri/ (*read*, *real*, *recent*, *relax*, *remarkable*, etc.); /rim/ narrows the cohort to words like *ream*, *remember*, *remark*, *remarkable*, etc.; /rimɑrk/ limits it to just *remark*, *remarks*, *remarked*, *remarking*, and *remarkable*; and /rimɑrkə/ finally reduces the cohort to *remarkable* (which is thus chosen). It is noteworthy that the various competing alternatives do not inhibit each other at each stage of the analysis. For example, activation of *remarkable* by /rim/ does not affect activation of *remember*, which falls out of the corpus when the stimulus analysis reaches /rimɑrk/. Bottom-down considerations are easily understood by considering how the cohort of possible alternatives is delimited by semantic and syntactic considerations when the word *remarkable* appears different sentences (e.g., *That magic trick was remarkable.* vs. *The remarkable event that I will describe is....*).

Trace Model

The **Trace model** of word recognition (Elman and McClelland, 1986; McClelland and Elman, 1986; Elman, 1989) is a connectionist model, which means that it involves interconnected elements that influence each other. The elements of a connectionist model are called *units*, which can be activated to a greater or lesser degree, and the connections between the units are called *links*. Signals from *excitatory* links increase a unit's activation level, and signals from *inhibitory* links decrease the activation level. The recognition of a particular word in the Trace model involves three levels of units (features, phonemes, and word), with interactions both within and across these levels. Interactions within a particular level are inhibitory so that the representation of one unit (e.g., /t/ at the phoneme level, or *bite* at the word level) prevents activation of competing units (e.g., /g/ or /m/ at the phoneme level, or *sight* at the word level). On the other hand, interactions across levels are excitatory (e.g., the representation of voicelessness at the feature level enhances the representation of voicelessness at other levels as well).

Shortlist Model

The **Shortlist model** (Norris, 1994) is a bottom-up connectionist approach to word recognition. In this two-stage model, the incoming speech signal activates a "short list" of viable word choices on a bottom-up basis (that is analogous to Cohort but unlike Trace), which are then subjected to inhibitory competition (that is unlike Cohort but similar to Trace).

[2] Consideration of the auditory and visual features of the stimulus allows FLMP to account for the McGurk effect.

[3] The term *logogen* was coined by Hallowell Davis based on *logos* for *word* and *genus* for *birth* (Morton (1969).

Neighborhood Activation Model

The **neighborhood activation model (NAM**; Luce, 1990; Kirk et al., 1995; Luce and Pisoni, 1998) attempts to account for how lexical neighborhoods and word frequency affect the identification of a word. The **lexical neighborhood** of a word is comprised of similar sounding alternatives. Words that have many similar sounding alternatives have *dense lexical neighborhoods*, and words with few similar sounding alternatives have *sparse lexical neighborhoods* (Luce, 1990; Kirk et al., 1995; Luce and Pisoni, 1998). Confusions are more likely to occur when there are many viable (similar sounding) alternatives, so that words with denser lexical neighborhoods are more difficult to recognize than other words with sparser lexical neighborhoods. **Word frequency** comes into play because we are significantly biased in favor of more frequently occurring words over those of lower frequency (e.g., Rosenzweig and Postman, 1957). According to the NAM, the sound patterns of a word are compared to acoustic–phonetic representations in the listener's memory. The probability of a representation being activated depends the degree to which it is similar to the stimulus. The next step is a lexical selection process among the words in memory that are potential matches to the stimulus, which is biased according to word frequency. A connectionist variation of NAM called **PARSYM** (Luce, Stephen, Auer, and Vitevitch, 2000; Auer and Luce, 2005) also accounts probabilities of occurrence of different allophones in a various positions.

Speech Intelligibility and Acoustical Measurements

Speech intelligibility under given conditions can be estimated or predicted using a number of acoustical methods, such as the articulation index and the speech transmission index. The **articulation index (AI)** was introduced by French and Steinberg (1947). The AI estimates speech intelligibility by considering how much of the speech signal is audible above the listener's threshold as well as the signal-to-noise ratio. In its original formulation, the basic concept of the AI involves the use of 20 contiguous frequency bands, each of which contributes the same proportion (0.05 or 5%) to the overall intelligibility of the message. These bands are then combined into a single number from 0 to 1.0, which is the articulation index.

In general, a given band is given full credit if all of the speech signal it contains is above threshold and also has a high enough signal-to-noise ratio. Assuming that the speech level in a band is well above threshold, then it would receive given full credit (0.05) if its SNR is at least +18 dB, and would receive partial credit for poorer SNRs down to −12 dB, where that band's contribution is zero. The resulting part of the band's potential value of 0.05 is its contribution; the sum of the 20 values (one from each band) becomes the AI, which therefore has a range between 0 and 1.0. (The interested reader should consult the sources mentioned in this section for details of how to calculate the various versions of the AI.)

French and Steinberg's original AI, which employed 20 equally weighted bandwidths, has been modified in various ways since its original description (e.g., Beranek, 1947; Kryter, 1962a, 1962b, 1985; ANSI S3.5-1969, [R1986]; Pavlovic, Studebaker, and Sherbecoe, 1986; Pavlovic, 1987; ANSI S3.5-1997[R2007]; Rhebergen and Versfeld, 2005). The current version, known as the **speech intelligibility index (SII)** involves the use of standard third-octave and octave-bands, adjustments in importance weightings given to each band, as well as other modifications (ANSI S3.5 1997[R2007]).

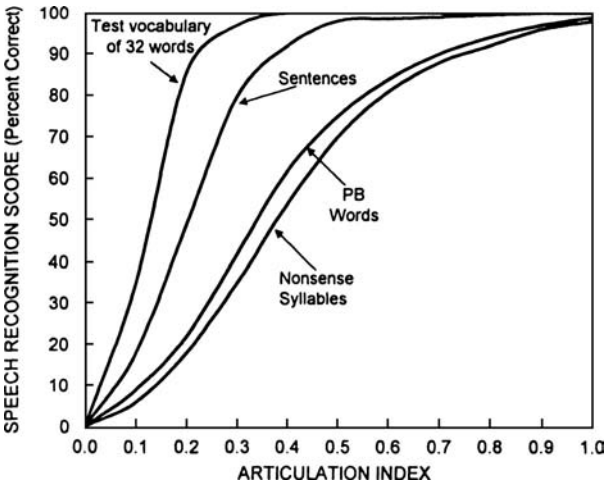

Figure 14.18 Relationship between the articulation index and speech recognition scores for selected speech materials. *Source*: Based on Kryter (1962b, 1985) and ANSI (1969, R1986).

The articulation index and the speech intelligibility are reasonably good predictors of actual speech recognition performance for a variety of speech materials (e.g., Kryter, 1962a, 1962b, 1985; Rhebergen and Versfeld, 2005). Figure 14.18 shows some examples of the way in which the AI is related to intelligibility scores for various kinds of speech materials. Notice that an AI of 0.5 corresponds to speech recognition scores of about 70% for nonsense syllables, 75% for phonetically balanced (PB) monosyllabic test words, 97% for sentences, and 100% when the test vocabulary is limited to only 32 words. The student should consider these differences in light of the discussion of the effects of test materials (and Fig. 14.17) earlier in this chapter. Beranek (1954/1986) proposed that the conditions for speech communication can be guesstimated as probably satisfactory when the AI is higher than 0.6 and most likely unsatisfactory when the AI is lower than 0.3. According to the material in Fig. 14.18 an AI of 0.6 is associated with speech recognition scores of approximately 98% for sentences and 85% for words, whereas speech intelligibility falls to about 80% for sentences and 45% for words when the AI is only 0.3.

The SII is based on the use a steady noise, but speech communication often occurs against a background of noise that fluctuates over time. To address this limitation, Rhebergen and Versfeld (2005) developed a modification of the SII for that can be used with fluctuating noises. Fundamentally, their

modification involves obtaining individual SII values within many successive time frames, which are then averaged to arrive at the overall SII value. They found that their approach produces more appropriate SII values for fluctuating noises and the same results as the standard SII for steady noises. With this method, Rhebergen and Versfeld showed that 50% correct sentence reception in noise occurs when the SII is 0.35, corresponding to SNRs of about −4.5 dB for steady noises and −12 dB for fluctuating noises.

The articulation index has been applied to the speech recognition of the hearing-impaired in a variety of ways. Although this topic is outside of the current scope, the interested student will find numerous papers dealing with this topic throughout the current literature (e.g., Pavlovic et al., 1986; Steeneken and Houtgast, 1980; Kamm et al., 1985; Humes et al., 1986; Pavlovic, 1988, 1991).

Beranek (1954/1986) introduced a simplified modification of the articulation index that estimates the amount of noise that will just allow speech communication to take place at various levels of vocal effort, and distances between the talker and listener is known as the **speech interference level (SIL)**. The SIL is simply the average of the noise levels that occur in three or four selected bands. The 500, 1000, 2000, and 4000 Hz octave bands are used in the current standard version of the SIL (ANSI, 1977, R1986). Several informative discussions of the SIL are available to the interested reader (e.g., Beranek, 1954/1986; Webster, 1978; ANSI-S3.14-1977 (R1997); Lazarus, 1987).

Another approach to estimate speech intelligibility from acoustical measurements is the **speech transmission index (STI)**, which is based upon the **modulation transfer function (MTF)**. This technique was originated by Steeneken and Houtgast (1980). In addition to their work, the interested reader should also refer to informative sources (e.g., Humes et al., 1986; Anderson and Kalb, 1987; Schmidt-Nielsen, 1987; Steeneken, 2006; van Wijngaarden and Drullman, 2008). An important advantage of the STI is that it accounts for the effects of all kinds of noises and distortions that affect the speech signal, including reverberation and other aberrations that occur over time.

Determining the STI begins by obtaining MTF results in the octave-bands from 125 to 8000 Hz. These results are used to produce a *transmission index* for each of the octave-bands, which are in turn adjusted by weighting factors that account for the importance of each band for speech communication. The weighted results are then combined to arrive at an STI, which can range in value from 0 to 1.0. Fig. 14.19 shows the STI and speech recognition performance for reprehensive kinds of test materials.

The **rapid speech transmission index (RASTI)** is an efficient and relatively simple method for making STI measurements using special instrumentation designed for this purpose (e.g., Bruel and Kjaer, 1985; IEC, 1987). A loudspeaker is placed in the room or other environment being tested at the location where a talker would be, and a microphone is placed at the listener's location. The signal arriving at the microphone thus incorporates all of the ways in which the original signal has been modified by the noise, reverberation, and other acoustical features of the room. The RASTI equipment then provides the results as a value ranging from 0 to 1.0. Orfield (1987) proposed that relative speech intelligibility may be inferred from the outcome of RASTI testing as follows: RASTI values of 0.75 or higher may be considered *excellent*; 0.6–0.74 are *good*; 0.45–0.59 are *fair*; 0.3–0.44 are *poor*; and values of 0.29 or lower are considered *bad*.

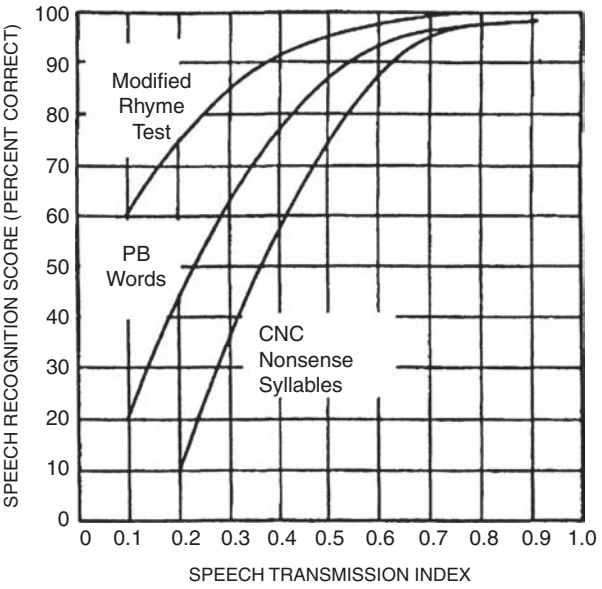

Figure 14.19 Relationship between the speech transmission index and speech recognition scores for selected speech materials. *Source*: Adapted from Anderson and Kalb (1987) with permission of *J. Acoust. Soc. Am.*

CLEAR SPEECH

We know from common experience that we sometimes modify the way we talk when trying to maximize the clarity of our speech. This kind of speech is appropriately called **clear speech** (Picheny, Durlach, and Braida, 1985). Clear speech is used when we are trying to make our speech as intelligible as possible for the benefit of a hearing-impaired listener, or perhaps when speaking under adverse acoustical conditions.

The acoustical characteristics and perceptual impact of clear speech have been described in considerable detail (Picheny et al., 1985, 1986, 1989; Moon and Lindblom, 1994; Payton et al., 1994; Schum, 1996; Uchanski et al., 1996; Liu and Zeng, 2004, 2006; Krause and Braida, 2002, 2004; Liu, Del Rio, Bradlow, and Zeng, 2004; Kain, Amano-Kusumoto, and Hosom, 2008). Several of the acoustical distinctions between clear and conversational speech may be seen by comparing the two spectrograms

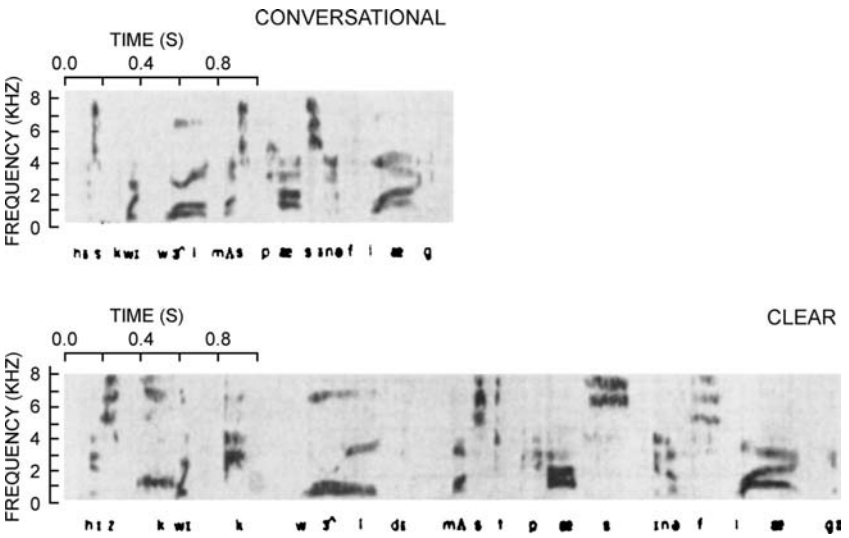

Figure 14.20 Spectrographic examples of conversational speech (above) and clear speech (below) produced by the same talker. *Source*: From Picheny, Durlach, and Braida (1986), with permission of the American Speech-Language-Hearing Association.

in Fig. 14.20. Clear speech involves a decrease in speaking rate so that it is slower or longer in duration than conversational speech. However, the durational difference is more than simply a matter of talking slowly. The increased duration of clear speech is due to *both* the insertion of more and longer pauses *and* increases in the durations of many of the individual speech sounds. In addition, instead of a uniform lengthening of all speech sounds, duration increases depend on the characteristics of the phonemes and their acoustical contexts. Moreover, there also tends to be greater variation in fundamental frequency and a greater degree of temporal modulation. Differences in nature of the phonemes produced during clear and conversational speech are seen, as well. For example, all stop bursts and most final position consonants are released in clear speech, whereas this usually does not occur in conversational speech. Also, there are much less vowel modifications (e.g., vowel reduction) in clear speech than in conversational speech.

Another distinction is that clear speech involves greater amplitudes in the higher frequencies, with increased intensity for obstruent sounds, especially for the stops, which can be as much as 10 dB higher than in conversational speech. Related to this is the point that the relative power of consonants to vowels (consonant-to-vowel ratio) increases during clear speech. This contrasts with what occurs during loud conversational speech, where the consonant-to-vowel ratio decreases.

Does clear speech successfully accomplish its goal of providing improved intelligibility? The answer to this question is clearly yes. Improved speech intelligibility for clear speech compared to conversational speech has been a uniform finding under a variety of listening situations, including such adverse conditions as noise and/or reverberation, and these advantages are enjoyed by normal-hearing as well as those with hearing loss and learning disabilities (Picheny et al., 1985, 1989; Payton et al., 1994; Schum, 1996; Uchanski et al., 1996; Helfer, 1997; Bradlow and Bent, 2002; Ferguson and Kewley-Port, 2002; Gagne et al., 1995; Krause and Braida, 2002; Bradlow, Kraus, and Hayes, 2003; Liu et al., 2004).

REFERENCES

Abramson, AS, Lisker, L. 1970. *Discriminability along the voicing continuum: Cross-language tests*. Proceedings of the Sixth International Congress on Phonetic Science. Prague, Czech Republic: Academia, 569–573.

American National Standards Institute (ANSI). 1986. *S3.5-1969 (R1986): American national standard methods for the calculation of the articulation index*. New York, NY: ANSI.

American National Standards Institute (ANSI). 1997. *ANSI-S3.14-1977(R1997): American national standard for rating noise with respect to speech interference*. New York, NY: ANSI.

American National Standards Institute (ANSI). 2002. S12.60-2002 *Acoustical performance criteria, design requirements, and guidelines for schools*. New York, NY: ANSI.

American National Standards Institute (ANSI). 2007. *S3.5-1997(R2007) American national standard methods for calculation of the speech intelligibility index*. New York, NY: ANSI.

American Speech-Language-Hearing Association (ASHA). 2004. Acoustics in educational settings: Position statement. Available at http://www.asha.org/about/leadership projects/LC/spring04/ LCAHS12004PS.htm.

Anderson, BW, Kalb, JT. 1987. English verification of the STI method for estimating speech intelligibility of a communications channel. *J Acoust Soc Am* 81, 1982–1985.

Arkebauer, JH, Hixon, TJ, Hardy, JC. 1967. Peak intraoral air pressure during speech. *J Speech Hear Res* 10, 196–208.

Auer, ET Jr, Luce, PA. 2005. Probabilistic phonotactics in spoken word recognition. In: DB Pisoni, RE Remez (eds), *The Handbook of Speech Perception.* Oxford, MA: Blackwell, 610–630.

Beasley, DS, Maki, JE. 1976. Time- and frequency-altered speech. In: EC Carterette, MP Friedman (eds.), *Handbook of Perception. Vol. 7: Language and Speech.* New York, NY: Academic Press, 419–458.

Beasley, DS, Schwimmer, S, Rintelmann, WF. 1972. Intelligibility of time-compressed CNC monosyllables. *J Speech Hear Res* 15, 340–350.

Behrens, S, Blumstein, SE. 1988. On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *J Acoust Soc Am* 84, 861–867.

Benson, R, Whalen, DH, Richardson, M, Swainson, B, Clark, VP, Lai, S, Liberman, AM. 2001. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain Lang* 78, 364–396.

Beranek, LL. 1947. The design of communication systems. *IRE Proc* 35, 880–890.

Beranek, LL. 1954/1986. *Acoustics.* New York, NY: Acoustical Society of America.

Berlin, CI, McNeil, MR. 1976. Dichotic listening. In: NJ Lass (ed.), *Contemporary Issues in Experimental Phonetics.* New York, NY: Academic Press, 327–387.

Berlin, CI, Lowe-Bell, SS, Cullen, JK, Thompson, CL, Loovis, CF. 1973. Dichotic speech perception: An interpretation of right-ear advantage and temporal-offset effects. *J Acoust Soc Am* 53, 699–709.

Bertoncini, J, Bijeljas-Babic, R, Jusczyk, P, Kennedy, L, Mehler, J. 1988. An investigation of young infants' perceptual representations of speech sounds. *J Exp Psychol: Gen* 117, 21–33.

Best, CT, McRoberts, GW, Sithole, NM. 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *J Exp Psychol Hum Percept Perform* 14, 345–360.

Blumstein, SE, Stevens, KN. 1979. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J Acoust Soc Am* 66, 1001–1017.

Blumstein, SE, Stevens, KN. 1980. Perceptual invariance and onset spectra for stop consonants indifferent vowel environments. *J Acoust Soc Am* 67, 648–662.

Blumstein, SE, Isaacs, E, Mertus, J. 1982. The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *J Acoust Soc Am* 72, 43–50.

Bolt, RH, MacDonald, AD. 1949. Theory of speech masking by reverberation. *J Acoust Soc Am* 21, 577–580.

Bradlow, AR, Bent, T. 2002. The clear speech effect for nonnative listeners. *J Acoust Soc Am* 112, 272–284.

Bradlow, AR, Kraus, N, Hayes, E. 2003. Speaking clearly for children with learning disabilities: Sentence perception in noise. *J Speech Lang Hear Res* 46, 80–97.

Broadbent, DE. 1954. The role of auditory localization in attention and memory span. *J Exp Psychol* 47,191–196.

Broadbent, DE. 1956. Successive responses to simultaneous stimuli. *Q J Exp Psychol* 8, 145–152.

Bruel and Kjaer. 1985. *The modulation transfer function in room acoustics. RASTI: A tool for evaluating auditoria.* Technical Review No. 3-1985. Massachusetts: Bruel & Kjaer.

Byrne, D, Dillon, H, Tran, K, Arlinger, K, Cox, R, Hagerman, B, Hetu, R, Kei, J, Lui, C, Kiessling, J, Kotby, MN, Nasser, N, El Kholy, AH, Nakanishi, Y, Oyer, H, Powell, R, Stephens, D, Meredith, R, Sirimanna, T, Tavartkiladze, G, Frolenkov, GI, Westerman, S, Ludvigsen, C. 1994. An international comparison of long-term average speech spectra. *J Acoust Soc Am* 96, 2108–2120.

Calearo, C, Lazzaroni, A. 1957. Speech intelligibility in relation to the speed of the message. *Laryngoscope* 67, 410–419.

Campbell, GA. 1910. Telephonic intelligibility. *Philos Mag* 19, 152–159.

Colin, C, Radeau, M, Soquet, A, Demolin, D, Colin, F, Deltenre, P. 2002. Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clin Neurophysiol* 113, 495–506.

Colin, C, Radeau, M, Soquet, A, Deltenre, P. 2004. Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clin Neurophysiol* 115, 1989–2000.

Cooper, FS, Delattre, PC, Liberman, AM, Borst, JM, Gerstman, LJ. 1952. Some experiments on the perception of synthetic speech sounds. *J Acoust Soc Am* 24, 597–617.

Cooper, WE, Blumstein, SA. 1974. A "labial" feature analyzer in speech perception. *Percept Psychophys* 15, 591–600.

Cox, RM, Moore, JN. 1988. Composite speech spectrum for hearing aid gain prescriptions. *J Speech Hear Res* 31, 102–107.

Cullen, JK, Thompsen, CL, Hughes, LF, Berlin, CI, Samson, DS. 1974. The effects of various acoustic parameters on performance in dichotic speech perception tasks. *Brain Lang* 1, 307–322.

Cutting, JE, Rosner, BS. 1974. Categories and boundaries in speech and music. *Percept Psychophys* 16, 564–570.

Darwin, CJ. 1971. Dichotic backward masking of complex sounds. *Q J Exp Psychol* 23, 386–392.

Delattre, PC, Liberman, AM, Cooper, FS. 1955. Acoustic loci and transitional cues for consonants. *J Acoust Soc Am* 27, 769–773.

Delattre, PC, Liberman, AM, Cooper, FS, Gerstman, LJ. 1952. An experimental study of the acoustic determinants of vowel color; Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word* 8, 195–210.

Diehl, RL, Lotto, AJ, Holt, LL. 2004. Speech perception. *Ann Rev Psychol* 55, 149–179.

Dooling, RJ, Best, CT, Brown, SD. 1995. Discrimination of synthetic full-formant and sinewave /ra-la/ continua by budgerigars (*Melopsittacus undulatus*) and zebra finches (*Taeniopygia guttata*). *J Acoust Soc Am* 97, 1839–1846.

Egan, JP. 1948. Articulation testing methods, *Laryngoscope* 58, 955–981.

Egan, JP, Wiener, FM. 1946. On the intelligibility of bands of speech in noise. *J Acoust Soc Am* 18, 435–441.

Eimas, P. 1974. Auditory and linguistic processing of cues for place of articulation by infants. *Percept Psychophys* 16, 513–521.

Eimas, P, Corbit, JD. 1973. Selective adaptation of linguistic feature detectors. *Cogn Psychol* 4, 99–109.

Eimas, P, Siqueland, P, Jusczyk, P, Vigorito, J. 1971. Speech perception in infants. *Science* 171, 303–306.

Elangovan, S, Stuart, A. 2008. Natural boundaries in gap detection are related to categorical perception of stop consonants. *Ear Hear* 29, 761–774.

Elman, JL. 1989. Connectionist approaches to acoustic/phonetic processing. In: WD Marslen-Wilson (ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press, 227–260

Elman, JL, McClelland, JL. 1986. Exploiting lawful variability in the speech wave. In: JS Perkell, DH Klatt (eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum, 360–380.

Fairbanks, G. 1958. Test of phonemic differentiation: The rhyme test. *J Acoust Soc Am* 30, 596–600.

Fairbanks, G, Kodman, F. 1957. Word intelligibility as a function of time compression. *J Acoust Soc Am* 29, 636–641.

Fant, G. 1970. *Acoustic Theory of Speech Perception.* The Hague: Mouton.

Ferguson, SH, Kewley-Port, D. 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 112, 259–271.

Ferrand, CT. 2007. *Speech Science.* 2nd ed. Boston, MA: Allyn and Bacon.

Fischer-Jorgensen, E. 1954. Acoustic analysis of stop consonants. *Misc Phonet* 2, 42–59.

Flanagan, JL. 1958. Some properties of the glottal sound source. *J Speech Hear Res* 1, 99–116.

Flanagan, JL. 1972. *Speech Analysis Synthesis and Perception.* New York, NY: Springer-Verlag.

Fletcher, H. 1953. *Speech and Hearing in Communication.* New York, NY: Van Nostrand.

Fletcher, H, Steinberg, JC. 1929. Articulation testing methods. *Bell Sys Tech J* 8, 848–852.

Formby, C, Barker, C, Abbey, H, Raney, JJ. 1993. Detection of silent temporal gaps between narrow-band noise markers having second-formant like properties of voiceless stop/vowel combinations. *J Acoust Soc Am* 93, 1023–1027.

Fowler, CA. 1986. An event approach to the study of speech perception from a direct realist perspective. *J Phon* 14, 3–28.

Fowler, CA. 1991. Auditory perception is not special: We see the world, we feel the world, we hear the world. *J Acoust Soc Am* 89, 2910–2915.

Fowler, CA. 1996. Listeners do hear sounds, not tongues. *J Acoust Soc Am* 99, 1730–1741.

Fowler, CA, Rosenblum, LD. 1991. Duplex perception: A comparison of monosyllables and slamming doors. *J Exp Psychol Hum Percept Perform* 16, 742–754.

French, NR, Steinberg, GC. 1947. Factors governing the intelligibility of speech. *J Acoust Soc Am* 19, 90–114.

Fry, D, Abramson, A, Eimas, P, Liberman, AM. 1962. The identification and discrimination of synthetic vowels. *Lang Speech* 5, 171–189

Fujimura, O. 1962. Analysis of nasal consonants. *J Acoust Soc Am* 34, 1865–1875.

Furui, S. 1986. On the role of spectral transition for speech perception. *J Acoust Soc Am* 80, 1016–1025.

Gagne, J, Querengesser, C, Folkeard, P, Munhall, K, Mastern, V. 1995. Auditory, visual and audiovisual speech intelligibility for sentence-length stimuli: An investigation of conversational and clear speech. *Volta Rev* 97, 33–51.

Galantucci, B, Fowler, CA, Turvey, MT. 2006. The motor theory of speech perception reviewed. *Psychon Bull Rev* 13, 361–377.

Gelfand, SA. 1998. Optimizing the reliability of speech recognition scores. *J Speech Lang Hear Res* 41, 1088–1102.

Gelfand, SA, Hochberg, I. 1976. Binaural and monaural speech discrimination under reverberation. *Audiology* 15, 72–84.

Gelfand, SA, Silman, S. 1979. Effects of small room reverberation upon the recognition of some consonant features. *J Acoust Soc Am* 66, 22–29.

Guerlekian, JA. 1981. Recognition of the Spanish fricatives /s/ and /f/. *J Acoust Soc Am* 79 1624–1627.

Haggard, MP, Ambler, S, Callow, M. 1970. Pitch as a voicing cue. *J Acoust Soc Am* 47, 613–617.

Halle, M, Hughes, GW, Radley, JPA 1957. Acoustic properties of stop consonants. *J Acoust Soc Am* 29, 107–116.

Harris, KS. 1958. Cues for the discrimination of American English fricatives in spoken syllables. *Lang Speech* 1, 1–7.

Harris, KS, Hoffman, HS, Liberman, AM, Delattre, PC, Cooper, FS. 1958. Effect of third formant transitions on the perception of the voiced stop consonants. *J Acoust Soc Am* 30, 122–126.

Harris, RW, Reitz, ML. 1985. Effects of room reverberation and noise on speech discrimination by the elderly. *Audiology* 24, 319–324.

Hawkins, JE, Stevens, SS. 1950. The masking of pure tones and of speech by white noise. *J Acoust Soc Am* 22, 6–13.

Heinz, JM, Stevens, N. 1961. On the properties of voiceless fricative consonants. *J Acoust Soc Am* 33, 589–596.

Helfer, KS. 1992. Aging and the binaural advantage in reverberation and noise. *J Speech Hear Res* 35, 1394–1401.

Helfer, KS. 1994. Binaural cues and consonant perception in reverberation and noise. *J Speech Hear Res* 37, 429–438.

Helfer, KS. 1997. Auditory and auditory-visual perception of clear and conversational speech. *J Speech Lang Hear Res* 40, 432–443.

Hillenbrand, J, Getty, LA, Clark, MJ, Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J Acoust Soc Am* 97, 3099–3111.

Hirsh, IJ, Davis, H, Silverman, SR, Reynolds, ER, Eldert, E, Benson, RW. 1952. Development of materials for speech audiometry. *J Speech Hear Dis* 17, 321–337.

House, AS. 1957. Analog studies of nasal consonants. *J Acoust Soc Am* 22, 190–204.

House, AS, Fairbanks, G. 1953. The influence of consonant environment upon the secondary characteristics of vowels. *J Acoust Soc Am* 25, 105–113.

House, A, Williams, C, Hecker, H, Kryter, A. 1965. Articulation-testing methods: Consonantal differentiations with a close-response set. *J Acoust Soc Am* 37,158–166.

Hughes, GW, Halle, M. 1956. Spectral properties of fricative consonants. *J Acoust Soc Am* 28, 303–310.

Humes, LE, Dirks, DD, Bell, TS, Ahlstrom, C, Kincaid, GE. 1986. Application of the Articulation index and the speech transmission index to the recognition of speech by normal hearing and hearing-impaired listeners. *J Speech Hear Res* 29, 447–462.

International Electrotechnical Commission (IEC). 1987. *Publ. 268: Sound system equipment part 16: Report on the RASTI method for objective rating of speech intelligibility in auditoria.*

Jassem, W. 1965. The formants of fricative consonants. *Lang Speech* 8, 1–16.

Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In: K Johnson, JW Mullennix (eds.), Talker Variability in Speech Processing. San Diego, CA: Academic, 145–165.

Jongman, A. 1985. Duration of fricative noise as a perceptual cue to place and manner of articulation in English fricatives. *J Acoust Soc Am* 77(Suppl 1), S26.

Josse, G, Mazoyer, B, Crivello, F, Tzourio-Mazoyer, N. 2003. Left planum temporale: An anatomical marker of left hemispheric specialization for language comprehension. *Brain Res Cogn Brain Res* 18, 1–14.

Jusczyk, PW, Luce, PA. 2002. Speech perception and spoken word recognition: Past and present. *Ear Hear* 23, 2–40.

Kain, A, Amano-Kusumoto, A, Hosom, JP. 2008. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *J Acoust Soc Am* 124, 2308–2319.

Kamm, CA, Dirks, DD, Bell, TS. 1985. Speech recognition and the articulation index for normal and hearing impaired listeners. *J Acoust Soc Am* 77, 281–288.

Kent, RD, Read, C. 2002. *The Acoustic Analysis of Speech*, 2nd ed. New York, NY: Delmar.

Kewley-Port, D. 1983. Time-varying features as correlates of place of articulation in stop consonants. *J Acoust Soc Am* 73, 322–335.

Kewley-Port, D, Luce, PA. 1984. Time-varying features of stop consonants in auditory running spectra: A first report. *Percept Psychophys* 35, 353–360.

Kimura, D. 1961. Cerebral dominance and the perception of verbal stimuli. *Can J Psychol* 15, 166–171.

Kimura, D. 1964. Left-right differences in the perception of melodies. *Q J Exp Psychol* 16, 355–358.

Kimura, D. 1967. Functional asymmetry of the brain in dichotic listening. *Cortex* 3, 163–178.

Kirk, KI, Pisoni, DB, Osberger, MJ. 1995. Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear Hear* 16, 470–481.

Knudsen, VO. 1929. The hearing of speech in auditoriums. *J Acoust Soc Am* 1, 56–82.

Koenig, W, Dunn, HK, Lacy, LY. 1946. The sound spectrograph. *J Acoust Soc Am* 17, 19–49.

Krause, JC, Braida, LD. 2002. Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *J Acoust Soc Am* 112, 2165–2172.

Krause, JC, Braida, LD. 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *J Acoust Soc Am* 115, 362–378.

Kryter, KD. 1962a. Methods for the calculation and use of the articulation index. *J Acoust Soc Am* 34, 1689–1697.

Kryter, KD. 1962b. Validation of the articulation index. *J Acoust Soc Am* 34, 1698–1702.

Kryter, KD. 1985. *The Effects of Noise on Man*, 2nd ed. New York, NY: Academic Press.

Kuhl, PK. 1991. Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Percept Psychophys* 50, 93–107.

Kuhl, PK, Miller, JD. 1975. Speech perception by the chinchilla: Voice–voiceless distinctions in alveolar plosive consonants. *Science* 190, 69–72.

Kuhl, PK, Miller, JD. 1978. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *J Acoust Soc Am* 63, 905–917.

Kuhl, PK, Padden, DM. 1982. Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Percept Psychophys* 32, 542–550.

Kuhl, PK, Padden, DM. 1983. Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J Acoust Soc Am* 73, 1003–1010.

Kuhl, PK, Williams, KA, Lacerda, F, Stevens, KN, Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606–608.

Kurowski, K, Blumstein, SE. 1987a. Acoustic properties for place of articulation in nasal consonants. *J Acoust Soc Am* 81, 1917–1927.

Kurowski, KM, Blumstein, SE. 1987b. In: MK Huffman, RA Krakow (eds.), *Nasals, Nasalization, and the Velum*. New York, NY: Academic, 197–222.

Kurowski, K, Blumstein, SE. 1993. Acoustic properties for the perception of nasal consonants. In: MK Huffman, R Krakow (eds.), *The Feature Nasal: Phonetic Bases and Phonological Implications*. Academic Press, 197–222.

Ladefoged, P, Broadbent, DE. 1957. Information conveyed by vowels. *J Acoust Soc Am* 29, 98–104.

Lazarus, H. 1987. Prediction of verbal communication in noise: A development of generalized SIL curves and the quality of communication, Part 2. *Appl Acoust* 20, 245–261.

Liberman, AM. 1996. *Speech: A Special Code*. Cambridge, MA: MIT Press.

Liberman, AM, Cooper, FS, Shankweiler, DB, Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychol Rev* 74, 431–461.

Liberman, AM, Delattre, PC, Cooper, FS, Gerstman, LJ. 1954. The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psychol Monogr* 68, 1–13.

Liberman, AM, Delattre, PC, Gerstman, LJ, Cooper, FS. 1956. Tempo of frequency as a cue for distinguishing classes of speech sounds. *J Exp Psychol* 52, 127–137.

Liberman, AM, Harris, KS, Hoffman, HS, Griffith, BC. 1957. The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol* 54, 358–368.

Liberman, AM, Harris, KS, Kinney, KA, Lane, HL. 1961. The discrimination of relative onset of certain speech and nonspeech patterns. *J Exp Psychol* 61, 379–388.

Liberman, AM, Mattingly, IG. 1985. The motor theory of speech perception revised. *Cognition* 21, 1–36.

Liberman, AM, Mattingly, IG. 1989. A specialization for speech perception. *Science* 243, 489–494.

Liberman, AM, Whalen, DH. 2000. On the relation of speech to language. *Trends Cogn Sci* 4, 187–196.

Licklider, JCR. 1946. Effects of amplitude distortion upon the intelligibility of speech. *J Acoust Soc Am* 18, 429–434.

Licklider, JCR, Bindra, D, Pollack, I. 1948. The intelligibility of rectangular speech waves. *Am J Psychol* 61, 1–20.

Licklider, JCR, Pollack, I. 1948. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *J Acoust Soc Am* 20, 42–51.

Lieberman, P. 1973. On the evaluation of language: A unified view. *Cognition* 2, 59–94.

Lindblom, B, Studdert-Kennedy, M. 1967. On the role of formant transitions in vowel recognition. *J Acoust Soc Am* 42, 830–843.

Lisker, L. 1957a. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language* 33, 42–49.

Lisker, L. 1957b. Minimal cues for separating /w,r,l,y/ in intervocalic position. *Word* 13, 256–267.

Lisker, L, Abramson, AS. 1964. Cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20, 384–422.

Lisker, L, Abramson, AS. 1967. Some effects of context on voice onset time in English stops. *Lang Speech* 10, 1–28.

Lisker, L, Abramson, AS. 1970. *The voicing dimension: Some experiments in comparative phonetics*. Proceedings of the Sixth International Congress on Phonetic Science. Prague, Czech Republic: Academia, 563–567.

Liu, S, Zeng, FG. 2006. Temporal properties in clear speech perception. *J Acoust Soc Am* 120, 424–432.

Liu, S, Rio, ED, Bradlow, AR, Zeng, F-G. 2004. Clear speech perception in acoustic and electric hearing. *J Acoust Soc Am* 116, 2374–2383.

Luce, PA. 1990. *Neighborhoods of Words in the Mental Lexicon*. Research on Speech Perception Tech. Report 6. Bloomington, IL: Indiana University.

Luce, PA, Pisoni, DB. 1998. Recognizing spoken words: The neighborhood activation model. *Ear Hear* 19, 1–36

Luce, PA, Stephen, DG, Auer, ET Jr, Vitevitch, MS. 2000. Phonetic priming, neighborhood activation, and PARSYN. *Percept Psychophys* 62, 615–625.

MacDonald, J, McGurk, H. 1978. Visual influences on speech perception processes. *Percept Psychophys* 24, 253–257.

Mäkelä, AM, Alku, P, May, PJ, Mäkinen, V, Tiitinen, H. 2005. Left-hemispheric brain activity reflects formant transitions in speech sounds. *Neuroreport* 16, 549–553.

Mäkelä, AM, Alku, P, Tiitinen, H. 2003. The auditory N1m reveals the left-hemispheric representation of vowel identity in humans. *Neurosci Lett* 353, 111–114.

Malecot, A. 1956. Acoustic cues for nasal consonants. *Language* 32, 274–284.

Marslen-Wilson, WD, Tyler, LK. 1980. The temporal structure of spoken language understanding. *Cognition* 8, 1–71.

Marslen-Wilson, WD, Welsh, A. 1978. Processing interactions during word-recognition in continuous speech. *Cogn Psychol* 10, 29–63.

Massaro, DW. 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, DW. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

Massaro, DW, Chen, TH. 2008. The motor theory of speech perception revisited. *Psychon Bull Rev* 15, 453–457.

Mattingly, IG, Liberman, AM. 1988. Specialized perceiving systems for speech and other biologically significant sounds. In: GMG Edelman, WE Gall, WM Cowen (eds.), *Auditory Function: Neurological Bases of Hearing*. New York, NY: Wiley, 775–793.

McClelland, JL, Elman, JL. 1986. The TRACE model of speech perception. *Cogn Psychol* 18, 1–86.

McGurk, H, MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746–747.

Miller, GA. 1947. The masking of speech. *Psychol Bull* 44, 105–129.

Miller, GA. 1951. *Language and Communication*. New York, NY: McGraw-Hill.

Miller, JD. 1989. Auditory-perceptual interpretation of the vowel. *J Acoust Soc Am* 85, 2114–2134.

Miller, GA, Licklider, JCR. 1950. The intelligibility of interrupted speech. *J Acoust Soc Am* 22, 167–173.

Miller, GA, Heise, GA, Lichten, W. 1951. The intelligibility of speech as a function of the context of the test materials. *J Ex Psychol* 41, 329–335.

Miller, JL, Kent, RD, Atal, BS. (eds.) 1991. *Papers in Speech Communication: Speech Perception*. New York, NY: Acoustical Society of America.

Miller, GA, Nicely, PA. 1955. An analysis of perceptual confusions among some English consonants. *J Acoust Soc Am* 27, 338–352.

Miller, MI, Sachs, MB. 1983. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *J Acoust Soc Am* 74, 502–517.

Miller, JD, Wier, CC, Pastore, RE, Kelly, WJ, and Dooling, RJ. 1976. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *J Acoust Soc Am* 60, 410–417.

Mirabile, PJ, Porter, RJ. 1975. *Dichotic and monotic interaction between speech and nonspeech sounds at different stimulus-onset-asynchronies*. Paper at 89th Meeting of the Acoustical Society of America, Austin, Texas.

Mitler, JD, Wier, CC, Pastore, R, Kelly, WJ, Dooling, RJ. 1976. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *J Acoust Soc Am* 60, 410–417.

Moon, S, Lindblom, B. 1994. Interaction between duration, context, and speaking style in English stressed vowels. *J Acoust Soc Am* 96, 40–55.

Morton, J. 1969. Interaction of information in word recognition. *Psychol Rev* 76, 165–178.

Möttönen, R, Krause, CM, Tiippana, K, Sams, M. 2002. Processing of changes in visual speech in the human auditory cortex. *Brain Res Cogn Brain Res* 13, 417–425.

Näätänen, R. 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology* 38, 1–21.

Nabelek, AK. 1976. Reverberation effects for normal and hearing-impaired listeners. In: SK Hirsh, DH Eldredge, IJ Hirsh, SR Silverman (eds.), *Hearing and Davis: Essays Honoring Hallowell Davis*. St. Louis, MO: Washington University Press, 333–341.

Nabelek, AK, Dagenais, PA. 1986. Vowel errors in noise and in reverberation by hearing impaired listeners. *J Acoust Soc Am* 80, 741–748.

Nabelek, AK, Letowski, TR. 1985. Vowel confusions of hearing-impaired listeners under reverberant and nonreverberant conditions. *J Speech Hear Dis* 50,126–131.

Nabelek, AK, Letowski, TR, Tucker, FM. 1989. Reverberant overlap- and self-masking in consonant identification. *J Acoust Soc Am* 86, 1259–1265.

Nabelek, AK, Mason, D. 1981. Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. *J Speech Hear Res* 24, 375–383.

Nabelek, AK, Nabelek, IV. 1994. Room acoustics and speech perception. In: J Katz (ed.), *Handbook of Clinical Audiology*, 4th ed. Baltimore, MD: Williams & Wilkins, 624–637.

Nabelek, AK, Pickett, JM. 1974. Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation and hearing aids. *J Acoust Soc Am* 56, 628–639.

Nabelek, AK, Robinette, LN. 1978. Reverberation as a parameter in clinical testing. *Audiology* 17, 239–259.

Nearey, TM. 1989. Static, dynamic, and relational properties in vowel perception. *J Acoust Soc Am* 2088–2113.

Nelson, DA, Marler, P. 1989. Categorical perception of a natural stimulus continuum: Birdsong. *Science* 244, 976–978.

Nelson, PB, Nittrouer, S, Norton, SJ. 1995. "Say-stay" identification and psychoacoustic performance of hearing-impaired listeners. *J Acoust Soc Am* 97, 1830–1838.

Nittrouer, S,. 1992 Age-related difference in perceptual effects of formant transitions with syllables and across phoneme boundaries. *J Phonet* 20, 351–382.

Nittrouer, S, Studdert-Kennedy, M. 1987. The role of coarticulatory effects in the perception of fricatives by children and adults. *J Acoust Soc Am* 86, 1266–1276.

Norris, DG. 1994. SHORTLIST: A connectionist model of continuous speech recognition. *Cognition* 52, 189–234.

O'Connor, JD, Gerstman, LJ, Liberman, AM, Delattre, PC, and Cooper, FS. 1957. Acoustic cues for the perception of initial /w,j,r,l/ in English. *Word* 13, 24–43.

Ogden, GC, Massaro, DW. 1978. Integration of featural information in speech perception. *Psychol Rev* 85, 172–191.

Ohde, RN, Haley, KL, Barnes, CW. 2006. Perception of the [m]-[n] distinction in consonant–vowel (CV) and vowel–consonant (VC) syllables produced by child and adult talkers. *J Acoust Soc Am* 119, 1697–1711.

Ohde, RN, Haley, KL, Vorperian, HK, McMahon, C. 1995. A developmental study of the perception of onset spectra for stop consonants in different vowel environments. *J Acoust Soc Am* 97, 3800–3812.

Orfield, SJ 1987. The RASTI method of testing relative intelligibility. *Sound Vibr* 21(12), 20–22.

Pardo, JS, Remez. 2006. The perception of speech. In: M Traxler, MA Gernsbacher (eds.), *The Handbook of Psycholinguistics*, 2nd ed. New York, NY: Academic, 201–248.

Pavlovic, CV. 1987. Derivation of primary parameters and procedures for use in speech intelligibility predictions. *J Acoust Soc Am* 82, 413–422.

Pavlovic, CV. 1988. Articulation index predictions of speech intelligibility in hearing aid selection. *ASHA* 30(6/7), 63–65.

Pavlovic, CV. 1991. Speech recognition and five articulation indexes. *Hear Inst* 42(9), 20–23.

Pavlovic, CV, Studebaker, GA, Sherbecoe, RL. 1986. An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. *J Acoust Soc Am* 80, 50–57.

Payton, KL, Uchanski, RM, Braida, LD. 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J Acoust Soc Am* 95(3), 1581–1592.

Pearsons, KS, Bennett, RL, Fidell, S. 1977. *Speech levels in various noise environments*. EPA Report 600/1-77-025. Washington, D.C.: Environmental Protection Agency.

Pegg, JE, Werker, JF. 1997. Adult and infant perception of two English phones. *J Acoust Soc Am* 102, 3742–3753.

Peterson, GE. 1952. The information-bearing elements of speech. *J Acoust Soc Am* 24, 629–637.

Peterson, GE, Barney, HL. 1952. Control methods used in a study of the vowels. *J Acoust Soc Am* 24, 175–184.

Peterson, GE, Lehiste, I. 1962. Revised CNC lists for auditory tests. *J Speech Hear Dis* 27, 62–70.

Phillips, DP. 1999. Auditory gap detection, perceptual channels, and temporal resolution in speech perception. *J Am Acad Audiol* 10, 343–354.

Phillips, DP, Taylor, T, Hass, SE, Carr, MM, Mossop, JE. 1997. Detection of silent intervals between noises activating different perceptual channels: Some properties of "central" gap detection. *J Acoust Soc Am* 101, 3694–3705.

Picheny, MA, Durlach, NI, Braida, LD. 1985. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *J Speech Hear Res* 28, 96–103.

Picheny, MA, Durlach, NI, Braida, LD. 1986. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *J Speech Hear Res* 29, 434–446.

Picheny, MA, Durlach, NI, Braida, LD. 1989. Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility of clear and conversational speech. *J Speech Hear Res* 32, 600–603.

Pickett, JM. 1999. *The Acoustics of Speech Communication.* Boston, MA: Allyn and Bacon.

Pisoni, DB. 1977. Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *J Acoust Soc Am* 61, 1352–1361.

Pisoni, DB, McNabb, SD. 1974. Dichotic interactions and phonetic feature processing. *Brain Lang* 1, 351–362.

Porter, RJ. 1975. Effect of delayed channel on the perception of dichotically presented speech and nonspeech sounds. *J Acoust Soc Am* 58, 884–892.

Potter, RK, Kopp, GA, Green, HC. 1947. *Visible Speech.* New York, NY: Van Nostrand.

Powers, GL, Speaks, C. 1973. Intelligibility of temporally interrupted speech. *J Acoust Soc Am* 54, 661–667.

Price, C, Thierry, G, Griffiths, T. 2005. Speech-specific auditory processing: Where is it? *Trends Cogn Sci* 9, 271–276.

Raphael, LJ. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. *J Acoust Soc Am* 51, 1296–1303.

Raphael, LJ, Borden, GJ, Harris, KS. 2007. *Speech Science Primer*, 5th ed. Baltimore, MD: Lippincott Williams & Wilkins.

Remez, RE. 2005. Perceptual organization of speech. In: DB Pisoni, RE Remez (eds.), *The Handbook of Speech Perception.* Oxford, MA: Blackwell, 28–50.

Remez, RE, Rubin, PE, Berns, SM, Pardo, JS, Lang, JM. 1994. On the perceptual organization of speech. *Psychol Rev* 101, 129–156.

Remez, RE, Rubin, PE, Pisoni, DB, Carrell, TD. 1981. Speech perception without traditional speech cues. *Science* 212, 947–950.

Repp, BR. 1984. Categorical perception: Issues, methods, findings. In: N Lass (ed.), *Speech and Language: Advances in Basic Research and Practice*, Vol. 10. New York, NY: Academic Press, 243–335.

Repp, BR, Lin, HB. 1989. Acoustic properties and perception of stop consonant release transients. *J Acoust Soc Am* 85, 379–396.

Rhebergen, KS, Versfeld, NJ. 2005. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J Acoust Soc Am* 117, 2181–2192.

Rosenzweig, MR, Postman, L. 1957. Intelligibility as a function of frequency of usage. *J Exp Psychol* 54, 412–422.

Ryalls, J. 1996. *A Basic Introduction to Speech Perception.* San Diego, CA: Singular.

Sacia, CE, Beck, CJ. 1926. The power of fundamental speech sounds. *Bell Sys Tech J* 5, 393–403.

Saint-Amour, D, DeSanctis, P, Molholm, S, Ritter, W, Foxe, JJ. 2007. Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologica* 45, 587–597.

Sams, M, Aulanko, R, Hämäläinen, M, Hari, R, Lounasmaa, OV, Lu, ST, Simola, J. 1991. Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Lett* 127, 141–145.

Sawusch, JR, Jusczyk, PW. 1981. Adaptation and contrast in the perception of voicing. *J Exp Psychol Hum Percept Perform* 7, 408–421.

Sawusch, JR, Pisoni, D. 1974. On the identification of place and voicing features in synthetic stop consonants. *J Phonet* 2, 181–194.

Schmidt-Nielsen, A. 1987. Comments on the use of physical measures to assess speech intelligibility. *J Acoust Soc Am* 81, 1985–1987.

Schouten, MEH, vanHessen, AJ. 1992. Modeling phoneme perception. I: Categorical perception. *J Acoust Soc Am* 92, 1841–1855.

Schum, DJ. 1996. Intelligibility of clear and conversational speech of young and elderly talkers. *J Am Acad Audiol* 7, 212–218.

Searle, CL, Jacobson, JZ, Rayment, SG. 1979. Stop consonant discrimination based on human audition. *J Acoust Soc Am* 79, 799–809.

Shankweiler, D, Studdert-Kennedy, M. 1967. Identification of consonants and vowels presented to the left and right ears. *Q J Exp Psychol* 19, 59–63.

Sharf, DJ. 1962. Duration of post-stress intervocalic stops preceding vowels. *Lang Speech* 5, 26–30.

Shtyrov, Y, Pihko, E, Pulvermüller, F. 2005. Determinants of dominance: Is language laterality explained by physical or linguistic features of speech? *Neuroimage* 27, 37–47.

Slis, IH. 1970. Articulatory measurements on voiced, voiceless and nasal consonants. *Phonetica* 24, 193–210.

Steeneken, HJM. 2006. *Speech Transmission Index (STI): Objective Speech Intelligibility Assessment*. Available at http://www.steeneken.nl/sti.html (accessed April 29, 2008).

Steeneken, HJM, Houtgast, T. 1980. A physical method for measuring speech-transmission quality. *J Acoust Soc Am* 67, 318–326.

Stevens, KN. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In: EE David, PB Denes (eds.), *Human Communication: A Unified View*. New York, NY: McGraw-Hill, 51–66.[check the surname]

Stevens, KN, Blumstein, SE. 1978. Invariant cues for place of articulation in stop consonants. *J Acoust Soc Am* 64, 1358–1368.

Stevens, KN, Halle, M. 1967. Remarks on analysis by synthesis and distinctive features. In: W Wathem-Dunn (ed.), *Models for the Perception of Speech and Visual Form*. Cambridge, MA: MIT Press, 88–102.

Stevens, KN, House, AS. 1955. Development of a quantitative description of vowel articulation. *J Acoust Soc Am* 27, 484–493.

Stevens, KN, House, AS. 1961. An acoustical theory of vowel productions and some of its implications. *J Speech Hear Res* 4, 302–320.

Stevens, KN, House, AS. 1963. Perturbation of vowel articulations by consonantal context: An acoustical study. *J Speech Hear Res* 6, 111–128.

Stevens, KN, Klatt, DH. 1974. The role of formant transitions in the voiced–voiceless distinctions for stops. *J Acoust Soc Am* 55, 653–659.

Stevens, SS, Miller, J, Truscott, I. 1946. The masking of speech by sine waves, square waves, and regular and modulated pulses. *J Acoust Soc Am* 18, 418–424.

Strange, W. 1989a. Dynamic specification of coarticulated vowels spoken in sentence context. *J Acoust Soc Am* 2135–2153.

Strange, W. 1989b. Evolving theories of vowel perception. *J Acoust Soc Am* 85, 2081–2087.

Strevens, P. 1960. Spectra of fricative noise in human speech. *Lang Speech* 3, 32–49.

Studdert-Kennedy, M, Shankweiler, D. 1970. Hemispheric specialization for speech perception. *J Acoust Soc Am* 48, 579–594.

Studdert-Kennedy, M, Shankweiler, D, Schulman, S. 1970. Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. *J Acoust Soc Am* 48, 599–602.

Takata, Y, Nabelek, AK. 1990. English consonant recognition in noise and in reverberation by Japanese and American listeners. *J Acoust Soc Am* 88, 663–666.

Tartter, VC, Eimas, PD. 1975. The role of auditory and phonetic feature detectors in the perception of speech. *Percept Psychophys* 18, 293–298.

Tervaniemi, M, Hugdahl, K. 2003. Lateralization of auditory-cortex functions. *Brain Res Brain Res Rev* 43, 231–246.

Tiffany, WR. 1953. Vowel recognition as a function of duration, frequency modulation and phonetic context. *J Speech Hear Disord* 18, 289–301.

Uchanski, RM, Choi, SS, Braida, LD, Reed, CM, Durlach, NI. 1996. Speaking clearly for the hard of hearing IV: Further studies of speaking rate. *J Speech Hear Res* 39, 494–509.

van Wijngaarden, SJ, Drullman, R. 2008. Binaural intelligibility prediction based on the speech transmission index. *J Acoust Soc Am* 123, 4514–4523.

Walley, AC, Carrell, TD. 1983. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *J Acoust Soc Am* 73, 1011–1022.

Webster, JC. 1978. Speech interference aspects of noise. In: DM Lipscomb (ed.), *Noise and Audiology*. Baltimore, MD: University of Park Press, 193–228.

Werker, JF, Gilbert, JHV, Humphrey, K, Tees, RC. 1981. Developmental aspects of cross language speech perception. *Child Dev* 52, 349–53.

Werker, JF, Tees, RC. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav Dev* 7, 49–63.

Whalen, DH, Liberman, AM. 1987. Speech perception takes precedence over nonspeech perception. *Science* 237, 169–171.

Whalen, DH, Benson, R, Richardson, M, Swainson, B, Clark, VP, Lai, S, Mencl, WE, Fulbright, RK, Constable, RT, Liberman, AM. 2006. Differentiation of speech and nonspeech processing within primary auditory cortex. *J Acoust Soc Am* 119, 575–581

Wood, CC. 1975. Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. *J Exp Psychol* 104, 133.

Wood, CC, Goff, WR, Day, RS. 1971. Auditory evoked potentials during speech perception. *Science* 173, 1248–1251.