5
The Perception of Speech Under Adverse Conditions

PETER ASSMANN and QUENTIN SUMMERFIELD

1. Introduction

Speech is the primary vehicle of human social interaction. In everyday life, speech communication occurs under an enormous range of different environmental conditions. The demands placed on the process of speech communication are great, but nonetheless it is generally successful. Powerful selection pressures have operated to maximize its effectiveness.

The adaptability of speech is illustrated most clearly in its resistance to distortion. In transit from speaker to listener, speech signals are often altered by background noise and other interfering signals, such as reverberation, as well as by imperfections of the frequency or temporal response of the communication channel. Adaptations for robust speech transmission include adjustments in articulation to offset the deleterious effects of noise and interference (Lombard 1911; Lane and Tranel 1971); efficient acoustic-phonetic coupling, which allows evidence of linguistic units to be conveyed in parallel (Hockett 1955; Liberman et al. 1967; Greenberg 1996; see Diehl and Lindblom, Chapter 3); and specializations of auditory perception and selective attention (Darwin and Carlyon 1995).

Speech is a highly efficient and robust medium for conveying information under adverse conditions because it combines strategic forms of redundancy to minimize the loss of information. Coker and Umeda (1974, p. 349) define redundancy as "any characteristic of the language that forces spoken messages to have, on average, more basic elements per message, or more cues per basic element, than the barest minimum [necessary for conveying the linguistic message]." This definition does not address the function of redundancy in speech communication, however. Coker and Umeda note that "redundancy can be used effectively; or it can be squandered on uneven repetition of certain data, leaving other crucial items very vulnerable to noise. . . . But more likely, if a redundancy is a property of a language and has to be learned, then it has a purpose." Coker and Umeda conclude that the purpose of redundancy in speech communication is to provide a basis for error correction and resistance to noise.

We shall review evidence suggesting that redundancy contributes to the perception of speech under adverse acoustic conditions in several different ways:

1. by limiting perceptual confusion due to errors in speech production;
2. by helping to bridge gaps in the signal created by interfering noise, reverberation, and distortions of the communication channel; and
3. by compensating for momentary lapses in attention and misperceptions on the part of the listener.

Redundancy is present at several levels in speech communication—acoustic, phonetic, and linguistic. At the acoustic level it is exemplified by the high degree of covariation in the pattern of amplitude modulation across frequency and over time. At the phonetic level it is illustrated by the many-to-one mapping of acoustic cues onto phonetic contrasts and by the presence of cue-trading relationships (Klatt 1989). At the level of phonology and syntax it is illustrated by the combinatorial rules that organize sound sequences into words, and words into sentences. Redundancy is also provided by semantic and pragmatic context.

This chapter discusses the ways in which acoustic, phonetic, and lexical redundancy contribute to the perception of speech under adverse conditions. By "adverse conditions" we refer to any perturbation of the communication process resulting from either an error in production by the speaker, channel distortion or masking in transmission, or a distortion in the auditory system of the listener. Section 2 considers the design features of speech that make it well suited for transmission in the presence of noise and distortion. The primary aim of this section is to identify perceptually salient properties of speech that underlie its robustness. Section 3 reviews the literature on the intelligibility of speech under adverse listening conditions. These include background noise of various types (periodic/random, broadband/narrowband, continuous/fluctuating, speech/nonspeech), reverberation, changes in the frequency response of the communication channel, distortions resulting from pathology of the peripheral auditory system, and combinations of the above. Section 4 considers strategies used by listeners to maintain, preserve, or enhance the intelligibility of speech under adverse acoustic conditions.

2. Design Features of Speech that Contribute to Robustness

We begin with a consideration of the acoustic properties of speech that make it well suited for transmission in adverse environments.

2.1 The Spectrum

The traditional starting point for studying speech perception under adverse conditions is the long-term average speech spectrum (LTASS) (Dunn and White 1940; French and Steinberg 1947; Licklider and Miller 1951; Fletcher 1953; Kryter 1985). A primary objective of these studies has been to characterize the effects of noise, filtering, and channel distortion on the LTASS in order to predict their impact on intelligibility. The short-term amplitude spectrum (computed over a time window of 10 to 30 ms) reveals the acoustic cues for individual vowels and consonants combined with the effects of distortion. The long-term spectrum tends to average out segmental variations. Hence, a comparison of the LTASS obtained under adverse conditions with the LTASS obtained in quiet can provide a clearer picture of the effects of distortion.

Figure 5.1 (upper panel) shows the LTASS obtained from a large sample of native speakers of 12 different languages reading a short passage from a story (Byrne et al. 1994). The spectra were obtained by computing the root mean square (rms) level in a set of one-third-octave-band filters over 125-ms segments of a 64-second recorded passage spoken in a "normal" speaking style. There are three important features of the LTASS. First, there is a 25-dB range of variation in average level across frequency, with the bulk of energy below 1 kHz, corresponding to the frequency region encompassing the first formant. Second, there is a gradual decline in spectrum level for frequencies above 0.5 kHz. Third, there is a clear distinction between males and females in the low-frequency region of the spectrum. This difference is attributable to the lower average fundamental frequency (f_0) of male voices. As a result, the first harmonic of a male voice contributes appreciable energy between 100 and 150 Hz, while the first harmonic of a female voice makes a contribution between 200 and 300 Hz.

The lower panel of Figure 5.1 shows the LTASS obtained using a similar analysis method from a sample of 15 American English vowels and diphthongs. After averaging, the overall spectrum level was adjusted to match that of the upper panel at 250 Hz in order to facilitate comparisons between panels. Compared to continuous speech, the LTASS of vowels shows a more pronounced local maximum in the region of f_0 (close to 100 Hz for males and 200 Hz for females). However, in other respects the pattern is similar, suggesting that the LTASS is dominated by the vocalic portions of the speech signal. Vowels and other voiced sounds occupy about half of the time waveform of connected speech, but dominate the LTASS because such segments contain greater power than the adjacent aperiodic segments.

The dashed line in each panel illustrates the variation in absolute sensitivity as a function of frequency for young adult listeners with normal hearing (Moore and Glasberg 1987). Comparison of the absolute threshold function with the LTASS shows that the decline in energy toward lower frequencies is matched by a corresponding decline in sensitivity. However, the

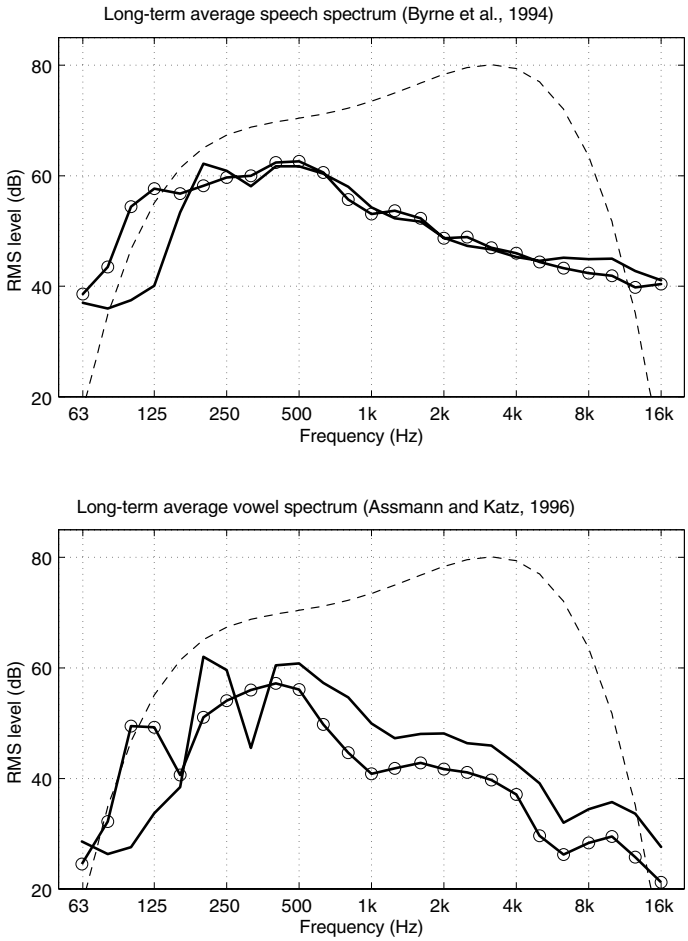

FIGURE 5.1. The upper panel shows the long-term average speech spectrum (LTASS) for a 64-second segment of recorded speech from 10 adult males and 10 adult females for 12 different languages (Byrne et al. 1994). The vertical scale is expressed in dB SPL (linear weighting). The lower panel shows the LTASS for 15 vowels and diphthongs of American English (Assmann and Katz 1998). Filled circles in each panel show the LTASS for adult males; unfilled circles show the LTASS for adult females. To facilitate comparisons, these functions were shifted along the vertical scale to match those obtained with continuous speech in the upper panel. The dashed line in each panel indicates the shape of the absolute threshold function for listeners with normal hearing (Moore and Glasberg 1987). The absolute threshold function is expressed on an arbitrary dB scale, with larger values indicating greater sensitivity.

speech spectrum has a shallower roll-off in the region above 4 kHz than the absolute sensitivity function and the majority of energy in the speech spectrum encompasses frequencies substantially lower than the peak in pure-tone sensitivity. This low-frequency emphasis may be advantageous for the transmission of speech under adverse conditions for several reasons:

1. The lowest three formants of speech, F_1 to F_3, generally lie below 3 kHz. The frequencies of the higher formants do not vary as much, and contribute much less to intelligibility (Fant 1960).
2. Phase locking in the auditory nerve and brain stem preserves the temporal structure of the speech signal in the frequency range up to about 1500 Hz (Palmer 1995). Greenberg (1995) has suggested that the low-frequency emphasis in speech may be linked to the greater reliability of information coding at low frequencies via phase locking.
3. To separate speech from background sounds, listeners rely on cues, such as a common periodicity and a common pattern of interaural timing (Summerfield and Culling 1995), that are preserved in the patterns of neural discharge only at low frequencies (Cariani and Delgutte 1996a,b; Joris and Yin 1995).
4. Auditory frequency selectivity is sharpest (on a linear frequency scale) at low frequencies and declines with increasing frequency (Patterson and Moore 1986).

The decline in auditory frequency selectivity with increasing frequency has several implications for speech intelligibility. First, auditory filters have larger bandwidths at higher frequencies, which means that high-frequency filters pass a wider range of frequencies than their low-frequency counterparts. Second, the low-frequency slope of auditory filters becomes shallower with increasing level. As a consequence, low-frequency maskers are more effective than high-frequency maskers, leading to an "upward spread of masking" (Wegel and Lane 1924; Trees and Turner 1986; Dubno and Ahlstrom 1995). In their studies of filtered speech, French and Steinberg (1947) observed that the lower speech frequencies were the last to be masked as the signal-to-noise ratio (SNR) was decreased.

Figure 5.2 illustrates the effects of auditory filtering on a segment of the vowel [I] extracted from the word "hid" spoken by an adult female talker. The upper left panel shows the conventional Fourier spectrum of the vowel in quiet, while the upper right panel shows the spectrum of the same vowel embedded in pink noise at an SNR of +6 dB. The lower panels show the "auditory spectra" or "excitation patterns" of the same two sounds. An excitation pattern is an estimate of the distribution of auditory excitation across frequency in the peripheral auditory system generated by a specific signal. The excitation patterns shown here were obtained by plotting the rms output of a set of gammatone filters[1] as a function of filter center frequency.

[1] The gammatone is a bandpass filter with an impulse response composed of two terms, one derived from the gamma function, and the other from a cosine function

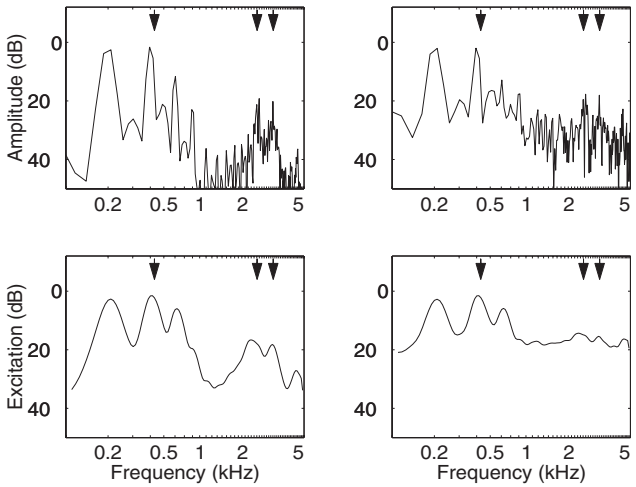

FIGURE 5.2. The upper left panel shows the Fourier amplitude spectrum of a 102.4-ms segment of the vowel [I] spoken by an adult female speaker of American English. The upper right panel shows the same segment embedded in pink noise at a signal-to-noise ratio (SNR) of +6 dB. Below each amplitude spectrum is its auditory excitation pattern (Moore and Glasberg 1983, 1987) simulated using a gammatone filter analysis (Patterson et al. 1986). Fourier spectra and excitation patterns are displayed on a log frequency scale. Arrows show the frequencies of the three lowest formants (F_1–F_3) of the vowel.

The three lowest harmonics are "resolved" as distinct peaks in the excitation pattern, while the upper harmonics are not individually resolved. In this example, the first formant (F_1) lies close to the second harmonic but does not coincide with it. In general, F_1 in voiced segments is not represented by a distinct peak in the excitation pattern and hence its frequency must be inferred, in all likelihood from the relative levels of prominent harmonics in this appropriate region (Klatt 1982; Darwin 1984; Assmann and Nearey 1986). The upper formants (F_2–F_4) give rise to distinct peaks in the excitation pattern when the vowel is presented in quiet. The addition of noise leads to a greater spread of excitation at high frequencies, and the spectral contrast (peak-to-valley ratio) of the upper formants is reduced.

The simulation in Figure 5.2 is based on data from listeners with normal hearing whose audiometric thresholds fall within normal limits and who

or "tone" (Patterson et al. 1992). The bandwidths of these filters increase with increasing center frequency, in accordance with estimates of psychophysical measures of auditory frequency selectivity (Moore and Glasberg 1983, 1987). Gammatone filters have been used to model aspects of auditory frequency selectivity as measured psychophysically (Moore and Glasberg 1983, 1987; Patterson et al. 1992) and physiologically (Carney and Yin 1988), and can be used to simulate the effects of auditory filtering on speech signals.

possess normal frequency selectivity. Sensorineural hearing impairments lead to elevated thresholds and are often associated with a reduction of auditory frequency selectivity. Psychoacoustic measurements of auditory filtering in hearing-impaired listeners often show reduced frequency selectivity compared to normal listeners (Glasberg and Moore 1986), and consequently these listeners may have difficulty resolving spectral features that could facilitate making phonetic distinctions among similar sounds. The reduction in spectral contrast can be simulated by broadening the bandwidths of the filters used to generate excitation patterns, such as those shown in Figure 5.2 (Moore 1995). Support for the idea that impaired frequency selectivity can result in poorer preservation of vocalic formant structure and lower identification accuracy comes from studies of vowel masking patterns (Van Tasell et al. 1987a; Turner and Henn 1989). In these studies, forward masking patterns were obtained by measuring the threshold of a brief sinusoidal probe at different frequencies in the presence of a vocalic masker to obtain an estimate of the "internal representation" of the vowel. Hearing-impaired listeners generally exhibit less accurate representations of the signal's formant peaks in their masking patterns than do normal-hearing listeners.

Many studies have shown that the intelligibility of masked, filtered, or distorted speech depends primarily on the proportion of the speech spectrum available to the listener. This principle forms the basis for the articulation index (AI), a model developed by Fletcher and his colleagues at Bell Laboratories in the 1920s to predict the effects of noise, filtering, and communication channel distortion on speech intelligibility (Fletcher 1953). Several variants of the AI have been proposed over the years (French and Steinberg 1947; Kryter 1962; ANSI S3.5 1969, 1997; Müsch and Buus 2001a,b).

The AI is an index between 0 and 1 that describes the effectiveness of a speech communication channel. An "articulation-to-intelligibility" transfer function can be applied to convert this index to predicted intelligibility in terms of percent correct. The AI model divides the speech spectrum into a set of up to 20 discrete frequency bands, taking into account the absolute threshold, the masked threshold imposed by the noise or distortion, and the long-term average spectrum of the speech. The AI has two key assumptions:

1. The contribution of any individual channel is independent of the contribution of other bands.
2. The contribution of a channel depends on the SNR within that band.

The predicted intelligibility depends on the proportion of time the speech signal exceeds the threshold of audibility (or the masked threshold, in conditions where noise is present) in each band. The AI is expressed by the following equation (Pavlovic and Studebaker 1984):

1

2

$$AI = P\int_{0}^{\infty} If)W(f)df \tag{1}$$

The term $I(f)$ is the importance function, which reflects the significance of different frequency bands to intelligibility. $W(f)$ is the audibility or weighting function, which describes the proportion of information associated with $I(f)$ available to the listener in the testing environment. The term P is the proficiency factor and depends on the clarity of the speaker's articulation and the experience of the listener (including such factors as the familiarity of the speaker's voice and dialect). Computation of the AI typically begins by dividing the speech spectrum into a set of n discrete frequency bands (Pavlovic 1987):

$$AI = P\sum_{i=1}^{n} I_i W_i \tag{2}$$

The AI computational procedure developed by French and Steinberg (1947) uses 20 frequency bands between 0.15 and 8 kHz, with the width of each band adjusted to make the bands equal in importance. These adjustments were made on the basis of intelligibility tests with low-pass and high-pass filtered speech, which revealed a maximum contribution from the frequency region around 2.5 kHz. Later methods have employed one-third octave bands (e.g., ANSI 1969) or critical bands (e.g., Pavlovic 1987) with nonuniform weights.[2]

The audibility term, W_i, estimates the proportion of the speech spectrum exceeding the masked threshold in the ith frequency band. The ANSI S3.5 model assumes that speech intelligibility is determined over a dynamic range of 30 dB, with the upper limit determined by the "speech peaks" (the sound pressure level exceeded 1% of the time by the speech energy integrated over 125-ms intervals—on average, about 12 dB above the mean level). The lower limit (representing the speech "valleys") is assumed to lie 18 dB below the mean level. The AI assumes a value of 1.0 under conditions of maximum intelligibility (i.e., when the 30-dB speech range exceeds the absolute threshold, as well as the masked threshold, if noise is present in every frequency band). If any part of the speech range lies below the threshold across frequency channels, or is masked by noise, the AI is reduced by the percentage of the area covered. The AI assumes a value of 0 when the speech is completely masked, or is below threshold, and hence

[2] Several studies have found that the shape of the importance function varies as a function of speaker, gender and type of speech material (e.g., nonsense CVCs versus continuous speech), and the procedure used (French and Steinberg 1947; Beranek 1947; Kryter 1962; Studebaker et al. 1987). Recent work (Studebaker and Sherbecoe 2002) suggests that the 30-dB dynamic range assumed in standard implementations may be insufficient, and that the relative importance assigned to different intensities within the speech dynamic range varies as a function of frequency.

unintelligible. As a final step, the value of the AI can be used to predict intelligibility with the aid of an empirically derived articulation-to-intelligibility transfer function (Pavlovic and Studebaker 1984). The shape of the transfer function differs for different speech materials and testing conditions (Kryter 1962; Studebaker et al. 1987).

The AI generates accurate predictions of average speech intelligibility over a wide range of conditions, including high- and low-pass filtering (French and Steinberg 1947; Fletcher and Galt 1950), different types of broadband noise (Egan and Wiener 1946; Miller 1947), bandpass-filtered noise maskers (Miller et al. 1951), and various distortions of the communication channel (Beranek 1947). It has also been used to model binaural masking level differences for speech (Levitt and Rabiner 1967) and loss of speech intelligibility resulting from sensorineural hearing impairments (Fletcher 1952; Humes et al. 1986; Pavlovic et al. 1986; Ludvigsen 1987; Rankovic 1995, 1998). The success of the AI model is consistent with the idea that speech intelligibility under adverse conditions is strongly affected by the audibility of the speech spectrum.[3] However, the AI was designed to accommodate linear distortions and additive noises with continuous spectra. It is less effective for predicting the effects of nonlinear or time-varying distortions, transmission channels with sharp peaks and valleys, masking noises with line spectra, and time-domain distortions, such as those created by echoes and reverberation. Some of these difficulties are overcome by a reformulation of AI theory—the speech transmission index—described below.

2.2 Formant Peaks

The vocal tract resonances (or "formants") provide both phonetic information (signaling the identity of the intended vowel or consonant) and source information (signaling the identity of the speaker). The frequencies of the lowest three formants, as well as their pattern of change over time, provide cues that help listeners ascertain the phonetic identities of vowels and consonants. Vocalic contrasts, in particular, are determined primarily by differences in the formant pattern (e.g., Peterson and Barney 1952; Nearey 1989; Hillenbrand et al. 1995; Hillenbrand and Nearey 1999; Assmann and Katz, 2000; see Diehl and Lindblom, Chapter 3).

[3] The AI generates a single number that can be used to predict the overall or average intelligibility of specified speech materials for a given communication channel. It does not predict the identification of individual segments, syllables, or words, nor does it predict the pattern of listeners' errors. Calculations are typically based on speech spectra accumulated over successive 125-ms time windows. A shorter time window and a short-time running spectral analysis (Kates 1987) would be required to predict the identification of individual vowels and consonants (and the confusion errors made by listeners) in tasks of phonetic perception.

The formant representation provides a compact description of the speech spectrum. Given an initial set of assumptions about the glottal source and a specification of the damping within the supralaryngeal vocal tract (in order to determine the formant bandwidths), the spectrum envelope can be predicted from a knowledge of the formant frequencies (Fant 1960). A change in formant frequency leads to correlated changes throughout the spectrum, yet listeners attend primarily to the spectral peaks in order to distinguish among different vocalic qualities (Carlson et al. 1979; Darwin 1984; Assmann and Nearey 1986; Sommers and Kewley-Port 1996).

One reason why spectral peaks are important is that spectral detail in the region of the formant peaks is more likely to be preserved in background noise. The strategy of attending primarily to spectral peaks is robust not only to the addition of noise, but also to changes in the frequency response of a communication channel and to some deterioration of the frequency resolving power of the listener (Klatt 1982; Assmann and Summerfield 1989; Roberts and Moore 1990, 1991; Darwin 1984, 1992; Hukin and Darwin 1995). In comparison, a whole-spectrum matching strategy that assigns equal weight to the level of the spectrum at all frequencies (Bladon 1982) or a broad spectral integration strategy (e.g., Chistovich 1984) would tend to incorporate noise into the spectral estimation process and thus be more susceptible to error. For example, a narrow band of noise adjacent to a formant peak could substantially alter the spectral center of gravity without changing the frequency of the peak itself.

While it is generally agreed that vowel quality is determined primarily by the frequencies of the two or three lowest formants (Pols et al. 1969; Rosner and Pickering 1994), there is considerable controversy over the mechanisms underlying the perception of these formants in vowel identification. Theories generally fall into one of two main classes—those that assert that the identity of a vowel is determined by a distributed analysis of the shape of the entire spectrum (e.g., Pols et al. 1969; Bakkum et al. 1993; Zahorian and Jagharghi 1993), and those that assume an intermediate stage in which spectral features in localized frequency regions are extracted (e.g., Chistovich 1984; Carlson et al. 1974). Consistent with the first approach is the finding that listeners rely primarily on the two most prominent harmonics near the first-formant peak in perceptual judgments involving front vowels (e.g., [i] and [ɛ]), which have a large separation of the lowest formants, F_1 and F_2. For example, listeners rely only on the most prominent harmonics in the region of the formant peak to distinguish changes in F_1 center frequency (Sommers and Kewley-Port 1996) as well as to match vowel quality as a function of F_1 frequency (Assmann and Nearey 1986; Dissard and Darwin 2000) and identify vowels along a phonetic continuum (Carlson et al. 1974; Darwin 1984; Assmann and Nearey 1986).

A different pattern of sensitivity is found when listeners judge the phonetic quality of back vowels (e.g., [u] and [o]), where F_1 and F_2 are close together in frequency. In this instance, harmonics remote from the F_1 peak

can make a contribution, and additional aspects of spectral shape (such as the center of spectral gravity in the region of the formant peaks or the relative amplitude of the formants) are taken into account (Chistovich and Lublinskaya 1979; Beddor and Hawkins 1990; Assmann 1991; Fahey et al. 1996).

The presence of competing sounds is a problem for models of formant estimation. Extraneous sounds in the F_1 region might change the apparent amplitudes of resolved harmonics and so alter the phonetic quality of the vowel. Roberts and Moore (1990, 1991a) demonstrated that this effect can occur. They found that additional components in the F_1 region of a vowel as well as narrow bands of noise could alter its phonetic quality. The shift in vowel quality was measured in terms of changes in the phonetic segment boundary along a continuum ranging from [I] to [ɛ] (Darwin 1984). Roberts and Moore hypothesized that the boundary shift was the result of excitation from the additional component being included in the perceptual estimate of the amplitudes of harmonics close to the first formant of the vowel.

How do listeners avoid integrating evidence from other sounds when making vowel quality judgments? Darwin (1984, 1992; Darwin and Carlyon 1995) proposed that the perception of speech is guided by perceptual grouping principles that exclude the contribution of sounds that originate from different sources. For example, Darwin (1984) showed that the influence of a harmonic component on the phoneme boundary was reduced when that harmonic started earlier or later than the remaining harmonics of the vowel. The perceptual exclusion of the asynchronous component is consistent with the operation of a perceptual grouping mechanism that segregates concurrent sounds on the basis of onset or offset synchrony. Roberts and Moore (1991a) extended these results by showing that segregation also occurs with inharmonic components in the region of F_1.

Roberts and Moore (1991b) suggested that the perceptual segregation of components in the F_1 region of vowels might benefit from the operation of a harmonic sieve (Duifhuis et al. 1982). The harmonic sieve is a hypothetical mechanism that excludes components whose frequencies do not correspond to integer multiples of a given fundamental. It accounts for the finding that a component of a tonal complex contributes less to its pitch when its frequency is progressively mistuned from its harmonic frequency (Moore et al. 1985). Analogously, a mistuned component near the F_1 peak makes a smaller contribution to its phonetic quality than that of its harmonic counterparts (Darwin and Gardner 1986).

The harmonic sieve utilizes a "place" analysis to group together components belonging to the same harmonic series, and thereby excludes inharmonic components. This idea has proved to have considerable explanatory power. However, it has not always been found to offer the most accurate account of the perceptual data. For example, computational models based on the harmonic sieve have not generated accurate predictions of listeners' identification of concurrent pairs of vowels with different f_0s (Scheffers

1983; Assmann and Summerfield 1990). The excitation patterns of "double vowels" often contain insufficient evidence of concurrent f_0s to allow for their segregation using a harmonic sieve. Alternative mechanisms, based on a temporal (or place-time) analysis, have been shown to make more accurate predictions of the pattern associated with listeners' identification responses (Assmann and Summerfield 1990; Meddis and Hewitt 1992).

Meddis and Hewitt (1991, 1992) describe a computational model that

1. carries out a frequency analysis of the signal using a bank of bandpass filters,
2. compresses the filtered waveforms using a model of mechanical-to-neural transduction,
3. performs a temporal analysis using autocorrelation functions (ACFs), and
4. sums the ACFs across the frequency channels to derive a summary autocorrelogram.

The patterning of peaks in the summary autocorrelogram is in accord with many of the classic findings of pitch perception (Meddis and Hewitt 1991). The patterning can also yield accurate estimates of the f_0s of concurrent vowels (Assmann and Summerfield 1991). Meddis and Hewitt (1992) segregated pairs of concurrent vowels by combining the ACFs across channels with a common periodicity to provide evidence of the first vowel, and then grouping the remaining ACFs to reconstruct the second segment. They showed that the portion of the summary autocorrelogram with short time lags (<4.5 ms) could be used to predict the phonetic identities of the vowels with reasonable accuracy.

The harmonic sieve and autocorrelogram embody different solutions to the problem of segregating a vowel from interfering sounds (including a second competing vowel). It can be complicated to compare models of vowel identification that incorporate these mechanisms because the models may differ not only in the technique used to represent the spectrum (or temporal pattern) of a vowel, but also in the approach to classifying the spectrum. Most models of categorization assume that the pattern to be classified is compared with a set of templates, and that the pattern is characterized as belonging to the set defined by the template to which it is most similar. "Similarity" is usually measured by an implicit representation of perceptual distance. The choice of distance metric can have a substantial effect on the accuracy with which a model predicts the pattern of vocalic identification made by a listener. Table 5.1 summarizes the results of several studies that have evaluated the efficacy of such perceptual distance metrics for vowels.

Three conclusions emerge from these comparisons. First, no single metric is optimal across all conditions. Different metrics appear to be best suited for different tasks. Second, metrics that highlight spectral peaks [and possibly also spectral "shoulders" (Assmann and Summerfield 1989; Lea and

TABLE 5.1. Sample of perceptual distance metrics for vowels

Factor	Pattern	Reference
Vowel similarity judgments	Dimensions derived from principal components analysis (PCA) of one-third-octave spectra	Pols et al. (1967)
Speaker quality judgments for normal and profoundly hearing-impaired talkers	One-third-octave spectra + PCA	Bakkum et al. (1993)
Talker normalization	Excitation patterns	Suomi (1984)
Vowel quality matching	Loudness density patterns	Bladon and Lindblom (1981)
Prediction of vowel systems	Loudness density patterns	Lindblom (1986)
Vowel similarity judgments	Weighted spectral slope metric	Carlson et al. (1979); Nocerino et al. (1985)
Vowel identification by hearing impaired listeners	Weighted spectral slope metric	Turner and Henn (1994)
Concurrent vowel identification	Negative second differential of excitation pattern; peak metric	Assmann and Summerfield (1989)
Discrimination of vowel formant frequencies (F_1 and F_2)	Peak-weighted excitation pattern, specific loudness difference	Sommers and Kewley-Port (1996); Kewley-Port and Zheng (1998)

Summerfield 1994)] perform best when the task is to phonetically identify vowels. Third, metrics that convey information about the entire shape of the spectrum are more appropriate when the task is to discriminate vowels acoustically, that is, on the basis of timbre rather than using differences in phonetic quality (Klatt 1982).

The fact that no single metric is optimal for all vowel tasks and that the sensitivity of perceptual distance metrics to distortion and noise is so highly variable suggests that a simple template-matching approach with fixed frequency weights is inappropriate for vowel perception. Similar conclusions have been reached in recent reviews of speech-recognition research (Gong 1994; Lippmann 1996a; see Morgan et al., Chapter 6). To a much greater extent than humans, most existing speech recognizers are adversely affected by transmission-channel distortion, noise, and reverberation. A major difficulty is that these types of distortion can obscure or mask weak formants and other aspects of spectral shape, resulting in the problem of "missing data" (Cooke et al. 1996; Cooke and Ellis 2001). They can introduce "spurious" peaks and alter the shape of the spectrum, resulting in greater than predicted perceptual distances. Adult listeners with normal hearing possess remarkable abilities to compensate for such distortions. Unlike machine-based speech recognizers, they do so without the need for explicit practice or "recalibration" (Watkins 1991; Dijkhuisen et al. 1987; Buunen et al. 1996; Lippmann 1996a).

The effects of two different types of noise on the spectrum of a vowel is illustrated in Figure 5.3. Panel A shows the Fourier amplitude spectrum of a 102.4-ms segment of the vowel in the word "head," spoken by an adult female speaker in a sound-attenuated recording chamber. Panel B shows the same signal combined with white noise at an SNR of 0 dB. The envelope of the spectrum [obtained by linear predictive coding (LPC) analysis (Markel and Gray 1976)] shows that the spectral contrast is greatly diminished and that the peaks generated by the higher formants (F_2, F_3, F_4) are no longer distinct. The harmonicity of the vowel is not discernible in the upper formants, but remains evident in the F_1 region.

In natural listening environments steady-state broadband noise with a flat spectrum is uncommon. A more common form of noise is created when several individuals talk at once, creating multispeaker babble. Panel C shows the amplitude spectrum of a 102.4-ms segment of such babble,

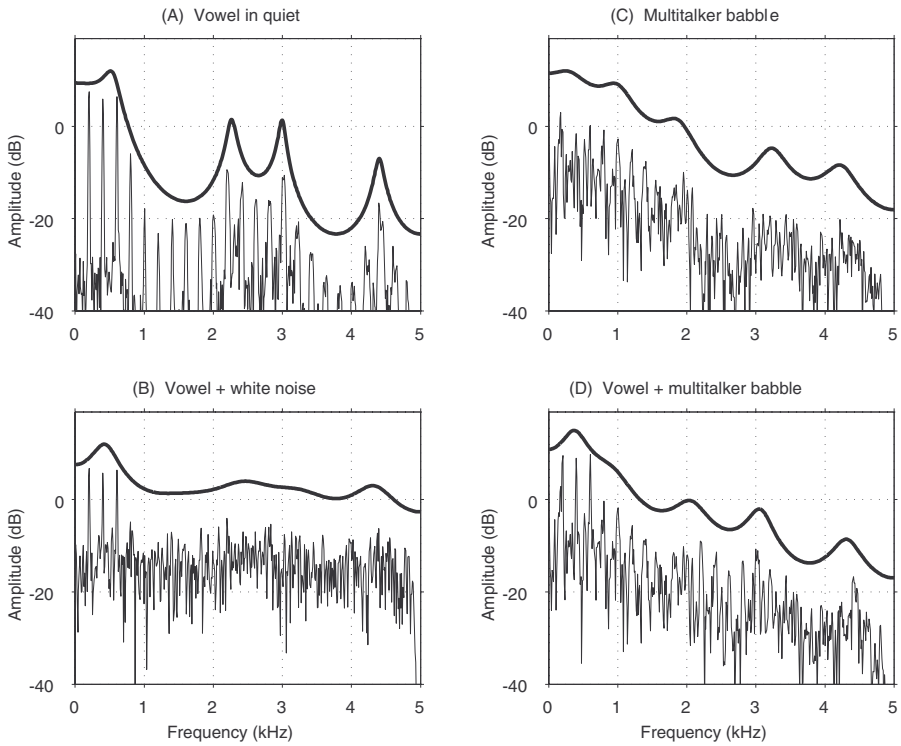

FIGURE 5.3. Effects of noise on formant peaks. A: The Fourier amplitude spectrum of a vowel similar to [ɛ]. The solid line shows the spectrum envelope estimated by linear predictive coding (LPC) analysis. B: White noise has been superimposed at an SNR of 0 dB. C: The spectrum of a sample of multitalker babble. D: The spectrum of the vowel mixed with the babble at an SNR of 0 dB.

created by mixing speech from four different speakers (two adult males, one adult female, and a child) at comparable intensities. In panel D the speech babble is combined with the vowel shown in panel A at an SNR of 0 dB. Compared with panel A, there is a reduction in the degree of spectral contrast and there are changes in the shape of the spectrum. There are additional spectral peaks introduced by the competing voices, and there are small shifts in the frequency locations of spectral peaks that correspond to formants of the vowel. The harmonicity of the vowel is maintained in the low-frequency region, and is preserved to some degree in the second and third formant regions. These examples indicate that noise can distort the shape of the spectrum, change its slope, and reduce the contrast between peaks and adjacent valleys. However, the frequency locations of the formant peaks of the vowel are preserved reasonably accurately in the LPC analysis in panel D, despite the fact that other aspects of spectral shape, such as spectral tilt and the relative amplitudes of the formants, are lost.

Figure 5.3 also illustrates some of the reasons why formant tracking is such a difficult engineering problem, especially in background noise (e.g., Deng and Kheirallah 1993). An example of the practical difficulties of locating particular formants is found in the design of speech processors for cochlear implants.[4] Explicit formant tracking was implemented in the processor developed by Cochlear PTY Ltd. during the 1980s, but was subsequently abandoned in favor of an approach that seeks only to locate spectral peaks without assigning them explicitly to a specific formant. The latter strategy yields improved speech intelligibility, particularly in noise (McKay et al. 1994; Skinner et al. 1994).

Listeners with normal hearing have little difficulty understanding speech in broadband noise at SNRs of 0 dB or greater. Environmental noise typically exhibits a sloping spectrum, more like the multispeaker babble of panels C and D than the white noise of panel B. For such noises, a subset of formants (F_1, F_2, and F_3) is often resolved, even at an SNR of 0 dB, and generates distinct peaks in the spectrum envelope. However, spectral contrast (the difference in dB between the peaks and their adjacent valleys) is reduced by the presence of noise in the valleys between formants. As a result, finer frequency selectivity is required to locate the peaks. Listeners with sensorineural hearing loss generally have difficulty understanding speech under such conditions. Their difficulties are likely to stem, at least in part, from reduced frequency selectivity (Simpson et al. 1990; Baer et al. 1993). This hypothesis has been tested by the application of digital signal processing techniques to natural speech designed to either (1) reduce the

[4] Cochlear implants provide a useful means of conveying auditory sensation to the profoundly hearing impaired by bypassing the malfunctioning parts of the peripheral auditory system and stimulating auditory-nerve fibers directly with electrical signals through an array of electrodes implanted within the cochlea (cf. Clark, Chapter 8).

spectral contrast by smearing the spectral envelope (Keurs et al. 1992, 1993a,b; Baer and Moore 1994) or (2) enhance the contrast by sharpening the formant peaks (Veen and Houtgast 1985; Simpson et al. 1990; Baer et al. 1993). Spectral smearing results in a degradation of speech intelligibility, particularly for vowels, as well as an elevation in the speech reception threshold (SRT) in noise (Plomp and Mimpen 1979). However, the magnitude of the reduction in spectral contrast is not closely linked to measures of frequency selectivity (Keurs 1993a,b). Conversely, attempts to enhance intelligibility by increasing spectral contrast have shown a modest improvement for listeners with cochlear hearing impairment [corresponding to an increase in SNR of up to about 4 dB (Baer et al. 1993)]. These results are consistent with the hypothesis that the difficulties experienced by the hearing-impaired when listening to speech in noise are at least partially due to the reduced ability to resolve formant peaks (cf. Edwards, Chapter 7).

2.3 Periodicity of Voiced Speech

The regularity with which the vocal folds open and close during voicing is one of the most distinctive attributes of speech—its periodicity (in the time domain) and corresponding harmonicity (in the frequency domain). This pattern of glottal pulsing produces periodicity in the waveform at rates between about 70 and 500 Hz. Such vocal fold vibrations are responsible for the perception of voice pitch and provide the basis for segmental distinctions between voiced and unvoiced sounds (such as [b] and [p]), as well as distinctions of lexical tone in many languages. At the suprasegmental level, voice pitch plays a primary role in conveying different patterns of intonation and prosody.

Evidence of voicing is broadly distributed across frequency and time, and is therefore a robust property of speech. Figure 5.4 illustrates the effects of background noise on the periodicity of speech. The left panel shows the waveforms generated by a set of gammatone filters in response to the syllable [ga] in quiet. In this example, the speaker closed her vocal tract about 30 ms after the syllable's onset and then released the closure 50 ms later. The frequency channels below 1 kHz are dominated by the fundamental frequency and the auditorily resolved, low-order harmonics. In the higher-frequency channels, filter bandwidths are broader than the frequency separation of the harmonics, and hence several harmonics interact in the passband of the filter to create amplitude modulation (AM) at the period of the fundamental. The presence of energy in the lowest-frequency channel during the stop closure provides evidence that the consonant is voiced rather than voiceless.

The panel on the right shows that periodicity cues are preserved to some extent in background noise at an SNR of +6 dB. The noise has largely obliterated the silent interval created by the stop consonant and has masked the burst. However, there is continued domination of the output of the low-

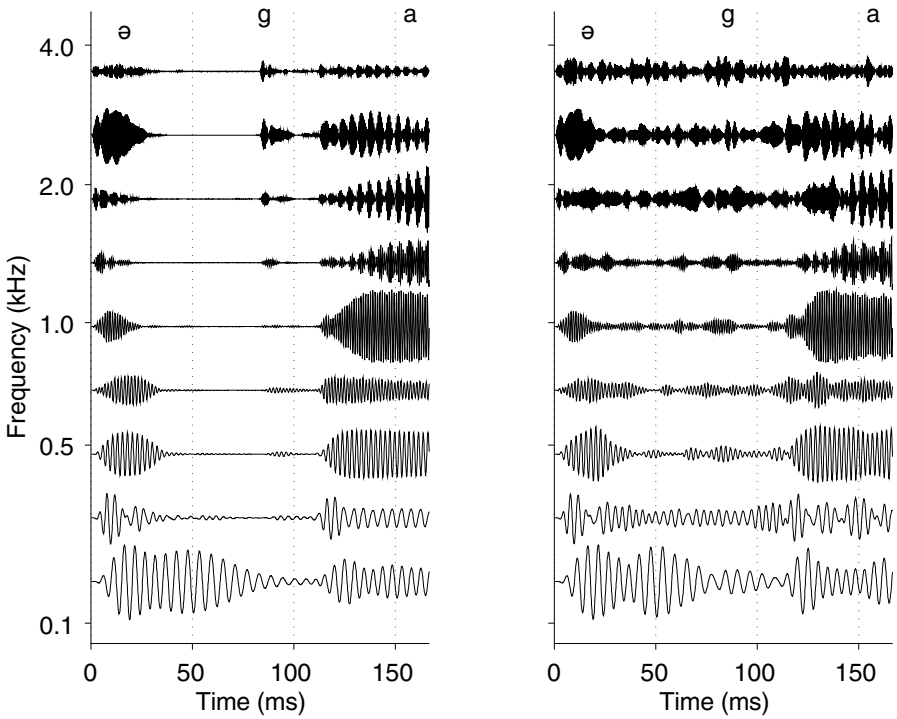

FIGURE 5.4. Effects of background noise on voicing periodicity. The left panel shows the results of a gammatone filter bank analysis (Patterson et al. 1986) of the voiced syllable [ga] spoken by an adult female talker. Filter center frequencies are equally spaced in ERB rate with bandwidths chosen to match auditory filter bandwidths measured psychophysically (Moore and Glasberg 1987) across the 0.1–4.0 kHz range. The panel on the right is an analysis of the same syllable combined with broadband (pink) noise at +6 dB SNR.

frequency channels by the individual harmonics and the modulation at the period of the fundamental remains in several of the higher-frequency channels.

It has been suggested that the presence of voicing underlies speech's robustness to noise. One source of evidence comes from a comparison of voiced and whispered speech. In the latter, the periodic glottal pulses are replaced with aperiodic turbulent noise, which has a continuous, rather than harmonic spectrum. Whispered speech is intelligible under quiet listening situations and is generally reserved for short-range communication, but can be less intelligible than voiced speech under certain conditions (Tartter 1991).

Periodicity cues in voiced speech may contribute to noise robustness via auditory grouping processes (Darwin and Carlyon 1995). A common

periodicity across frequency provides a basis for associating speech components originating from the same larynx and vocal tract (Scheffers 1983; Assmann and Summerfield 1990; Bregman 1990; Darwin 1992; Langner 1992; Meddis and Hewitt 1992). Compatible with this idea, Brokx and
7 Nooteboom (1982), Bird and Darwin (1998), and Assmann (1999) have shown that synthesized target sentences are easier to understand in the presence of a continuous speech masker if targets and maskers are synthesized with different f_0s, than with the same f_0. Similarly, when pairs of synthesized vowels are presented concurrently, listeners are able to identify them more accurately if they are synthesized with different fundamental frequencies, compared to the case where

1. both have the same fundamental (Scheffers 1983; Chalikia and Bregman 1989; Assmann and Summerfield 1990),
2. one is voiced and the other is noise-excited (Scheffers 1983), or
3. both are noise-excited (Scheffers 1983; Lea 1992).[5]

A further source of evidence for a contribution of voicing periodicity to speech intelligibility comes from studies of sine-wave speech (Remez et al. 1981). Sine-wave speech uses frequency-modulated sinusoids to model the movements of F_1, F_2, and F_3 from a natural speech signal, and thus lacks harmonic structure. Despite this spectral reduction, it can be understood, to a certain extent, under ideal listening conditions, though not in background noise. Carrell and Opie (1992), however, have shown that sine-wave speech is easier to understand when it is amplitude modulated at a rate similar to that imposed by the vocal folds during voicing. Thus, common, coherent AM may help listeners to group the three sinusoidal formant together to distinguish them from background noise.

2.4 Rapid Spectral Changes

Stevens (1980, 1983) has emphasized that consonants are differentiated from vowels and other vocalic segments (glides, liquids) by their rate of change in the short-time spectrum. The gestures accompanying consonantal closure and release result in rapid spectral changes (associated with bursts and formant transitions) serving as landmarks or pointers to regions of the signal where acoustic evidence for place, manner, and voicing are concentrated (Liu 1996). Stevens proposed that the information density in speech is highest during periods when the vocal tract produces this sort of rapid opening or closing gestures associated with consonants.

---

[5] However, if one vowel is voiced and the other is noise-excited, listeners can identify the noise-excited (or even an inharmonic) vowel at lower SNRs than its voiced counterpart (Lea 1992). Similar results are obtained using inharmonic vowels whose frequency components are randomly displaced in frequency (Cheveigné et al. 1995). These findings suggest that harmonicity or periodicity may provide a basis for "subtracting" interfering sounds, rather than selecting or enhancing target signals.

Stop consonants are less robust than vowels in noise and more vulnerable to distortion. Compared to vowels, they are brief in duration and low in intensity, making them particularly susceptible to masking by noise (e.g., Miller and Nicely 1955), temporal smearing via reverberation (e.g., Gelfand and Silman 1979), and attenuation and masking in hearing impairment (e.g., Walden et al. 1981). Given their high susceptibility to distortion, it is surprising that consonant segments contribute more to overall intelligibility than vowels, particularly in view of the fact that the latter are more intense, longer in duration, and less susceptible to masking. In natural environments, however, there are several adaptations that serve to offset, or at least partially alleviate, these problems. One is a form of auditory enhancement resulting from peripheral or central adaptation, which increases the prominence of spectral components with sudden onsets (e.g., Delgutte 1980, 1996; Summerfield et al. 1984, 1987; Summerfield and Assmann 1987; Watkins 1988; Darwin et al. 1989). A second factor is the contribution of lipreading, that is, the ability to use visually apparent articulatory gestures to supplement and/or complement the information provided by the acoustic signal (Summerfield 1983, 1987; Grant et al. 1991, 1994). Many speech gestures associated with rapid spectral changes provide visual cues that make an important contribution to intelligibility when the SNR is low.

2.5 Temporal Envelope Modulations

Although the majority of speech perception studies have focused on acoustic cues identified in the short-time Fourier spectrum, an alternative (and informative) way to describe speech is in terms of temporal modulations of spectral amplitude (Plomp 1983; Haggard 1985). The speech waveform is considered as the sum of amplitude-modulated signals contained within a set of narrow frequency channels distributed across the spectrum. The output of each channel is described as a carrier signal that specifies the waveform fine structure and a modulating signal that specifies its temporal envelope. The carrier signals span the audible frequency range between about 0.5 and 8 kHz, while the modulating signals represent fluctuations in the speech signal that occur at slower rates between 5 and 50 events per second—too low to evoke a distinctive sensation of pitch (Hartmann 1997) though they convey vital information for segmental and suprasegmental distinctions in speech.

Rosen (1992) summarized these ideas by proposing that the temporal structure of speech could be partitioned into three distinct levels based on their dominant fluctuation rates:

1. Envelope cues correspond to the slow modulations (at rates below 50 Hz) that are associated with changes in syllabic and phonetic-segment constituents.

2. Periodicity cues, at rates between about 70 and 500 Hz, are created by the opening and closing of the vocal folds during voiced speech.
3. Fine-structure cues correspond to the rapid modulations (above 250 Hz) that convey information about the formant pattern.

Envelope cues contribute to segmental (phonetic) distinctions that rely on temporal patterning (such as voicing and manner of articulation in consonants), as well as suprasegmental information for stress assignment, syllabification, word onsets and offsets, speaking rate, and prosody. Periodicity cues are responsible for the perception of voice pitch, and fine-structure cues are responsible for the perception of phonetic quality (or timbre).

One advantage of analyzing speech in this way is that the reduction in intelligibility caused by distortions such as additive, broadband noise, and reverberation can be modeled in terms of the corresponding reduction in temporal envelope modulations (Houtgast and Steeneken 1985). The capacity of a communication channel to transmit modulations in the energy envelope of speech is referred to as the temporal modulation transfer function (TMTF), which tends to follow a low-pass characteristic, with greatest sensitivity to modulations below about 20 Hz (Viemeister 1979; Festen and Plomp 1981).

Because the frequency components in speech are constantly changing, the modulation pattern of the *broadband* speech signal underestimates the information carried by spectrotemporal changes. Steeneken and Houtgast (1980) estimated that 20 bands are required to adequately represent variation in the formant pattern over time. They obtained the modulation (temporal envelope) spectrum of speech by

1. filtering the speech waveform into octave bands whose center frequencies range between 0.25 and 8 kHz;
2. squaring and low-pass-filtering the output (30-Hz cutoff); and
3. analyzing the resulting intensity envelope with a set of one-third octave, bandpass filters with center frequencies ranging between 0.63 and 12.5 Hz.

The output in each filter was divided by the long-term average of the intensity envelope and multiplied by $\sqrt{2}$ to obtain the modulation index. The modulation spectrum (modulation index as a function of modulation frequency) showed a peak around 3 to 4 Hz, reflecting the variational frequency of individual syllables in speech, as well as a gradual decline in magnitude at higher frequencies.

The modulation spectrum is sensitive to the effects of noise, filtering, nonlinear distortion (such as peak clipping), as well as time-domain distortions (such as those introduced by reverberation) imposed on the speech signal (Houtgast and Steeneken 1973, 1985; Steeneken and Houtgast 2002). Reverberation tends to attenuate the rapid modulations of speech by filling in the less-intense portions of the waveform. It has a low-pass filtering effect on the

TMTF.[6] Noise, on the other hand, attenuates all modulation frequencies to approximately the same degree. Houtgast and Steeneken showed that the extent to which modulations are preserved by a communication channel can be expressed by the TMTF and summarized using a numerical index of transmission fidelity, the speech transmission index (STI). 9

The STI measures the overall reduction in modulations present in the intensity envelope of speech and is obtained by a band-weighting method similar to that used in computing the AI. The input is either a test signal (sinusoidal intensity-modulated noise) or any complex modulated signal such as speech. The degree to which modulations are preserved by the communication channel is determined by analyzing the signal with 7 one-octave band filters whose center frequencies range between 0.125 and 8 kHz. Within each band, the modulation index is computed for 14 modulation frequencies between 0.63 and 12.5 Hz. Each index is transformed into an SNR, truncated to a 30-dB range, and averaged across the 14 modulation frequencies. Next, the octave bands are combined into a single measure, the STI, using band weightings in a manner similar to that used in computing the AI. The STI assumes a value of 1.0 when all modulations are preserved and 0 when they are observed no longer.

Houtgast and Steeneken showed that the reduction in intelligibility caused by reverberation, noise, and other distortions could be predicted accurately by the reduction of the TMTF expressed as the STI. As a result, the technique has been applied to characterizing the intelligibility of a wide range of communication channels ranging from telephone systems to individual seating positions in auditoria. The STI accounts for the effects of nonlinear signal processing in a way that makes it a useful alternative to the AI (which works best for linear distortions and normal hearing listeners). However, both methods operate on the long-term average properties of speech, and therefore do not account for effects of channel distortion on individual speech sounds or predict the pattern of confusion errors.

Figure 5.5 shows the analysis of a short declarative sentence, "The watchdog gave a warning growl." The waveform is shown at the top. The four traces below show the amplitude envelopes (left panels) and modulation spectra (right panels) in four frequency channels centered at 0.5, 1, 2, and 4 kHz. Distinctions in segmental and syllable structure are revealed by the modulation patterns in different frequency bands. For example, the affricate [ç] generates a peak in the amplitude envelope of the 2- and 4-kHz channels, but not in the lower channels. The sentence contains eight syllables, with an average duration of about 200 ms, but only five give rise to distinct peaks in the amplitude envelope in the 1-kHz channel.

[6] In addition to suppressing modulations at low frequencies (less than 4 Hz), room reverberation may introduce spurious energy into the modulation spectrum at frequencies above 16 Hz (Lundlin 1982) as a result of harmonics and formants rapidly crossing the room resonances (Haggard 1985). 45

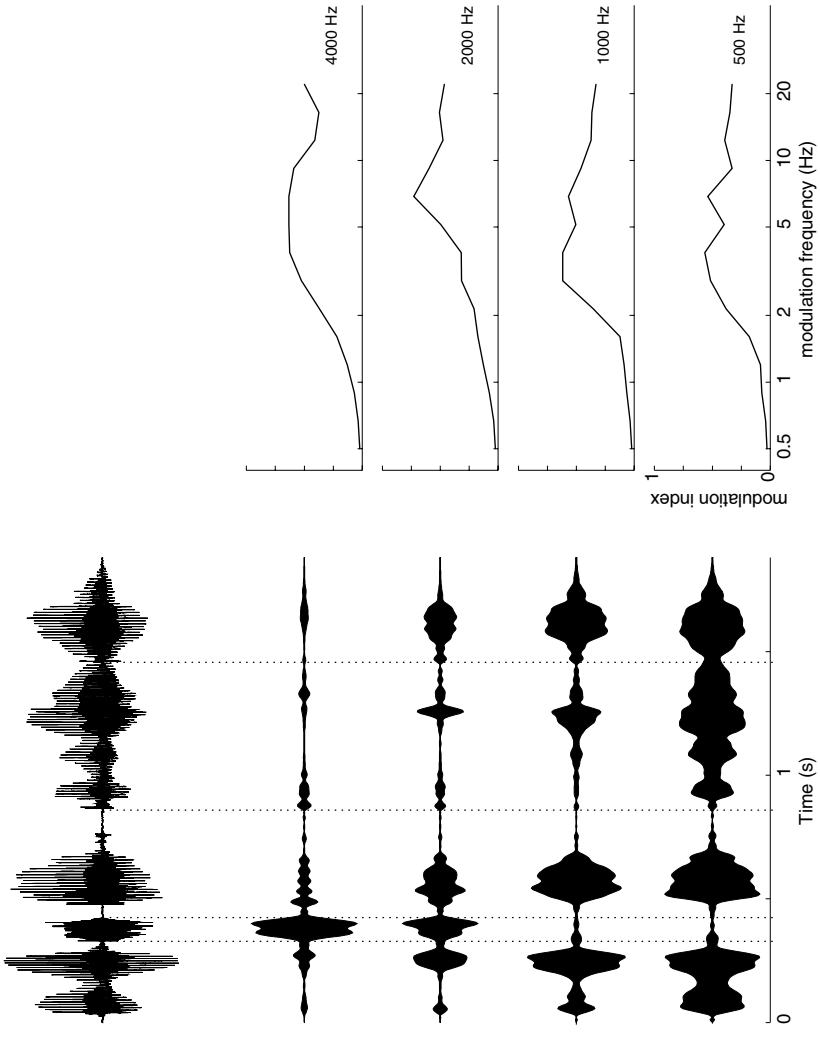

FIGURE 5.5. The upper trace shows the waveform of the sentence, "The watchdog gave a warning growl," spoken by an adult male. The lower traces on the left show the amplitude envelopes in four one-octave frequency bands centered at 0.5, 1, 2, and 4 kHz. The envelopes were obtained by (1) bandpass filtering the speech waveform (elliptical filters; one-octave bandwidth, 80 dB/oct slopes), (2) half-wave rectifying the output, and (3) low-pass filtering (elliptical filters; 80 dB/oct slopes, 30-Hz cutoff). On the right are envelope spectra (modulation index as a function of modulation frequency) corresponding to the four filter channels. Envelope spectra were obtained by (1) filtering the waveforms on the left with a set of bandpass filters at modulation frequencies between 0.5 and 22 Hz (one-third-octave bandwidth, 60 dB/oct slopes), and (2) computing the normalized root-mean-square (rms) energy in each filter band.

A powerful demonstration of the perceptual contribution of temporal envelope modulations to the robustness of speech perception was provided by Shannon et al. (1995). They showed that the rich spectral structure of speech recorded in quiet could be replaced by four bands of random noise that retained only the temporal modulations of the signal, eliminating all evidence of voicing and details of spectral shape. Nonetheless, intelligibility was reduced only marginally, both for sentences and for individual vowels and consonants in an [ɑCɑ] context (where C = any consonant). Subsequent studies showed that the precise corner frequencies and degree of overlap among the filter bands had a relatively minor effect on intelligibility (Shannon et al. 1998). Similar results were obtained when the noise carrier signals were replaced by amplitude-modulated sinusoids with fixed frequencies, equal to the center frequency of each filter band (Dorman et al. 1997). These findings illustrate the importance of the temporal modulation structure of speech and draw attention to the high degree of redundancy in the spectral fine-structure cues that have traditionally been regarded as providing essential information for phonetic identification. The results indicate that listeners can achieve high intelligibility scores when speech has been processed to remove much of the spectral fine structure, provided that the temporal envelope structure is preserved in a small number of broad frequency channels.

It is important to note that these results were obtained for materials recorded in quiet. Studies have shown greater susceptibility to noise masking for processed than unprocessed speech (Dorman et al. 1998; Fu et al. 1998). While performance in quiet reaches an asymptote with four or five bands, the decline in intelligibility as a function of decreasing SNR can be offset, to some degree, by increasing the number of spectral bands up to 12 or 16. Informal listening suggests that there is a radical loss of intelligibility when the speech is mixed with competing sounds. The absence of the spectrotemporal detail denies listeners access to cues such as voicing periodicity, which they would otherwise use to separate the sounds produced by different sources.

Although some global spectral shape information is retained in a four-band approximation to the spectrum, the precise locations of the formant peaks are generally not discernible. Reconstruction of the speech spectrum from just four frequency bands can be viewed as an extreme example of smearing the spectrum envelope. In several studies it has been demonstrated that spectral smearing over half an octave or more (thus exceeding the ear's critical bandwidth) results in an elevation of the SRT for sentences in noise (Keurs et al. 1992, 1993a,b; Baer and Moore 1993). These results are consistent with the notion that the spectral fine structure of speech plays a significant role in resisting distortion and noise.

Some investigators have studied the contribution of temporal envelope modulations in speech processed through what amounts to a *single*-channel, wideband version of the processor described by Shannon and colleagues.

These studies were motivated, in part, by the observation that temporal envelope cues are well preserved in the stimulation pattern of some present-day cochlear implants. Signal-correlated noise is created when noise is modulated by the temporal envelope of the wideband speech signal. It is striking that even under conditions where all spectral cues are removed, listeners can still recover some information for speech intelligibility. Grant et al. (1985, 1991) showed that this type of modulated noise could be an effective supplement to lip-reading for hearing-impaired listeners. Van Tasell et al. (1987b) generated signal-correlated noise versions of syllables in [aCa] context and obtained consonant identification scores from listeners with normal hearing. The temporal patterning preserved in the stimuli was sufficient for the determination of voicing, burst, and amplitude cues, although overall identification accuracy was low. Turner et al. (1995) created two-channel, signal-correlated noise by summing two noise bands, one modulated by low-frequency components, the other by high frequencies (the cutoff frequency was 1500 Hz). They found that two bands were more intelligible for normal listeners (40% correct syllable identification) than a single band (25% correct). This result is consistent with the findings of Shannon et al. (1995), who showed a progressive improvement in intelligibility as the number of processed channels increased from one to four (four bands yielding intelligibility comparable to unprocessed natural speech). Turner et al. (1995) reported similar abilities of normal and sensorineural hearing-impaired listeners to exploit the temporal information in the signal, provided that the reduced audibility of the signal for the hearing impaired was adequately compensated for. Taken together, these studies indicate that temporal envelope cues contribute strongly to intelligibility, but their contribution must be combined across a number of distinct frequency channels.

An alternative approach to studying the role of the temporal properties of speech was adopted in a series of studies by Drullman et al. (1994a,b). They filtered the modulation spectrum of speech to ascertain the contribution of different modulation frequencies to speech intelligibility. The speech waveform was processed with a bank of bandpass filters whose center frequencies ranged between 0.1 and 6.4 kHz. The amplitude envelope in each band (obtained by means of the Hilbert transform) was then low-pass filtered with cutoff frequencies between 0 and 64 Hz. The original carrier signal (waveform fine structure in each filter) was modulated by the modified envelope function. All of the processed waveforms were then summed using appropriate gain to reconstruct the wideband speech signal.

Drullman et al. found that low-pass filtering the temporal envelope of speech with cutoff frequencies below 8 Hz led to a substantial reduction in intelligibility. Low-pass filtering with cutoff frequencies above 8 Hz or high-pass filtering below 4 Hz did not lead to substantially altered SRTs for sentences in noise, compared to unprocessed speech. The intermediate range of modulation frequencies (4–16 Hz) made a substantial contribution to speech intelligibility, however. Removing high-frequency modulations in this range resulted in higher SRTs for sentences in noise and increased

errors in phoneme identification, especially for stop consonants. Removing the low-frequency modulations led to poorer consonant identification, but stops (which are characterized by more rapid modulations) were well preserved, compared to other consonant types. Place of articulation was affected more than manner of articulation. Diphthongs were misclassified as monophthongs. Confusions between long and short vowels (in Dutch) were more prevalent when the temporal envelope was high-pass filtered.

The bandwidth of the analyzing filter had little effect on the results, except with filter cutoffs below 4 Hz. Listeners had considerable difficulty understanding speech from which all but the lowest modulation frequencies (0–2 Hz) had been removed. For these stimuli, the effect of temporal smearing was less deleterious when the bandwidths of the filters were larger (one octave rather than one-quarter octave). Drullman et al. interpreted this outcome in terms of a greater reliance on within-channel processes for low modulation rates. At higher modulation rates listeners may rely to a greater extent on across-channel processes. The cutoff was around 4 Hz, close to the mean rate of syllable and word alternation. If the analysis of temporal modulations relies on across-channel coupling, this would lead to the prediction that phase-shifting the carrier bands would disrupt the coupling and also result in lower intelligibility. However, this does not seem to be the case: Greenberg and colleagues (Greenberg 1996; Arai and Greenberg 1998; Greenberg and Arai 1998) reported that temporal desynchronization of frequency bands by up to 120 ms had relatively little effect on intelligibility on connected speech. Instead, the temporal modulation structure appears to be processed independently in different frequency bands (as predicted by the STI). In comparison, spectral transformations that involve frequency shifts (i.e., applying the temporal modulations to bands with different center frequencies) are extremely disruptive (Blesser 1972). One implication of the result is the importance of achieving the correct relationship between frequency and place within the cochlea when tuning multichannel cochlear-implant systems (Dorman et al. 1997; Shannon et al. 1998).

A further aspect of the temporal structure of speech was investigated by Greenberg et al. (1998). They partitioned the spectrum of spoken English sentences into one-third-octave bands and carried out intelligibility tests on these bands, alone and in combination. Even with just three bands, intelligibility remained high (up to 83% of the words were identified correctly). However, performance was severely degraded when these bands were desynchronized by more than 25 ms. Speech is limited to a small number of narrow bands. In contrast, previous findings by Arai and Greenberg (1998), as well as Greenberg and Arai (1998), show that listeners are relatively insensitive to temporal asynchrony when a larger number (19) of one-quarter-octave bands is presented in combination to create an approximation to (temporally desynchronized) full-bandwidth speech. This suggests that listeners rely on across-channel integration of the temporal structure to improve their recognition accuracy. Greenberg et al. suggested that listeners are sensitive to the phase properties of the modulation spectrum of

speech, and that this sensitivity is revealed most clearly when the spectral information in speech is limited to a small number of narrow bands.

When speech is presented in a noisy background, it undergoes a reduction in intelligibility, in part because the noise reduces the modulations in the temporal envelope. However, the decline in intelligibility may also result from distortion of the temporal fine structure and the introduction of spurious envelope modulations (Drullman 1995a,b; Noordhoek and
12 Drullman 1997). A limitation of the TMTF and STI methods is that they do not consider degradations in speech quality resulting from the introduction of spurious modulations absent from the input (Ludvigsen et al. 1990). These modulations can obscure or mask the modulation pattern of speech, and obliterate some of the cues for identification. Drullman's work suggests that the loss of intelligibility is mainly due to noise present in the temporal envelope troughs (envelope minima) rather than at the peaks (envelope maxima). Drullman (1995b) found that removing the noise from the speech peaks (by transmitting only the speech when the amplitude envelope in each band exceeded a threshold) had little effect on intelligibility. In comparison, removing the noise from the troughs (transmitting speech alone when the envelope fell below the threshold) led to a 2-dB elevation of the SRT.

In combination, these studies show that:

1. an analysis of the temporal structure of speech can make a valuable contribution to describing the perception of speech under adverse conditions;
2. the pattern of temporal amplitude modulation within a few frequency bands provides sufficient information for speech perception; and
3. a qualitative description of the extent to which temporal amplitude modulation is lost in a communication channel (but also, in the case of noise and reverberation, augmented by spurious modulations) is an informative way of predicting the loss of intelligibility that occurs when speech passes through that channel.

2.6 Speaker Adaptations Designed to Resist Noise and Distortion

The previous section considered several built-in properties of speech that help shield against interference and distortion. In addition, speakers actively adjust the parameters of their speech to offset reductions in intelligibility due to masking and distortion. In the current section, we consider specific strategies adopted by speakers under adverse conditions to promote successful communication under adverse conditions. These include the so-called Lombard effect, the use of distinct speaking styles such as "clear" and "shouted" speech, styles used to address hearing-impaired listeners and foreigners, as well as speech produced under high cognitive workload.

When speakers are given explicit instructions to "speak as clearly as possible," their speech differs in several respects from normal conversational speech. Clear speech is produced with higher overall amplitude, a higher mean f_0, and longer segmental durations (Picheny et al. 1985, 1986; Payton et al. 1994; Uchanski et al. 1994). Clear speech is more intelligible than conversational speech under a variety of conditions, including noise, reverberation, and hearing, impairment. Clear speech and conversational speech have similar long-term spectra, but differ with respect to their spectrotemporal patterning, which produces different TMTFs (Payton et al. 1994).

In long-distance communication, speakers often raise the overall amplitude of their voice by shouting. Shouted speech is produced with a reduced spectral tilt, higher mean f_0, and longer vocalic durations (Rostolland 1982). Despite its effectiveness in long-range communication, shouted speech is less intelligible than conversational speech at the same SNR (Pickett 1956; Pollack and Pickett 1958; Rostolland 1985).

When speech communication takes place in noisy backgrounds, such as crowded rooms, speakers modify their vocal output in several ways. The most obvious change is an increase in loudness, but there are a number of additional changes. Collectively these changes are referred to as the Lombard reflex (Lombard 1911).

The conditions that result in Lombard speech have been used to investigate the role of auditory feedback in speech production. Ladefoged (1967) used intense noise designed to mask both airborne and bone-conducted sounds from the speaker's own voice. His informal observations suggested that elimination of auditory feedback and its replacement by intense random noise have a disruptive effect, giving rise to inappropriate nasalization, distorted vowel quality, more variable segment durations, a narrower f_0 range, and an increased tendency to use simple falling intonation patterns. Dreher and O'Neill (1957) and Summers et al. (1988) have extended this work to show that speech produced under noisy conditions (flat-spectrum broadband noise) is more intelligible than speech produced in quiet conditions when presented at the same SNR. Thus, at SNRs where auditory feedback is not entirely eliminated, speakers adjust the parameters of their speech so as to preserve its intelligibility.

Table 5.2 summarizes the results of several studies comparing the production of speech in quiet and in noise. These studies have identified changes in a number of speech parameters. Taken together, the adjustments have two major effects: (1) improvement in SNR; and (2) a reduction in the information rate, allowing more time for decoding. Such additional time is needed, in view of demonstrations by Baer et al. (1993) that degradation of SNR leads to increases in the latency with which listeners make decisions about the linguistic content of a speech signal.

Researchers in the field of automatic speech recognition have sought to identify systematic properties of Lombard speech to improve the recogni-

TABLE 5.2. Summary of changes in the acoustic properties of speech produced in background noise (Lombard speech) compared to speech produced in quiet

Change	Reference
Increase in vocal intensity (about 5 dB increase in speech for every 10 dB increase in noise level)	Dreher and O'Neill (1957)
Decrease in speaking rate	Hanley and Steer (1949)
Increase in average f_0	Summers et al. (1988)
Increase in segment durations	Pisoni et al. (1985)
Reduction in spectral tilt (boost in high-frequency components)	Summers et al. (1988)
Increase in F_1 and F_2 frequency (inconsistent across talkers)	Summers et al. (1988); Junqua and Anglade (1990); Young et al. (1993)

53

tion accuracy of recognizers in noisy backgrounds (e.g., Hanson and Applebaum 1990; Gong 1994) given that Lombard speech is more intelligible than speech recorded in quiet (Summers et al. 1988). Lindblom (1990) has provided a qualitative account of the idea that speakers monitor their speech output and systematically adjust its acoustic parameters to maximize the likelihood of successful transmission to the listener. The hypospeech and hyperspeech (H & H) model assumes a close link between speech production and perception. The model maintains that speakers employ a variety of strategies to compensate for the demands created by the environment to ensure that their message will be accurately received and decoded. When the constraints are low (e.g., in quiet conditions), fewer resources are allocated to speech production, with the result that the articulators deviate less from their neutral positions and hypospeech is generated. When the demands are high (e.g., in noisy environments), speech production assumes a higher degree of flexibility, and speakers produce a form of speech known as hyperspeech.

Consistent with the H & H model, Lively et al. (1993) documented several changes in speech production under conditions of high cognitive work load (created by engaging the subject in a simultaneous task of visual information processing). Several changes in the acoustic correlates of speech were observed when the talker's attention was divided in this way, including increased amplitude, decreased spectral tilt, increased speaking rate, and more variable f_0. A small (2–5%) improvement in vowel identification was observed for syllables produced under such conditions. However, there were substantial differences across speakers, indicating that speaker adaptation under adverse conditions are idiosyncratic, and that it may be difficult to provide a quantitative account of their adjustments. Lively et al. did not inform the speakers that the intelligibility of their speech would be measured. The effects of work load might be greater in conditions where speakers are explicitly instructed to engage in conversation with listeners.

2.7 Summary of Design Features

In this section we have proposed that speech communication incorporates several types of shielding to protect the signal from distortion. The acoustic properties of speech suggest coding principles that contribute to noise reduction and compensation for communication channel distortion. These include the following:

1. The short-term amplitude spectrum dominated by low-frequency energy (i.e., lower than the region of maximum sensitivity of human hearing) and characterized by resonant peaks (formants) whose frequencies change gradually and coherently across time;
2. Periodicity in the waveform at rates between 50 and 500 Hz (along with corresponding harmonicity in the frequency domain) due to vocal fold vibration, combined with the slow fluctuations in the repetition rate that are a primary correlate of prosody;
3. Slow variations in waveform amplitude resulting from the alternation of vowels and consonants at a rate of roughly 3–5 Hz;
4. Rapid spectral changes that signal the presence of consonants.

To this list we can add two additional properties:

1. Differences in relative intensity and time of arrival at the two ears of a target voice and interfering sounds, which provide a basis for the spatial segregation of voices;

2. The visual cues provided by lip-reading provide temporal synchronization between the acoustic signal and the visible movements of the articulators (lips, tongue, teeth, and jaw); Cross-modal integration of acoustic and visual information can improve the effective by about 6 dB (MacLeod and Summerfield 1987).

Finally, the studies reviewed in section 2.6 suggest yet a different form of adaptation; under adverse conditions, speakers actively monitor the SNR and adjust the parameters of their speech to offset the effects of noise and distortion, thereby partially compensating for the reduction of intelligibility. The most salient modifications include an increase in overall amplitude and segmental duration, as well as a reduction in spectral tilt.

3. Speech Intelligibility Under Adverse Conditions

3.1 Background Noise

Speech communication nearly always takes place under conditions where some form of background noise is present. Traffic noise, competing voices, and the noise of fans in air conditioners and computers are common forms of interference. Early research on the effects of noise demonstrated that listeners with normal hearing can understand speech in the presence of white

noise even when the SNR is as low as 0 dB (Fletcher 1953). However, under natural conditions the distribution of noise across time and frequency is rarely uniform. Studies of speech perception in noise can be grouped according to the type of noise maskers used. These include tones and narrowband noise, broadband noise, interrupted noise, speech-shaped noise, multispeaker babble, and competing voices. Each type of noise has a somewhat different effect on speech intelligibility, depending on its acoustic form and information content, and therefore each is reviewed separately.

The effects of different types of noise on speech perception have been compared in several ways. The majority of studies conducted in the 1950s and 1960s compared overall identification accuracy in quiet and under several different levels of noise (e.g., Miller et al. 1951). This approach is time-consuming, because it requires separate measurements of intelligibility for different levels of speech and noise. Statistical comparisons of conditions can be problematic if the mean identification level approaches either 0% or 100% correct in any condition. An alternative method, developed by Plomp and colleagues (e.g., Plomp and Mimpen 1979) avoids these difficulties by measuring the SRT. The SRT is a masked identification threshold, defined as the SNR at which a certain percentage (typically 50%) of the syllables, words, or sentences presented can be reliably identified. The degree of interference produced by a particular noise can be expressed in terms of the difference in dB between the SRT in quiet and in noise. Additional studies have compared the effects of different noises by conducting closed-set phonetic identification tasks and analyzing confusion matrices. The focus of this approach is phonetic perception rather than overall intelligibility, and its primary objective is to identify those factors responsible for the pattern of errors observed within and between different phonetic classes (e.g., Miller and Nicely 1955; Wang and Bilger 1973).

3.2 Narrowband Noise and Tones

A primary factor in determining whether a sound will be an effective masker is its frequency content and the extent of spectral overlap between masker and speech signal. In general, low-frequency noise (20–250 Hz) is more pervasive in the environment, propagates more efficiently, and is more
13 disruptive than high-frequency interference (Berglund et al. 1996). At high intensities, noise with frequencies as low as 20 Hz can reduce the intelligibility of speech (Pickett 1957). Speech energy is concentrated between 0.1 and 6 kHz (cf. section 2.1), and noise with spectral components in this region is the most effective masker of speech. Within this spectral range, lower-frequency interference produces more masking than their higher-frequency counterparts (Miller 1947).

When speech is masked by narrowband maskers, such as pure tones and narrowband noise, low frequencies (<500 Hz) are more disruptive than higher frequencies (Stevens et al. 1946). As the sound pressure level

increases, there is a progressive shift toward lower frequencies (300 Hz), presumably as the result of upward spread of masking by low frequencies (Miller 1947). Complex tonal maskers equated for sound pressure level (square and rectangular waves) are more effective maskers than sinusoids of comparable frequency, with little variation in masking effectiveness as a function of f_0 in the low-frequency (80–400 Hz) range. For frequencies above 1 kHz, neither pure tones nor square waves are effective maskers of speech (Stevens et al. 1946).

Licklider and Guttman (1957) varied the number and frequency spacing of sinusoidal components in a complex tonal masker, holding the overall power constant. Maskers, whose spectral energy is distributed across frequency in accordance with the "equal importance function" (proportional to the critical bandwidth), are more effective speech maskers than those with energy uniformly distributed. Masking effectiveness increased as the number of components was increased from 4 to 40, but there was little further change as the number of components increased beyond 40. Even with 256 components, the masking effectiveness of the complex was about 3 dB less than pink noise with the same frequency range and power.

3.3 Broadband Noise

When speech is masked by broadband noise with a uniform spectrum, its intelligibility is a linear function of SNR as long as the sound pressure level of the noise is greater than about 40 dB (Kryter 1946; Hawkins and Stevens 1950). For listeners with normal hearing, speech communication remains unhampered, unless the SNR is less than +6 dB. Performance remains above chance, even when the SNR is as low as –18 dB (Licklider and Miller 1951). The relationship between SNR and speech intelligibility is affected by context (e.g., whether the stimuli are nonsense syllables, isolated words, or words in sentences), by the size of the response set, and by the entropy associated with the speech items to be identified (Miller et al. 1951). In closed-set identification, the larger the response set the greater the susceptibility to noise. In open-set tasks the predictability of words within the sentence is a significant factor. Individual words in low-predictability sentences are more easily masked than those in high-predictability or neutral sentences (Kalikow et al. 1977; Elliot 1995).

Miller and Nicely (1955) examined the effects of broadband (white) noise on the identification of consonants in CV (consonant-vowel) syllables. They classified consonants in terms of such phonetic features as voicing, nasality, affrication, duration, and place of articulation. For each subgroup they examined overall error rates and confusion patterns, as well as a measure of the amount of information transmitted. Their analysis revealed that noise had the greatest effect on place of articulation. Duration and frication were somewhat more resistant to noise masking. Voicing and nasality were transmitted fairly successfully, and preserved to some extent, even at an SNR of

–12 dB. The effects of noise masking were similar to those of low-pass filtering, but did not resemble high-pass filtering, which resulted in a more random pattern of errors. They attributed the similarity in effects of low-pass filtering and noise to the sloping long-term spectrum of speech, which tends to make the high-frequency portion of the spectrum more susceptible to noise masking.

Pickett (1957) and Nooteboom (1968) examined the effects of broadband noise on the perception of vowels. Pickett suggested that vowel identification errors might result when phonetically distinct vowels exhibited similar formant patterns. An analysis of confusion matrices for different noise conditions revealed that listeners frequently confused front vowels (such as [i], with a high second formant) with a corresponding back vowel (e.g., [u], with a low F_2). When the F_2 peak is masked, the vowel is identified as a back vowel with a similar F_1. This error pattern supports the hypothesis that listeners rely primarily on the frequencies of formant peaks to identify vowels (rather than the entire shape of the spectrum), and are predicted by a formant-template model of vowel perception (Scheffers 1983). Scheffers (1983) found that the identification thresholds for synthesized vowels masked by pink noise could be predicted fairly well by the SNR in the region of the second formant. Scheffers found that unvoiced (whispered) vowels had lower thresholds than voiced vowels. He also showed that vowels were easier to identify when the noise was on continuously, or was turned on 20 to 30 ms before the onset of the vowel, compared with a condition where vowels and noise began together.

Pickett (1957) reported that duration cues (differences between long and short vowels) had a greater influence on identification responses when one or more of the formant peaks was masked by noise. This finding serves as an example of the exploitation of signal redundancy to overcome the deleterious effects of spectral masking. It has not been resolved whether results like these reflect a "re-weighting" of importance in favor of temporal over spectral cues or whether the apparent importance of cue B automatically increases when cue A cannot be detected.

3.4 Interrupted Speech and Noise

Miller and Licklider (1950) observed that under some condition the speech signal could be turned on and off periodically without substantial loss of intelligibility. Two factors, the interruption rate and the speech-time fraction, were found to be important. Figure 5.6 shows that intelligibility was lowest for interruption rates below 2 Hz (and a speech-time fraction of 50%), where large fragments of each word are omitted. If the interruption rate was higher (between 10 and 100 interruptions per second), listeners identified more than 80% of the monosyllabic words correctly. Regular, aperiodic, or random interruptions produced similar results, as long as the same constant average interruption rate and speech-time fraction were

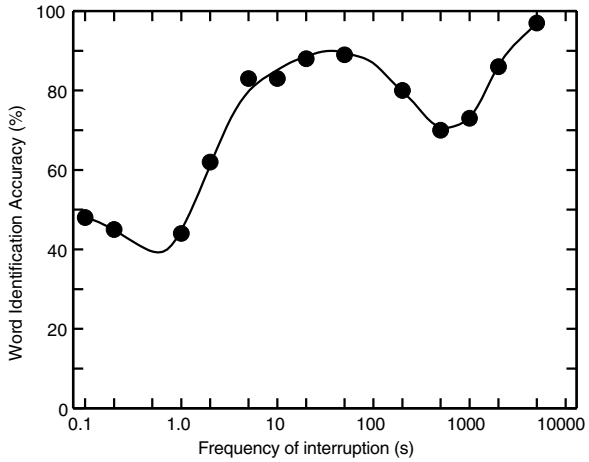

FIGURE 5.6. Word identification accuracy as a function of the rate of interruption for a speech-time fraction of 50%. (After Miller and Licklider 1950.)

maintained. The high intelligibility of interrupted speech is remarkable, considering that near-perfect identification is obtained in conditions where close to half of the power in the speech signal has been omitted. At the optimal (i.e., most intelligible) interruption rate (about 10 interruptions per second), listeners were able to understand conversational speech. Miller and Licklider suggested that listeners were able to do this by "patching together" successive glimpses of signal to reconstruct the intended message. For the speech materials in their sample (phonetically balanced monosyllabic words), listeners were able to obtain, on average, one "glimpse" per phonetic segment (although phonetic segments are not uniformly spaced in time). Miller and Licklider's findings were replicated and extended by Huggins (1975), who confirmed that the optimum interruption rate was around 10 Hz (100 ms) and demonstrated that the effect was at least partially independent of speaking rate. Huggins interpreted the effect in terms of a "gap-bridging" process that contributes to the perception of speech in noise.

Miller and Licklider (1950) also investigated the masking of speech by interrupted noise. They found that intermittent broadband noise maskers interfered less with speech intelligibility than did continuous maskers. An interruption rate of around 15 noise bursts per second produced the greatest release from masking. Powers and Wilcox (1977) have shown that the greatest benefit is observed when the interleaved noise and speech are comparable in loudness.

Howard-Jones and Rosen (1993) examined the possibility that the release from masking by interrupted noise might benefit from an independent analysis of masker alternations in different frequency regions. They

14 proposed that listeners might benefit from a process of "un-comodulated" glimpsing in which glimpses are patched together across different frequency regions at different times. To test this idea they used a "checkerboard" masking noise. The noise was divided into 2, 4, 8, or 16 frequency bands of equal power. The noise bands were switched on and off at a rate of 10 Hz, either synchronously in time ("comodulated" interruptions) or asynchronously, with alternating odd and even bands ("un-comodulated" interruptions) to create a masker whose spectrogram resembled a checkerboard. Evidence for a contribution of un-comodulated glimpsing was obtained when the masker was divided into either two or four bands, resulting in a release from masking of 16 and 6 dB, respectively (compared to 23 dB for fully comodulated bands). The conclusion from this study is that listeners can benefit from un-comodulated glimpsing to integrate speech cues from different frequency bands at different times in the signal.

When speech is interrupted periodically by inserting silent gaps, it assumes a harsh, unnatural quality, and its intelligibility is reduced. Miller and Licklider (1950), using monosyllabic words as stimuli, noted that this harsh quality could be eliminated by filling the gaps with broadband noise. Although adding noise restored the natural quality of the speech, it did not improve intelligibility. Subsequent studies with connected speech found both greater naturalness and higher intelligibility when the silent portions of interrupted speech were filled with noise (Cherry and Wiley 1967; Warren et al. 1997). One explanation is that noise-filled gaps more effectively engage the listener's ability to exploit contextual cues provided by syntactic and semantic continuity (Warren and Obusek 1971; Bashford et al. 1992; Warren 1996).

3.5 Competing Speech

While early studies of the effects of noise on speech intelligibility often used white noise (e.g., Hawkins and Stevens 1950), later studies were interested in exploring more complex forms of noise that are more representative of noisy environments such as cafeterias and cocktail parties (e.g., Duquesnoy 1983; Festen and Plomp 1990; Darwin 1990; Festen 1993; Howard-Jones 15 16 and Rosen 1993; Bronkhorst 2000; Brungart 2001; Brungart et al. 2001). Research on the perceptual separation of speech from competing spoken material has received particular attention because

1. the acoustic structure of the target and masker are similar,
2. listeners with normal hearing perform the separation of voices successfully and with little apparent effort, and
3. listeners with sensorineural hearing impairments find competing speech to be a major impediment to speech communication.

Accounting for the ability of listeners to segregate a mixture of voices and attend selectively to one of them has been described as the "cocktail

party problem" by Cherry (1953). This ability is regarded as a prime example of auditory selective attention (Broadbent 1958; Bregman 1990).

The interfering effect of competing speech is strongly influenced by the number of competing voices present. Figure 5.7 illustrates the effects of competing voices on syllable identification with data from Miller (1947). Miller obtained articulation functions (percent correct identification of monosyllabic words as a function of intensity of the masker) in the presence of one, two, four, six, or eight competing voices. The target voice was always male, while the interfering voices were composed of equal numbers of males and females. A single competing (male) voice was substantially less effective as a masker than two competing voices (one male and one female). Two voices were less effective than four, but there was little subsequent change in masking effectiveness as the number was increased to six and eight voices. When a single competing voice is used as a masker, variation in its overall amplitude creates dips or gaps in the waveform that enable the listener to hear out segments of the target voice. When several voices are present, the masker becomes more nearly continuous in overall amplitude and the opportunity for "glimpsing" the target voice no longer arises.

When speech and nonspeech broadband maskers were compared in terms of their masking effect, competing speech maskers from a single speaker and amplitude-modulated noise were found to produce less masking than steady-state broadband noise (Speaks et al. 1967; Carhart et al. 1969; Gustafsson and Arlinger 1994). The advantage for speech over nonspeech maskers disappeared when several speakers were combined. Mixing

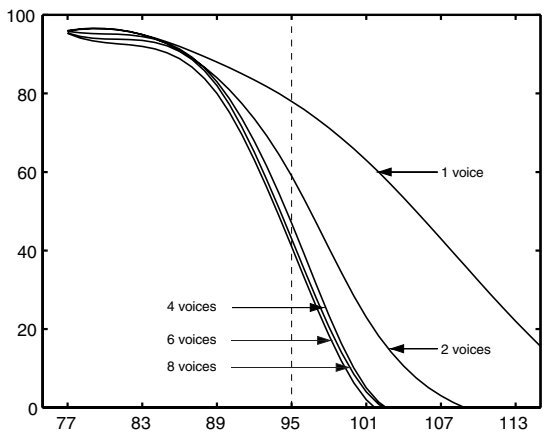

FIGURE 5.7. Syllable identification accuracy as a function of the number of competing voices. The level of the target speech (monosyllabic nonsense words) was held constant at 95 dB. (After Miller 1947.)

the sounds of several speakers produces a signal with a more continuous amplitude envelope and a more uniform spectrum (cf. Fig. 5.1). The masking effect of a mixture of speakers, or a mixture of samples of recorded music (Miller 1947), was similar to that of broadband noise.

A competing voice may interfere with speech perception for at least two reasons. First, it may result in spectral and temporal overlap, which leads to auditory masking. Second, it may interrupt the linguistic processing of the target speech. Brungart and colleagues (Brungart 2001; Brungart et al. 2001) measured the intelligibility of a target phrase masked by one, two, or three competing talkers as a function of SNR and masker type. Performance was generally worse with a single competing talker than with temporally modulated noise with the same long-term average spectrum as the speech. Brungart et al. suggested that part of the interference produced by a masking voice is due to informational masking, distinct from energetic masking caused by spectral and temporal overlap of the signals.

In contrast with these results, Dirks and Bower (1969) and Hygge et al. (1992) obtained similar results for speech maskers played forward or backward. In these studies there was little evidence that the masking effect was enhanced by semantic or syntactic interference of the masker. Their results suggest that the interfering effects of speech maskers can be partially alleviated by temporal dips in the masker that permit the listener to "glimpse" the acoustic structure of the target voice.

Support for the idea that listeners with normal hearing can exploit the temporal modulations associated with a single competing voice comes from studies that compared speech maskers with steady-state and amplitude-modulated noise maskers. There is a large difference in masking effect of a steady-state noise and a modulated noise (or a single interfering voice), as measured by the SRT. Up to 8 dB of masking release is provided by an amplitude-modulated noise masker, compared to the steady-state masker (Duquesnoy 1983; Festen and Plomp 1990; Gustafsson and Arlinger 1994). Speech-spectrum–shaped noise is a more effective masker than a competing voice (Festen and Plomp 1990; Festen 1993; Peters et al. 1998). Speech reception thresholds for sentences in modulated noise are 4 to 6 dB lower than comparable sentences in unmodulated noise. For sentences masked by a competing voice, the masking difference increased to 6 to 8 dB. However, masker modulation does not appear to play a significant role in masking of isolated nonsense syllables or spondee words (Carhart et al. 1969), and hence may be related to the syllable structure of connected speech.

For hearing-impaired listeners the benefits of modulation are reduced, and both types of maskers are equally disruptive (Dirks et al. 1969; Festen and Plomp 1990; Gustafsson and Arlinger 1994). This result is attributable to reduced temporal resolution associated with sensorineural hearing loss (Festen and Plomp 1990). Festen and Plomp (1990) suggested two possible bases for the effect: (1) listening in the temporal dips of the masker, providing a locally favorable SNR; and (2) comodulation masking release

(CMR). Festen (1993) described experiments in which across-frequency coherence of masker fluctuations was disrupted. He concluded that across-frequency processing of masker fluctuations (CMR) makes only a small (about 1.3 dB) contribution to the effect. The effect of masker fluctuation is level-dependent, in a manner consistent with an alternative explanation based on forward masking (the modulation is expected to produce less masking release at low sensation levels because the decay in forward masking is more gradual near threshold).

When listening to a mixture of two voices, listeners with normal hearing have exceptional abilities to hear out components of the composite signal that stem from the same larynx and vocal tract. For example, when a target sentence is combined with an interfering sentence spoken by a different speaker, listeners can correctly identify 70% to 80% of the target words at an SNR of 0 dB (Stubbs and Summerfield 1991). One factor that contributes to intelligibility is auditory grouping and segregation on the basis of f_0 (Brokx and Nooteboom 1982; Scheffers 1983; Assmann and Summerfield 1990, 1994; Darwin and Carlyon 1995; Bird and Darwin 1998). Summerfield 19
and Culling (1992) demonstrated that listeners can exploit simultaneous differences in f_0 to segregate competing voices even at disadvantageous SNRs when the formants of a target voice do not appear as distinct peaks in the composite spectrum. They determined masked identification thresholds for target vowels in the presence of vowel-like maskers. Thresholds were about 15 dB lower when the masker and target differed in f_0 by two semitones (about 12%). At threshold, the formants of the target did not appear as distinct peaks in the composite spectrum envelope but rather as small bumps or "shoulders." An autocorrelation analysis, based on Meddis and Hewitt's (1992) model, revealed that the periodicity of the masker was stronger than that of the target in the majority of frequency channels. Summerfield and Culling proposed that the identity of the target vowel was determined on the basis of the disruption it produced in the periodicity of the masker, rather than on the basis of its own periodicity. This explanation is consistent with models of source segregation that remove the evidence of an interfering voice on the basis of its periodicity (Meddis and Hewitt 1992; Cheveigné 1997).

During voiced speech the pulsing of the vocal folds gives rise to a consistent pattern of periodicity in the waveform and harmonicity in the spectrum. In a mixture of two voices, the periodicity or harmonicity associated with the target voice provides a basis for grouping together signal components with the same f_0. Time-varying changes in f_0 also provide a basis for tracking properties of the voice over time.

Brokx and Nooteboom (1982) demonstrated benefits of differences in average f_0 using LPC-resynthesized speech. Brokx and Nooteboom analyzed natural speech using an LPC vocoder to artificially modify the characteristics of the excitation source and create synthesized, monotone versions of a set of 96 semantically anomalous sentences. They then varied

the difference in fundamental frequency between the target sentence and a continuous speech masker. Identification accuracy was lowest when the target and masker had the same f_0, and gradually improved as a function of increasing difference in f_0. Identification accuracy was lower when the two voices were exactly an octave apart, a condition where every second harmonic of the higher-pitched voice overlaps with a harmonic of the lower f_0.
20 These results were replicated and extended by Bird and Darwin (1998) who used monotone versions of short declarative sentences consisting of entirely voiced sounds. They presented the sentences concurrently in pairs, with one long masker sentence and a short target sentence in each pair. They found an improvement in intelligibility with differences in f_0 between ±2 and ±8 semitones. Using a similar method, Assmann (1999) confirmed the benefits of f_0 difference using both monotone sentence pairs (in which f_0 was held constant) and sentence pairs with natural intonation (in which the natural variation in f_0 was preserved in each sentence, but shifted up or down along the frequency scale to produce the corresponding mean difference in f_0). An unexpected result was that sentences with natural intonation were not significantly more intelligible than monotone sentences, suggesting that f_0 differences are more important for segregating competing speech sounds than time-varying changes in f_0.

In natural environments, competing voices typically originate from different locations in space. A number of laboratory studies have confirmed that a difference in spatial separation can aid the perceptual segregation of competing voices (e.g., Cherry 1953). Yost et al. (1996) presented speech (words, letters, or numbers) to listeners with normal hearing in three listening conditions. In one condition the listener was seated in a sound-deadened room and signals were presented over loudspeakers arranged in a circle around the listener. In a second condition, speech was presented in the free field as in the first condition, but was recorded using a stationary KEMAR manikin and delivered binaurally over headphones to a listener in a remote room. In the third condition, a single microphone was used and the sounds presented monaurally. Sounds were presented individually, in pairs, or in triplets from different randomly chosen subsets of the loudspeakers. Identification scores were highest when the free-field conditions were comparable to the monaural. Intermediate scores were observed under conditions where the binaural recordings were made with the KEMAR manikin. Differences among the conditions were reduced substantially when only two, rather than three, utterances were presented simultaneously, suggesting that listening with two ears in free field is most effective when more than two concurrent sound sources are present.

3.6 Binaural Processing and Noise

When a sound source is located directly in front of an observer in the free field, the acoustic signals reaching the two ears are nearly identical. When

the source is displaced to one side or the other, each ear receives a slightly different signal. Interaural level differences (ILDs) in sound pressure level, which are due to head shadow, and interaural time differences (ITDs) in the time of arrival provide cues for sound localization and can also contribute to the intelligibility of speech, especially under noisy conditions. When speech and noise come from different locations in space, interaural disparities can improve the SRT by up to 10 dB (Carhart 1965; Levitt and Rabiner 1967; Dirks and Wilson 1969; Plomp and Mimpen 1981). Some benefit is derived from ILDs and ITDs, even when listening under monaural conditions (Plomp 1976). This benefit is probably a result of the improved SNR at the ear ipsilateral to the signal.

Bronkhorst and Plomp (1988) investigated the separate contributions of ILDs and ITDs using free-field recordings obtained with a KEMAR manikin. Speech was recorded directly in front of the manikin, and noise with the same long-term spectrum as the speech was recorded at seven different angles in the azimuthal plane, ranging from 0 to 180 degrees in 30-degree steps. Noise samples were processed to contain only ITD or only ILD cues. The binaural benefit was greater for ILDs (about 7 dB) than for ITDs (about 5 dB). In concert, ILDs and ITDs yielded a 10-dB binaural gain, comparable to that observed in earlier studies.

The binaural advantage is frequency dependent (Kuhn 1977; Blauert 1996). Low frequencies are diffracted around the head with relatively little attenuation (a consequence of the wavelength of such signals being appreciably longer than the diameter of the head), while high frequencies (>4 kHz for human listeners) are attenuated to a much greater extent (thus providing a reliable cue based on ILDs in the upper portion of the spectrum). The encoding of ITDs is based on neural phase-locking, which declines appreciably above 1500 Hz (in the upper auditory brain stem). Thus, ITD cues are generally not useful for frequencies above this limit, except when high-frequency carrier signals are modulated by low frequencies. Analysis of the pattern of speech errors in noise suggests that binaural listening may provide greater benefits at low frequencies. For example, in binaural conditions listeners made fewer errors involving manner-of-articulation features, which rely predominantly on low-frequency cues, and they were better able to identify stop consonants with substantial low-frequency energy, such as the velar stops [k] and [g] (Helfer 1994).

3.7 Effects of Reverberation

When speech is spoken in a reverberant environment, the signal emanating from the mouth is combined with reflections that are time-delayed, scaled versions of the original. The sound reaching the listener is a mixture of direct and reflected energy, resulting in temporal "smearing" of the speech signal. Echoes tend to fill the dips in the temporal envelope of speech and increase the prominence of low-frequency energy that masks

the speech spectrum. Sounds with time-invariant cues, such as steady-state vowels, suffer little distortion, but the majority of speech sounds are characterized by changing formant patterns. For speech sounds with time-varying spectra, reverberation leads to a blurring of spectral detail. Hence, speech sounds with rapidly changing spectra (such as stop consonants) are more likely to suffer deleterious effects of reverberation than segments with more stationary formants. Factors that affect speech intelligibility include volume of the enclosure, reverberation time, ambient noise level, the speaker's vocal output level, as well as the distance between speaker and listener. Hearing-impaired listeners are more susceptible to the effects of reverberation than listeners with normal hearing (Finitzo-Hieber and Tillman 1978; Duquesnoy and Plomp 1983; Humes et al. 1986).

An illustration of the effects of reverberation on the speech spectrogram is shown in Figure 5.8. Overall, the most visible effect is the transformation of dynamic features of the spectrogram into more static features. Differences between the spectrogram of the utterance in quiet and in reverberation include:

1. Reverberation fills the gaps and silent intervals associated with vocal-tract closure in stop consonants. For example, the rapid alternation of noise

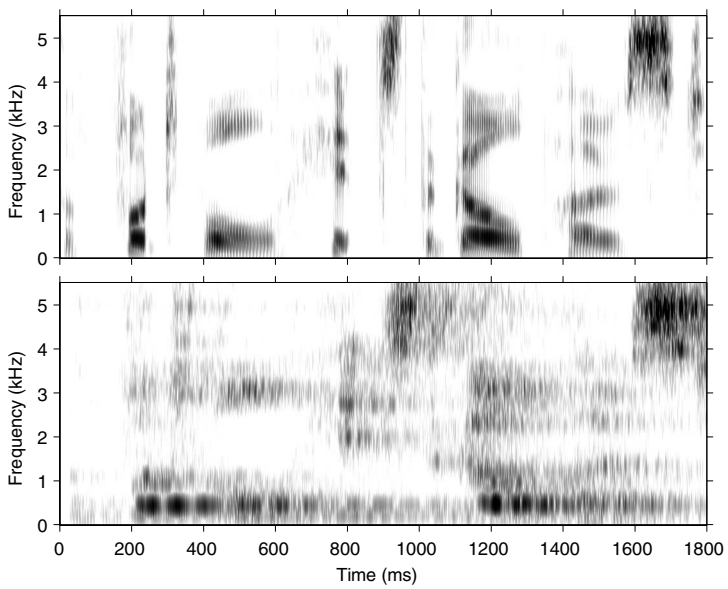

FIGURE 5.8. The upper panel displays the spectrogram of a wideband sentence, "The football hit the goal post," spoken by an adult male. The lower panel shows the spectrogram of a version of the sentence in simulated reverberation, modeling the effect of a highly reverberant enclosure with a reverberation time of 1.4 seconds at a location 2 m from the source.

and silence surrounding the [t] burst in "football" (occurring at about the 300-ms frame on the spectrogram) is blurred under reverberant conditions (lower spectrogram).

2. Both onsets and offsets of syllables tend to be blurred, but the offsets are more adversely affected.

3. Noise bursts (fricatives, affricates, stop bursts) are extended in duration. This is most evident in the [t] burst of the word "hit" (cf. the 900-ms frame in the upper spectrogram).

4. Reverberation blurs the relationship between temporal events, such as the voice onset time (VOT), the time interval between stop release and the onset of voicing. Temporal offsets are blurred, making it harder to determine the durations of individual speech segments, such as the [U] in "football" (at approximately the 200-ms point in the upper spectrogram).

5. Formant transitions are flattened, causing diphthongs and glides to appear as monophthongs, such as the [o^{w}] in "goal" (cf. the 1100-ms frame).

6. Amplitude modulations associated with f_0 are reduced, smearing the vertical striation pattern in the spectrogram during the vocalic portions of the utterance (e.g., during the word "goal").

In a reverberant sound field, sound waves reach the ears from many directions simultaneously and hence their sound pressure levels and phases vary as a function of time and location of both the source and receiver. Plomp and Steeneken (1978) estimated the standard deviation in the levels of individual harmonics of complex tones and steady-state vowels to be about 5.6 dB, while the phase pattern was effectively random in a diffuse sound field (a large concert hall with a reverberation time of 2.2 seconds). This variation is smaller than that associated with phonetic differences between pairs of vowels, and is similar in magnitude to differences in pronunciations of the same vowel by different speakers of the same age and gender (Plomp 1983). Plomp and Steeneken showed that the effects of reverberation on timbre are well predicted by differences between pairs of amplitude spectra, measured in terms of the output levels of a bank of one-third-octave filters. Subsequent studies have confirmed that the intelligibility of spoken vowels is not substantially reduced in a moderately reverberant environment for listeners with normal hearing (Náblek and Letowski 1985). 21

Nábelek (1988) suggested two reasons why vowels are typically well preserved in reverberant environments. First, the spectral peaks associated with formants are generally well defined in relation to adjacent spectral troughs (Leek et al. 1987). Second, the time trajectory of the formant pattern is relatively stationary (Nábelek and Letowski 1985; Nábelek and Dagenais 1986). 21 While reverberation has only a minor effect on steady-state speech segments and monophthongal vowels, diphthongs are affected more

dramatically (as illustrated in Fig. 5.8). Nábelek et al. (1994) noted that reverberation often results in confusions among diphthongs such as [ai] and [au]. Frequently, diphthongs are identified as monophthongs whose onset formant pattern is similar to the original diphthong (e.g., [ai] and [a]). Nábelek et al. proposed that the spectral changes occurring over the final portion of the diphthong are obscured in reverberant conditions by a temporal-smearing process they refer to as "reverberant self-masking." Errors can also result from "reverberant overlap-masking," which occurs when the energy originating from a preceding segment overlaps a following segment. This form of distortion often leads to errors in judging the identity of a syllable-final consonant preceded by a relatively intense vowel, but rarely causes errors in vowel identification per se (Nábelek et al. 1989).

Reverberation tends to "smear" and prolong spectral-change cues, such as formant transitions, smooth out the waveform envelope, and increase the prominence of low-frequency energy capable of masking higher frequencies. Stop consonants are more susceptible to distortion than other consonants, particularly in syllable-final position (Nábelek and Pickett 1974; Gelfand and Silman 1979). When reverberation is combined with background noise, final consonants are misidentified more frequently than initial consonants. Stop consonants, in particular, may undergo "filling in" of the silent gap during stop closure (Helfer 1994). Reverberation tends to obscure cues that specify rate of spectral change (Nábelek 1988), and hence can create ambiguity between stop consonants and semivowels (Liberman et al. 1956). Reverberation results in "perseveration" of formant transitions, and formant transitions tend to be dominated by their onset frequencies. Relational cues, such as the frequency slope of the second formant from syllable onset to vocalic midpoint (Sussman et al. 1991), may be distorted by reverberation, and this distortion may contribute to place-of-articulation errors.

When listening in the free field, reverberation diminishes the interaural coherence of speech because of echoes reaching the listener from directions other than the direct path. Reverberation also reduces the interaural coherence of sound sources and tends to randomize the pattern of ILDs and ITDs. The advantage of binaural listening under noisy conditions is reduced, but not eliminated in reverberant environments (Moncur and Dirks 1967; Nábelek and Pickett 1974). Plomp (1976) asked listeners to adjust the intensity of a passage of read speech until it was just intelligible in the presence of a second passage from a competing speaker. Compared to the case where both speakers were located directly in front of the listener, spatial separation of the two sources produced a 6-dB advantage in SNR. This advantage dropped to about 1 dB in a room with a reverberation time of 2.3 seconds. The echo suppression process responsible for this binaural advantage is referred to as binaural squelching of reverberation and is particularly pronounced at low frequencies (Bronkhorst and Plomp 1988).

The deterioration of spatial cues in reverberant environments may be one reason why listeners do not use across-frequency grouping by common ITD to separate sounds located at different positions in the azimuthal plane. Culling and Summerfield (1995) found no evidence that listeners could exploit the pattern of ITDs across frequency for the purpose of grouping vocalic formants that share the same ITD as a means of segregating them from formants with different ITDs. Their results were corroborated by experiments showing that listeners were unable to segregate a harmonic from a vowel when the remaining harmonics were assigned a different ITD (Darwin and Hukin 1997). Some segregation was obtained when ITDs were combined with other cues (e.g., differences in f_0 and onset asynchrony), but the results suggest that ITDs exert their influence by drawing attention to sounds that occupy a specific location in space, rather than by grouping frequency components that share a common pattern of ITD (Darwin 1997; Darwin and Hukin 1998).

Binaural processes that minimize the effects of reverberation are supplemented by monaural processes to offset the deleterious effects of reverberation (Watkins 1988, 1991; Darwin 1990). In natural environments high frequencies are often attenuated by obstructions, and the longer wavelengths of low-frequency signals allow this portion of the spectrum to effectively bend around corners. Darwin et al. (1989) examined the effects of imposing different spectral slopes on speech signals to simulate such effects of reverberant transmission channels. A continuum of vowels was synthesized, ranging from [ɪ] to [ɛ] within the context of various [bVt] words, and the vowels filtered in such a manner as to impose progressively steeper slopes in the region of the first formant. When the filtered signals were presented in isolation, the phonemic boundary between the vocalic categories shifted in accordance with the apparent shift in the F_1 peak. However, when the filtered speech was presented after a short carrier phrase filtered in comparable fashion, the magnitude of the boundary shift was reduced. This result is consistent with the idea that listeners perceptually compensate for spectral tilt. However, this compensation may occur only under extreme conditions, since it was present only with extreme filtering (30-dB change in spectral slope) and did not completely eliminate the effects of filtering.

Watkins (1991) used an LPC vocoder to determine the ability of listeners to compensate for distortions of the spectrum envelope. He synthesized a set of syllables along a perceptual continuum ranging from [Iç] ("itch") to [ɛç] ("etch") by varying the F_1 frequency of the vowel and processing each segment with a filter whose transfer function specified the *difference* between the spectral envelopes of the two end-point vowels (i.e., [ɪ] minus [ɛ], as well as its inverse). The two filtering operations resulted in shifts of the phonemic boundary associated with F_1 toward a higher apparent formant peak when the first form of subtractive filter was used, and toward a lower apparent peak for the second type of filter. However, when the signals were embedded in a short carrier phrase processed in a compara-

ble manner, the magnitude of the shift was reduced, suggesting that listeners are capable of compensating for the effects of filtering if given sufficiently long material with which to adapt. The shifts were not entirely eliminated by presenting the carrier phrase and test signals to the opposing ears or by using different apparent localization directions (by varying the ITD). Subsequent studies (Watkins and Makin 1994, 1996) showed that perceptual compensation was based on the characteristics of the following, as well as those of the preceding, signals. The results indicate that perceptual compensation does not reflect peripheral adaptation directly, but is based on some form of central auditory process.

When harmonically rich signals, such as vowels and other voiced segments, are presented in reverberation, echoes can alter the sound pressure level of individual harmonics and scramble the original phase pattern, but the magnitude of these changes is generally small relative to naturally occurring differences among vocalic segments (Plomp 1983). However, when the f_0 is nonstationary, the echoes originating from earlier time points overlap with later portions of the waveform. This process serves to diffuse cues relating to harmonicity, and could therefore reduce the effectiveness of f_0 differences to segregate competing voices. Culling et al. (1994) confirmed this supposition by simulating the passage of speech from a speaker to the ears of a listener in a reverberant room. They measured the benefit afforded by f_0 differences under reverberant conditions sufficient to counteract the effects of spatial separation (produced by a 60-degree difference in azimuth). They showed that this degree of reverberation reduces the ability of listeners to use f_0 differences in segregating pairs of concurrent vowels under conditions where f_0 is changing, but not in the condition where both masker and target had stationary f_0s. When f_0 is modulated by an amount equal to or greater than the difference in f_0 between target and masker, the benefits of using a difference in f_0 are no longer present.

The effects of reverberation on speech intelligibility are complex and not well described by a spectral-based approach such as the AI. This is illustrated in Figure 5.8, which shows that reverberation radically changes the appearance of the speech spectrogram and eliminates or distorts many traditional speech cues such as formant transitions, bursts, and silent intervals. Reverberation thus provides an illustration of perceptual constancy in speech perception. Perceptual compensation for such distortions is based on a number of different monaural and binaural "dereverberation" processes acting in concert. Some of these processes operate on a local (syllable-internal) basis (e.g., Nábelek et al. 1989), while others require prior exposure to longer stretches of speech (e.g., Watkins 1988; Darwin et al. 1989).

A more quantitatively predictive means of studying the impact of reverberation is afforded by the TMTF, and an accurate index of the effects of reverberation in intelligibility is provided by the STI (Steeneken and Houtgast 1980; Humes et al. 1987). Such effects can be modeled as a low-

pass filtering of the modulation spectrum. Although the STI is a good predictor of overall intelligibility, it does not attempt to model processes underlying perceptual compensation. In effect, the STI transforms the effects of reverberation into an equivalent change in SNR. However, several properties of speech intelligibility are not well described by this approach. First, investigators have noted that the pattern of confusion errors is not the same for noise and reverberation. The combined effect of reverberation and noise is more harmful than noise alone (Nábelek et al. 1989; Takata and Nábelek 1990; Helfer 1994). Second, some studies suggest there may be large individual subject differences in susceptibility to the effects of reverberation (Nábelek and Letowski 1985; Helfer 1994). Third, children are affected more by reverberation than adults, and such differences are observed up to age 13, suggesting that acquired perceptual strategies contribute to the ability of compensating for reverberation (Finitzo-Hieber and Tillman 1978; Nábelek and Robinson 1982; Neuman and Hochberg 1983). Fourth, elderly listeners, with normal sensitivity, are more adversely affected by reverberation than younger listeners, suggesting that aging may lead to a diminished ability to compensate for reverberation (Gordon-Salant and Fitzgibbons 1995; Helfer 1992).

3.8 Frequency Response of the Communication Channel

In speech perception, the vocal-tract resonances provide phonetic and lexical information, as well as information about the source, such as personal identity, gender, age, and dialect of the speaker (Kreiman 1997). However, under everyday listening conditions the spectral envelope is frequently distorted by properties of the transmission channel. Indoors, sound waves are reflected and scattered by various surfaces (e.g., furniture and people), while room resonances and antiresonances introduce peaks and valleys into the spectrum. Outdoor listening environments contain potential obstructions such as buildings and trees, and exhibit atmospheric distortions due to wind and water vapor. For this reason damping is not uniform as a function of frequency. In general, high-frequency components tend to be absorbed more rapidly than their low-frequency counterparts. As a result of the need to communicate efficiently in all of these conditions, listeners compensate for a variety of distortions of the communication channel rapidly and without conscious effort.

3.8.1 Low-Pass and High-Pass Filtering

Fletcher (1953) studied the effects of low-pass (LP) and high-pass (HP) filtering on the intelligibility of nonsense syllables. His objective was to measure the independent contribution of the low- and high-frequency channels. Eliminating the high frequencies reduced the articulation scores of consonants more than vowels, while eliminating the low-frequency portion

of the spectrum had the opposite effect. Fletcher noted that the articulation scores for both the LP and HP conditions did not actually sum to the full-band score. He developed a model, the AI (Fletcher and Galt 1950), as a means of transforming the partial articulation scores (Allen 1994) into an additive form (cf. section 2.1). Accurate predictions of phone and syllable articulation were obtained using a model that assumed that (1) spectral information is processed independently in each frequency band and (2) is combined in an "optimal" way to derive recognition probabilities. As discussed in section 2.1, the AI generates accurate and reliable estimates of the intelligibility of filtered speech based on the proportion of energy within the band exceeding the threshold of audibility and the width of the band. One implication of the model is that speech "features" (e.g., formant peaks) are extracted from each frequency band independently, a strategy that may contribute to noise robustness (Allen 1994).

Figure 5.9 illustrates the effects of LP and HP filtering on speech intelligibility. Identification of monosyllabic nonsense words remains high when LP-filtered at a cutoff frequency of 3 kHz or greater, or HP-filtered at a cutoff frequency of 1 kHz or lower. For a filter cutoff around 2 kHz, the effects of LP and HP filtering are similar, resulting in intelligibility of around 68% (for nonsense syllables).

When two voices are presented concurrently it is possible to improve the SNR by restricting the bandwidth of one of the voices. Egan et al. (1954) found that HP-filtering either voice with a cutoff frequency of 500 Hz led to improved articulation scores. Spieth and Webster (1955) confirmed that

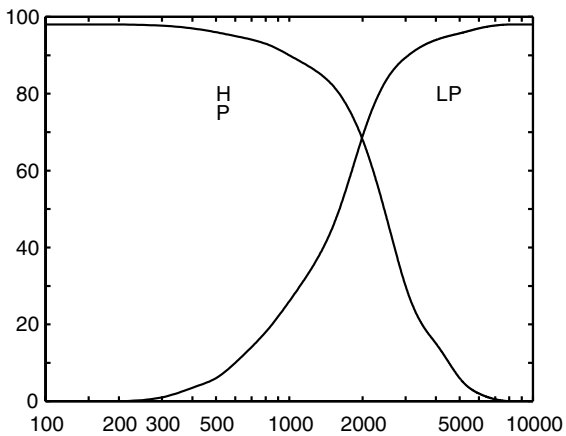

FIGURE 5.9. Effects of high-pass and low-pass filtering on the identification of monosyllabic nonsense words. (After French and Steinberg 1947.)

differential filtering led to improved scores whenever one of the two voices was filtered, regardless of whether such filtering was imposed on the target or interfering voice. Intelligibility was higher when one voice was LP-filtered and the other HP-filtered, compared to the case where both voices were unfiltered. The effectiveness of the filtering did not depend substantially on the filter-cutoff frequency (565, 800, or 1130 Hz for the HP filter, and 800, 1130, and 1600 Hz for the LP filter). Egan et al. (1954) found that intensity differences among the voices could be beneficial. Slight attenuation of the target voice provided a small benefit, offset, in part, by the increased amount of masking exerted by the competing voice. Presumably, such benefits of attenuation are a consequence of perceptual grouping processes sensitive to common amplitude modulation. Webster (1983) suggested that any change in the signal that gives one of the voices a "distinctive sound" could lead to improved intelligibility.

3.8.2 Bandpass and Bandstop Filtering

Several studies have examined the effects of narrow, bandpass (ca. one-third octave) filtering on the identification of vowels (Lehiste and Peterson 1959; Carterette and Møller 1964; Castle 1964). Two conclusions emerge 28
from these studies. First, vowel identification is substantially reduced, but remains above chance, when the signals are subjected to such bandpass filtering. Second, the pattern of errors is not uniform but varies as a function of the intended vowel category—a conclusion not in accord with template theories of vowel perception.[7] For example, when the filter is centered near the first formant, a front vowel may be confused for a back vowel with similar F_1 (e.g., American English [e] is heard as [o]), consistent with the observation that back vowels (e.g., [o]) can be approximated using only a single formant, while front vowels (e.g., [e]) cannot (Delattre et al. 1952).

The studies of LP and HP filtering, reviewed in section 3.8.1, indicate that speech intelligibility is not substantially reduced by removing that portion of the spectrum below 1 kHz or above 3 kHz. In addition, speech that is band limited between 0.3 and 3.4 kHz (i.e., telephone bandwidth) is only marginally less intelligible than full-spectrum speech. These findings suggest that frequencies between 0.3 and 3.4 kHz provide the bulk of the information in speech. However, several studies have shown that speech can withstand disruption of the midfrequency region without substantial loss of intelligibility. Lippmann (1996b) filtered CVC nonsense syllables to remove the frequency band between 0.8 and 3 kHz and found that speech intelligi-

[7] Note that this applies equally to "whole-spectrum" and feature-based models that classify vowels on the basis of template matching using the frequencies of the two or three lowest formants.

bility was not substantially reduced (better than 90% correct consonant identification from a 16-item set). Warren et al. (1995) reported high intelligibility for everyday English sentences that had been filtered using narrow bandpass filters, a condition they described as "listening through narrow spectral slits." With one-third-octave filter bandwidths, about 95% of the words could be understood in sentences filtered at center frequencies of 1100, 1500, and 2100 Hz. Even when the bandwidth was reduced to 1/20th of an octave, intelligibility was about 77% for the 1500-Hz band.

The high intelligibility of spectrally limited sentences can be attributed, in part, to the ability of listeners to exploit the linguistic redundancy in everyday English sentences. Stickney and Assmann (2001) replicated Warren et al.'s findings using gammatone filters (Patterson et al. 1992) with
29 bandwidths of one ERB, chosen to match psychophysical estimates of auditory filter bandwidth (Moore and Glasberg 1987). Listeners identified the final keywords in high-predictability sentences from the Speech Perception in Noise (SPIN) test (Kalikow et al. 1977) at rates similar to those reported by Warren et al. (between 82% and 98% correct for bands centered at 1500, 2100, and 3000 Hz). However, performance dropped by about 20% when low-predictability sentences were used, and by a further 23% when the filtered, final keywords were presented in isolation. These findings highlight the importance of linguistic redundancy (provided both within each sentence, and in the context of the experiment where reliable expectations about the prosody, syntactic form, and semantic content of the sentences are established). Context helps to sustain a high level of intelligibility even when the acoustic evidence for individual speech sounds is extremely sparse.

4. Perceptual Strategies for Retrieving Information from Distorted Speech

The foregoing examples demonstrate that speech communication is a remarkably robust process. Its resistance to distortion can be attributed to many factors. Section 2 described acoustic properties of speech that contribute to its robustness and discussed several strategies used by speakers to improve intelligibility under adverse listening conditions. Section 3 reviewed the spectral and temporal effects of distortions that arise naturally in everyday environments and discussed their perceptual consequences. The overall conclusion is that the information in speech is shielded from distortion in several ways. First, peaks in the envelope of the spectrum provide robust cues for the identification of vowels and consonants even when the spectral valleys are obscured by noise. Second, periodicity in the waveform reflects the fundamental frequency of voicing, allowing listeners to group together components that stem from the same voice across frequency and time in order to segregate them from competing signals (Brokx

and Nooteboom 1982; Bird and Darwin 1998). Third, at disadvantageous SNRs, the formants of voiced sounds can exert their influence by disrupting the periodicity of competing harmonic signals or by disrupting the interaural correlation of a masking noise (Summerfield and Culling 1992; Culling and Summerfield 1995). Fourth, the amplitude modulation pattern across frequency bands can serve to highlight informative portions of the speech signal, such as prosodically stressed syllables. These temporal modulation patterns are redundantly specified in time and frequency, making it possible to remove large amounts of the signal via gating in the time domain (e.g., Miller and Licklider 1950) or filtering in the frequency domain (e.g., Warren et al. 1995). Even when the spectral details and periodicity of voiced speech are eliminated, intelligibility remains high if the temporal modulation structure is preserved in a small number of bands (Shannon et al. 1995). However, speech processed in this manner is more susceptible to interference by other signals.

In this section we consider the perceptual and cognitive strategies used by listeners to facilitate the extraction of information from speech signals corrupted by noise and other distortions of the communication channel. Background noise and distortion generally lead to a reduction in SNR, as portions of the signal are rendered inaudible or are masked by other signals. Masking, the inability to resolve auditory events closely spaced in time and frequency, is a consequence of the fact that the auditory system has limited frequency selectivity and temporal resolution (Moore 1995). The processes described below can be thought of as strategies used by listeners to overcome these limitations.

In sections 4.1 and 4.2 we consider the role of two complementary strategies for recovering information from distorted speech: glimpsing and tracking. Glimpsing exploits moment-to-moment fluctuations in SNR to focus auditory attention on temporal regions of the composite signal where the target voice is best defined. Tracking processes exploit moment-to-moment correlations in fundamental frequency, amplitude envelope, and formant pattern to group together components of the signal originating from the same voice.

Glimpsing and tracking are low-level perceptual processes that require an ongoing analysis of the signal within a brief temporal window, and both can be regarded as *sequential* processes. Perceptual grouping also involves *simultaneous* processes (Bregman 1990), as when a target voice is separated from background signals on the basis of either a static difference in fundamental frequency (Scheffers 1983), or differences in interaural timing and level (Summerfield and Culling 1995).

In the final subsections of the chapter we consider additional processes (both auditory and linguistic) that help listeners to compensate for distortions of the communication channel. In section 4.3 we examine the role of perceptual grouping and adaptation in the enhancement of signal onsets. In section 4.4 we review evidence for the existence of central processes

that compensate for deformations in the frequency responses of communication channels, and we consider their time course. Finally, in section 4.5, we briefly consider how linguistic and pragmatic context helps to resolve the ambiguities created by gaps and discontinuities in the signal, and thereby contributes to the intelligibility of speech under adverse acoustic conditions.

4.1 Glimpsing

In vision, glimpsing occurs when an observer perceives an object based on fragmentary evidence (i.e., when the object is partly obscured from view). It is most effective when the object is highly familiar (e.g., the face of a friend) and when it serves as the focus of attention. Visual objects can be glimpsed from a static scene (e.g., a two-dimensional image). Likewise, auditory glimpsing involves taking a brief "snapshot" from an ongoing temporal sequence. It is the process by which distinct regions of the signal, separated in time, are linked together when intermediate regions are masked or deleted. Empirical evidence for the use of a glimpsing strategy comes from a variety of studies in psychoacoustics and speech perception. The following discussion offers some examples and then considers the mechanism that underlies glimpsing in speech perception.

In comodulation masking release, the masked threshold of a tone is lower in the presence of an amplitude-modulated masker (with correlated amplitude envelopes across different and widely separated auditory channels) compared to the case where the modulation envelopes at different frequencies are uncorrelated (Hall et al. 1984). Buus (1985) proposed a model of CMR that implements the strategy of "listening in the valleys" created by the masker envelope. The optimum time to listen for the signal is when the envelope modulations reach a minimum. Consistent with this model is the finding that CMR is found only during periods of low masker energy, that is, in the valleys where the SNR is highest (Hall and Grose 1991).

Glimpsing has been proposed as an explanation for the finding that modulated maskers produce less masking of connected speech than unmodulated maskers. Section 3.5 reviewed studies showing that listeners with normal hearing can take advantage of the silent gaps and amplitude minima in a masking voice to improve their identification of words spoken by a target voice. The amplitude modulation pattern associated with the alternation of syllable peaks in a competing sentence occur at rates between 4 and 8 Hz (see section 2.5). During amplitude minima of the masker, entire syllables or words of the target voice can be glimpsed.

Additional evidence for glimpsing comes from studies of the identification of concurrent vowel pairs. When two vowels are presented concurrently, they are identified more accurately if they differ in f_0 (Scheffers 1983). When the difference in f_0 is small (less than one semitone, 6%), cor-

responding low-frequency harmonics from the two vowels occupy the same auditory filter and beat together, alternately attenuating and then reinforcing one another. As a result, there can be segments of the signal where the harmonics defining the F_1 of one vowel are of high amplitude and hence are well defined, while those of the competing vowel are poorly defined. The variation in identification accuracy as a function of segment duration suggests that listeners can select these moments to identify the vowels (Culling and Darwin 1993a, 1994; Assmann and Summerfield 1994).

Supporting evidence for glimpsing comes from a model proposed by Culling and Darwin (1994). They applied a sliding temporal window across the vowel pair, and assessed the strength of the evidence favoring each of the permitted response alternatives for each position of the window. Because the window isolated those brief segments where beating resulted in a particularly favorable representation of the two F_1s, strong evidence favoring the vowels with those F_1s was obtained. In effect, their model was a computational implementation of glimpsing. Subsequently, their model was extended to account for the improvement in identification of a target vowel when the competing vowel is preceded or followed by formant transitions (Assmann 1995, 1996). These empirical studies and modeling results suggest that glimpsing may account for several aspects of concurrent vowel perception.

The ability to benefit from glimpsing depends on two separate processes. First, the auditory system must perform an analysis of the signal with a sliding time window to search for regions where the property of the signal being sought is most evident. Second, the listener must have some basis for distinguishing target from masker. In the case of speech, this requires some prior knowledge of the structure of the signal and the masker (e.g., knowledge that the target voice is female and the masker voice is male). Further research is required to clarify whether glimpsing is the consequence of a unitary mechanism or a set of loosely related strategies. For example, the time intervals available for glimpsing are considerably smaller for the identification of concurrent vowel pairs (on the order of tens of milliseconds) compared to pairs of sentences, where variation in SNR provides intervals of 100 ms or longer during which glimpsing could provide benefits.

4.2 Tracking

Bregman (1990) proposed that the perception of speech includes an early stage of auditory scene analysis in which the components of a sound mixture are grouped together according to their sources. He suggested that listeners make use of gestalt grouping principles such as proximity, good continuation, and common fate to link together the components of signals and segregate them from other signals. Simultaneous grouping processes make use of co-occuring properties of signals, such as the frequency spacing

of harmonics and the shape of the spectrum envelope, in order to group together components of sound that emanate from the same source. Sequential grouping is used to forge links over time with the aid of tracking processes. Tracking exploits correlations in signal characteristics across time and frequency to group together acoustic components originating from the same larynx and vocal tract.

Two properties of speech provide a potential basis for tracking. First, changes in the rate of vocal-fold vibration during voiced speech tend to be graded, giving rise to finely granulated variations in pitch. Voiced signals have a rich harmonic structure, and hence changes in f_0 generate a pattern of correlated changes across the frequency spectrum. Second, the shape of the vocal tract tends to change slowly and continuously during connected speech, causing the trajectories of formant peaks to vary smoothly in time and frequency. When the trajectories of the formants and f_0 are partially obscured by background noise and other forms of distortion, the perceptual system is capable of recovering information from the distorted segments by a process of tracking (or trajectory extrapolation).

4.2.1 Fundamental Frequency Tracking

Despite the intuitive appeal of the idea that listeners track a voice through background noise, the empirical support for such a tracking mechanism, sensitive to f_0 modulation, is weak (Darwin and Carlyon 1995). Modulation of f_0 in a target vowel can increase its prominence relative to a steady-state masker vowel (McAdams 1989; Marin and McAdams 1991; Summerfield
32 and Culling 1992; Culling and Summerfield 1995). However, there is little evidence that listeners can detect the coherent (across-frequency) changes produced by f_0 modulation (Carlyon 1994). Gardner et al. (1989) were able to induce alternative perceptual groupings of subsets of formants by synthesizing them with different stationary f_0s, but not with different patterns
32 of f_0 modulation. Culling and Summerfield (1995) found that coherent f_0 modulation improved the identification of a target vowel presented in a background of an unmodulated masker vowel. However, the improvement occurred both for coherent (sine phase) and incoherent (random phase) sinusoidal modulation of the target. Overall, these results suggest that f_0 modulation can affect the perceptual prominence of a vowel but does not provide any benefit for sound segregation. In continuous speech, the benefits of f_0 modulation may have more to do with momentary differences in instantaneous f_0 between two voices (providing opportunities for simultaneous grouping processes and glimpsing) than with correlated changes in different frequency regions. One reason why f_0 modulation does not help may be that the harmonicity in voiced speech obviates the need for a computationally expensive operation of tracking changes in the frequencies of harmonics (Summerfield 1992). A further reason is that in enclosed environments, reverberation tends to blur the pattern of modulation created by

changes in the frequencies of the harmonics, making f_0 modulation an unreliable source of information (Gardner et al. 1989; Culling et al. 1994).

4.2.2 Formant Tracking

Bregman (1990) suggested that listeners might exploit the trajectories of formant peaks to track the components of a voice through background noise. In section 2.1 it was suggested that peaks in the spectrum envelope provide robust cues because they are relatively impervious to the effects of background noise, as well as to modest changes in the frequency response of communication channels and deterioration in the frequency selectivity of the listener. A complicating factor is that the trajectories of different formants are often uncorrelated (Bregman 1990). For example, during the transition from the consonant to the vowel in the syllable [da], the frequency of the first formant increases while the second formant decreases. Moreover, in voiced speech the individual harmonics also generate peaks in the fine structure of the spectrum. Changes in the formant patterns are independent of changes in the frequencies of harmonics, and thus listeners need to distinguish among different types of spectral peaks in order to track formants over time. The process is further complicated by the limited frequency selectivity in hearing [i.e., the low-order harmonics of vowels and other voiced signals are individually resolved, while the higher harmonics are not (Moore and Glasberg 1987)].

Despite the intuitive plausibility of the idea that listeners track formants through background noise, there is little direct evidence to support its role in perceptual grouping. Assmann (1995) presented pairs of concurrent vowels in which one member of the pair had initial or final flanking formant transitions that specified a [w], [j], or [l] consonant. He found that the addition of formant transitions helped listeners identify the competing vowel, but did not help identify the vowel to which they were linked. The results are not consistent with a formant-tracking process, but instead support an alternative hypothesis: formant transitions provide a time window over which the formant pattern of a competing vowel can be glimpsed.

Indirect support for perceptual extrapolation of frequency trajectories comes from studies of frequency-modulated tones that lie on a smooth temporal trajectory. When a frequency-modulated sinusoid is interrupted by noise or a silent gap, listeners hear a continuous gliding pitch (Ciocca and Bregman 1987; Kluender and Jenison 1992). This illusion of continuity is also obtained when continuous speech is interrupted by brief silent gaps or noise segments (Warren et al. 1972). In natural environments speech is often interrupted by extraneous impulsive noise, such as slamming doors, barking dogs, and traffic noise, that masks portions of the speech signal. Warren et al. describe a perceptual compensatory mechanism that appears to "fill in," or restore, the masked portions of the original signal. This process is called auditory induction and is thought to occur at an unconscious level

since listeners are unaware that the perceptually restored sound is actually missing.

Evidence for auditory induction comes from a number of studies that
33 have examined the effect of speech interruptions (Verschuure and Brocaar 1983; Bashford et al. 1992; Warren et al. 1997). These studies show that the intelligibility of interrupted speech is higher when the temporal gaps are filled with broadband noise. Adding noise provides benefits for conditions with high-predictability sentences, as well as for low-predictability sentences, but not with isolated nonsense syllables (Miller and Licklider 1950; Bashford et al. 1992; Warren et al. 1997). Warren and colleagues (Warren 1996; Warren et al. 1997) attributed these benefits of noise to a "spectral restoration" process that allows the listener to "bridge" noisy or degraded portions of the speech signal. Spectral restoration is an unconscious and automatic process that takes advantage of the redundancy of speech to minimize the interfering effects of extraneous signals. It is likely that spectral restoration involves the evocation of familiar or overlearned patterns from long-term memory (or schemas; Bregman 1990) rather than the operation of tracking processes or trajectory extrapolation.

4.3 Role of Adaptation and Grouping in Enhancing Onsets

A great deal of information is conveyed in temporal regions where the speech spectrum is changing rapidly (Stevens 1980). The auditory system is particularly responsive to such changes, especially when they occur at
34 the onsets of signals (Palmer 1996; see Palmer and Shamma, Chapter 4). For example, auditory-nerve fibers show increased rates of firing at the onsets of syllables and during transient events such as stop consonant bursts (Delgutte 1996). Such "adaptation" is associated with a decline in discharge rate observed over a period of prolonged stimulation and is believed to arise because of the depletion of neurotransmitter in the synaptic junction between inner hair cells and the auditory nerve (Smith 1979). The result is a sharp increase in firing rate at the onset of each pitch pulse, syllable or word, followed by a gradual decline to steady-state levels. It has been suggested that adaptation plays an important role in enhancing the spectral contrast between successive signals, and increases the salience of a stimulus immediately following its onset (Delgutte and Kiang 1984; Delgutte 1996).

Adaptation has also been suggested as an explanation for the phenomenon of psychophysical enhancement. *Enhancement* is the term used to describe the increase in perceived salience of a frequency component omitted from a broadband sound when it is subsequently reintroduced (Viemeister 1980; Viemeister and Bacon 1982). Its relevance for speech was demonstrated by Summerfield et al. (1984, 1987), who presented a sound whose spectral envelope was the "complement" of a vowel (i.e., formant peaks were replaced by valleys and vice versa) followed by a tone complex

with a flat amplitude spectrum. The flat-spectrum sound was perceived as having a timbral quality similar to the vowel whose complement had preceded it. Summerfield and Assmann (1989) showed that the identification of a target vowel in the presence of a competing masker vowel was substantially improved if the vowel pair was preceded by a precursor with the same spectral envelope as the masker. By providing prior exposure to the spectral peaks of the masker vowel, the precursor served to enhance the spectral peaks in the target vowel. These demonstrations collectively illustrate the operation of an auditory mechanism that enhances the prominence of spectral components subjected to sudden changes in amplitude. It may also play an important role in compensating for distortions of the communication channel by emphasizing frequency regions containing newly arriving energy relative to background components (Summerfield et al. 1984, 1987). Enhancement is thus potentially attributable to the reduction in discharge rate in auditory-nerve fibers whose characteristic frequencies (CFs) are close to spectral peaks of the precursor. Less adaptation will appear in fibers whose CFs occur in the spectral valleys. Hence, newly arriving sounds generate higher discharge rates when their spectral components stimulate unadapted fibers (tuned to the spectral valleys of the precursor) than when they stimulate adapted fibers (tuned to the spectral peaks). In this way the neural response to newly arriving signals could be greater than the response to preexisting components.

An alternative explanation for enhancement assumes that this perceptual phenomenon is the result of central grouping processes that link auditory channels with similar amplitude envelopes (Darwin 1984; Carlyon 1989). According to this grouping account, the central auditory system selectively enhances the neural response in channels that display abrupt increases in level. Central grouping processes have been invoked to overcome several problems faced by the peripheral adaptation account (or a related account based on the adaptation of suppression; Viemeister and Bacon 1982). First, under some circumstances, enhancement has been found to persist for as long as 20 seconds, a longer time period than the recovery time constants for adaptation of nerve fibers in the auditory periphery (Viemeister 1980). Second, while adaptation is expected to be strongly level-dependent, Carlyon (1989) demonstrated a form of enhancement whose magnitude was *not* dependent on the level of the enhancing stimulus (but cf. Hicks and Bacon 1992). Finally, in physiological recordings from peripheral auditory-nerve fibers, Palmer et al. (1995) found no evidence for an increased gain in the neural responses of fibers tuned to stimulus components that evoke enhancement. The conclusion is that peripheral adaptation contributes to the enhancement effect, but does not provide a complete explanation for the observed effects. This raises the question, What is grouping and how does it relate to peripheral adaptation? A possible answer is that adaptation in peripheral analysis highlights frequency channels in which abrupt increments in spectral level have occurred

(Palmer et al. 1995). Central grouping processes must then establish whether these increments have occurred concurrently in different frequency channels. If so, the "internal gain" in those channels is elevated relative to other channels.

4.4 Compensation for Communication Channel Distortions

Nonuniformities in the frequency response of a communication channel can distort the properties of the spectrum envelope of speech, yet unintelligi-
35 bility is relatively unaffected by manipulations of spectral tilt (Dijkhuisen et al. 1987; Klatt 1989) or by the introduction of a broad peak into the frequency response of a hearing aid (Buuren et al. 1996). A form of perceptual compensation for spectral-envelope distortion was demonstrated by Watkins and colleagues (Watkins 1991; Watkins and Makin 1994, 1996), who found that listeners compensate for complex changes in the frequency response of a communication channel when identifying a target word embedded in a brief carrier sentence. They synthesized a continuum of sounds whose end points defined one of two test words. They showed that the phoneme boundary shifted when the test words followed a short carrier phrase that was filtered using the inverse of the spectral envelope of the vowel to simulate a transmission channel with a complex frequency response. The shift in perceived quality was interpreted as a form of perceptual compensation for the distortion introduced by the filter. Watkins and Makin showed that the effect persists, in reduced form, in conditions where the carrier phrase follows the test sound, when the carrier is presented to the opposite ear, and when a different pattern of interaural timing is applied. For these reasons they attributed the perceptual shifts to a central (as opposed to peripheral) auditory (rather than speech-specific) process that compensates for distortions in the frequency responses of communication channels.

The effects described by Watkins and colleagues operate within a very brief time window, one or two syllables at most. There are indications of more gradual forms of compensation for changes in the communication channel. Perceptual acclimatization is a term often used to describe the long-term process of adjustment to a hearing aid (Horwitz and Turner 1997). Evidence for perceptual acclimatization comes from informal observations of hearing-aid users who report that the benefits of amplification are greater after a period of adjustment, which can last up to several weeks in duration. Gatehouse (1992, 1993) found that some listeners fitted with a single hearing aid understand speech more effectively with their aided ear at high presentation levels, but perform better with their unaided ear at low sound pressure levels. He proposed that each ear performs best when receiving a pattern of stimulation most like the one it commonly receives. The internal representation of the spectrum is assumed to change in a fre-

quency-dependent way to adapt to alterations of the stimulation pattern. Such changes have been observed to take place over periods as long as 6 to 12 weeks. In elderly listeners, this may involve a process of relearning the phonetic interpretation of (high-frequency) speech cues that were previously inaudible. Reviews of the contribution of perceptual acclimatization have concluded, however, that the average increase in hearing-aid benefit over time is small at best (Turner and Bentler 1998); the generality of this phenomenon bears further study.

Sensorineural hearing loss is often associated with elevated thresholds in the high-frequency region. It has been suggested (Moore 1995) that there may be a remapping of acoustic cues in speech perception by hearing-impaired listeners, with greater perceptual weight placed on the lower frequencies and on the temporal structure of speech. An extreme form of this remapping is seen with cochlear-implant users, for whom the spectral fine structure and tonotopic organization of speech is greatly reduced (Rosen 1992; see Clark, Chapter 8). For such listeners, temporal cues may play an enhanced role. Most cochlear implant users show a gradual process of adjustment to the device, accompanied by improved speech recognition performance. This suggests that acclimatization processes may shift the perceptual weight assigned to different aspects of the temporal structure of speech preserved by the implant.

Shannon et al. (1995) showed that listeners with normal hearing could achieve a high degree of success in understanding speech that retained only the temporal information in four broad frequency channels and lacked both voicing information and spectral fine structure (see section 2.5). Rosen et al. (1998) used a similar processor to explore the effects of shifting the bands so that each temporal envelope stimulated a frequency band between 1.3 and 2.9 octaves higher in frequency than the one from which it was originally obtained. Similar shifts may be experienced by multichannel cochlear implants when the apical edge of the electrode reaches only part of the way down the cochlea. Consistent with other studies (Dorman et al. 1997; Shannon et al. 1998), Rosen et al. found a sharp decline in intelligibility of frequency-shifted speech presented to listeners with normal hearing. However, over the course of a 3-hour training period, performance improved significantly, indicating that some form of perceptual reorganization had taken place. Their findings suggest that (1) a coarse temporal representation may, under some circumstances, provide sufficient cues for understanding speech with little or no need for training; and (2) a period of perceptual adjustment may be needed when the bands are shifted from their expected locations along the tonotopic array.

4.5 Use of Linguistic Context

The successful recovery of information from distorted speech depends on properties of the signal. Nonuniformities in the distribution of energy across

time and frequency enable listeners to glimpse the target voice, while regularities in time and frequency allow for the operation of perceptual grouping principles. Intelligibility is also determined by the ability of the listener to exploit various aspects of linguistic and pragmatic context, especially when the signal is degraded (Treisman 1960, 1964; Warren 1996). For example, word recognition performance in background noise is strongly affected by such factors as the size of the response set (Miller et al. 1951), lexical status, familiarity of the stimulus materials and word frequency (Howes 1957; Pollack et al. 1959; Auer and Bernstein 1997), and lexical neighborhood similarity (Luce et al. 1990; Luce and Pisoni 1998).

Miller (1947) reported that conversational babble in an unfamiliar language was neither more nor less interfering than babble in the native language of the listeners (English). He concluded that the spectrum of a masking signal is the crucial factor, while the linguistic content is of secondary importance. A different conclusion was reached by Treisman (1964), who used a shadowing task to show that the linguistic content of an interfering message was an important determinant of its capacity to interfere with the processing of a target message. Most disruptive was a competing message in the same language and similar in content, followed by a foreign language familiar to the listeners, followed by reversed speech in the native language, followed by an unfamiliar foreign language. Differences in task demands (the use of speech babble or a single competing voice masker), the amount of training, as well as instructions to the subjects may underlie the difference between Triesman's and Miller's results. The importance of native-language experience was demonstrated by Gat and Keith (1978) and Mayo et al. (1997). They found that native English listeners could understand monosyllabic words or sentences of American English at lower SNRs than could nonnative students who spoke English as a second language. In addition, Mayo et al. found greater benefits of linguistic context for native speakers of English and for those who learned English as a second language before the age of 6, compared to bilinguals who learned English as a second language in adulthood. Other studies have confirmed that word recognition by nonnative listeners can be severely reduced in conditions where fine phonetic discrimination is required and background noise is present (Bradlow and Pisoni 1999).

When words are presented in sentences, the presence of semantic context restricts the range of plausible possibilities. This leads to higher intelligibility and greater resistance to distortion (Kalikow et al. 1977; Boothroyd and Nittrouer 1988; Elliot 1995). The SPIN test (Kalikow et al. 1977) provides a clinical measure of the ability of a listener to take advantage of context to identify the final keyword in sentences, which are either of low or high predictability.

Boothroyd and Nittrouer (1988) presented a model that assumes that the effects of context are equivalent to providing additional, statistically independent channels of sensory information. First, they showed that the prob-

ability of correct recognition of speech units (phones or words) in context (p_c) could be predicted from their identification without context (p_i) from the following relationship:

$$p_c = 1 - (1 - p_i)^k \tag{3}$$

The factor k is a constant that measures the use of contextual information. It is computed from the ratio of the logarithms of the error probabilities:

$$k = \frac{\log(1 - p_c)}{\log(1 - p_i)} \tag{4}$$

Boothroyd and Nittrouer extended this model to show that the recognition of complex speech units (e.g., words) could be predicted from the identification of their component parts (phones). Their model was based on earlier work by Fletcher (1953) showing that the probability of correct identification of individual consonants and vowels within CVC nonsense syllables could be accurately predicted by assuming that the recognition of the whole depends on prior recognition of the component parts, and that the probabilities of recognizing the parts are statistically independent. According to this model, the probability of recognizing the whole (p_w) depends on the probability of identifying the component parts (p_p):

$$p_w = p_p^j \tag{5}$$

where $1 < j < n$, and n is the number of parts. The factor j is computed from 36
the ratio of the logarithms of the recognition probabilities:

$$j = \frac{\log(1 - p_w)}{\log(1 - p_p)} \tag{6}$$

The value of j ranges between 1 (in situations where context plays a large role) and n (where context has no effect on recognition). For nonsense syllables and nonmeaningful sentences, the value of j is assumed to be equal to n, the number of component parts.

Boothroyd and Nittrouer applied these models to predict context effects in CVC syllables and in sentences. They included high-predictability and low-predictability sentences differing in the degree of semantic context, as well as zero-predictability sentences in which the words were presented randomly so that neither semantic nor syntactic context was available. They found values of k ranging between 1.3 for CVCs and 2.7 for high-predictability sentences and values of j ranging from 2.5 in nonsense CVC syllables to 1.6 in four-word, high-predictability sentences. The derived j and k factors were constant across a range of probabilities, supporting the assumption that these measures provide good quantitative measures of the effects of linguistic context.

Another modeling approach was used by Rooij and Plomp (1991), who characterized the effects of linguistic context on sentence perception in

terms of linguistic entropy, a measure derived from information theory (Shannon and Weaver 1949). The entropy, H, of an information source (in bits) represents the degree of uncertainty in receiving a given item from a vocabulary of potential elements, and is defined as

$$H = -\sum_{i=1}^{n} \log p_i \tag{7}$$

where p_i is the probability of selecting item i from a set of N independent items. The entropy increases as a function of the number of items in the set and is dependent on the relative probabilities of the individual items in the set. The degree of linguistic redundancy is inversely proportional to its entropy. Rooij and Plomp estimated the linguistic entropy of a set of sentences (originally chosen to be as similar as possible in overall redundancy) by means of a visual letter-guessing procedure proposed by Shannon (1951). They estimated the entropy in bits per character (for individual letters in sentences) from the probability of correct guesses made by subjects who were given successive fragments of each sentence, presented one letter at a time. After each guess the subject was told the identity of the current letter and all those that preceded it. Rooij and Plomp showed that estimates of the linguistic entropy of a set of sentences (presented auditorily) could predict the susceptibility of the sentences to masking by speech-shaped noise. Differences in linguistic entropy had an effect of about 4 dB on the SRT and followed a linear relationship for listeners with normal hearing. Despite the limitations of this approach (e.g., the assumption that individual letters are equi-probable, and the use of an orthographic measure of linguistic entropy, rather than one based on phonological, morphological, or lexical units), this study illustrates the importance of linguistic factors in accounting for speech perception abilities in noise. The model has been extended to predict speech recognition in noise for native and nonnative listeners (van Wijngaarden et al. 2002).

Listeners exploit their knowledge of linguistic constraints to restrict the potential interpretations that can be placed on the acoustic signal. The process involves the active generation of possible interpretations, combined with a method for filtering or restricting lexical candidates (Klatt 1989; Marslen-Wilson 1989). When speech is perceived under adverse conditions, the process of restricting the set of possible interpretations requires a measure of quality or "goodness of fit" between the candidate and its acoustical support. The process of evaluating and assessing the reliability of incoming acoustic properties depends both on the signal properties (including some measure of distortion) and the strength of the linguistic hypotheses that are evoked. If the acoustic evidence is weak, then the linguistic hypotheses play a stronger role. If the signal provides clear and unambiguous evidence for a given phonetic sequence, then linguistic plausibility makes little or no contribution (Warren et al. 1972). One challenge for

future research is to describe how this attentional switching is achieved on-line by the central nervous system.

5. Summary

The overall conclusion from this review is that the information in speech is shielded from distortion in several ways. First, peaks in the envelope of the spectrum provide robust cues for the identification of vowels and consonants, even when the spectral valleys are obscured by noise. Second, periodicity in the waveform reflects the fundamental frequency of voicing, allowing listeners to group together components that stem from the same voice across frequency and time in order to segregate them from competing signals (Brokx and Nooteboom 1982; Bird and Darwin 1998). Third, at 37
disadvantageous SNRs, the formants of voiced sounds can exert their influence by disrupting the periodicity of competing harmonic signals or by disrupting the interaural correlation of a masking noise (Summerfield and Culling 1992; Culling and Summerfield 1995). Fourth, the amplitude- 38
modulation pattern across frequency bands can serve to highlight informative portions of the speech signal, such as prosodically stressed syllables. These temporal modulation patterns are redundantly specified in time and frequency, making it possible to remove large amounts of the signal via gating in the time domain (e.g., Miller and Licklider 1950) or filtering in the frequency domain (e.g., Warren et al. 1995). Even when the spectral details and periodicity of voiced speech are eliminated, intelligibility remains high if the temporal modulation structure is preserved in a small number of bands (Shannon et al. 1995). However, speech processed in this manner is more susceptible to interference by other signals (Fu et al. 1998).

Competing signals, noise, reverberation, and other imperfections of the communication channel can eliminate, mask, or distort the information-providing segments of the speech signal. Listeners with normal hearing rely on a range of perceptual and linguistic strategies to overcome these effects and bridge the gaps that appear in the time-frequency distribution of the distorted signal. Time-varying changes in the SNR allow listeners to focus their attention on temporal and spectral regions where the target voice is best defined, a process described as glimpsing. Together with complementary processes such as perceptual grouping and tracking, listeners use their knowledge of linguistic contraints to fill in the gaps in the signal and arrive at the most plausible interpretations of the distorted signal.

Glimpsing and tracking depend on an analysis of the signal within a sliding temporal window, and provide effective strategies when the distortion is intermittent. When the form of distortion is relatively stationary (e.g., a continuous, broadband noise masker, or the nonuniform frequency response of a large room), other short-term processes such as adaptation and perceptual grouping can be beneficial. Adaptation serves to emphasize

newly arriving components of the signal, enhancing syllable onsets and regions of the signal undergoing rapid spectrotemporal change. Perceptual grouping processes link together acoustic components that emanate from the same sound source. Listeners may also benefit from central auditory processes that compensate for distortions of the frequency response of the channel. The nature and time course of such adaptations remain topics of current interest and controversy.

List of Abbreviations

ACF	autocorrelation function
AI	articulation index
AM	amplitude modulation
CF	characteristic frequency
CMR	comodulation masking release
f_0	fundamental frequency
F_1	first formant
F_2	second formant
F_3	third formant
HP	high pass
ILD	interaural level difference
ITD	interaural time difference
LP	low pass
LPC	linear predictive coding
LTASS	long-term average speech spectrum
rms	root mean square
SNR	signal-to-noise ratio
SPIN	speech perception in noise test
SRT	speech reception threshold
STI	speech transmission index
TMTF	temporal modulation transfer function
VOT	voice onset time

References

Allen JB (1994) How do humans process and recognize speech? IEEE Trans Speech Audio Proc 2:567–577.

ANSI (1969) Methods for the calculation of the articulation index. ANSI S3.5-1969. New York: American National Standards Institute.

ANSI (1997) Methods for the calculation of the articulation index. ANSI S3.5-1997. New York: American National Standards Institute.

Arai T, Greenberg S (1998) Speech intelligibility in the presence of cross-channel spectral asynchrony, IEEE Int Conf Acoust Speech Signal Proc, pp. 933–936.

Assmann PF (1991) Perception of back vowels: center of gravity hypothesis. Q J Exp Psychol 43A:423–448.

Assmann PF (1995) The role of formant transitions in the perception of concurrent vowels. J Acoust Soc Am 97:575–584.
Assmann PF (1996) Modeling the perception of concurrent vowels: role of formant transitions. J Acoust Soc Am 100:1141–1152.
Assmann PF (1999) Fundamental frequency and the intelligibility of competing voices. Proceedings of the 14th International Congress of Phonetic Sciences, pp. 179–182.
Assmann PF, Katz WF (2000) Time-varying spectral change in the vowels of children and adults. J Acoust Soc Am 108:1856–1866.
Assmann PF, Nearey TM (1986) Perception of front vowels: the role of harmonics in the first formant region. J Acoust Soc Am 81:520–534.
Assmann PF, Summerfield AQ (1989) Modeling the perception of concurrent vowels: vowels with the same fundamental frequency. J Acoust Soc Am 85: 327–338.
Assmann PF, Summerfield AQ (1990) Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J Acoust Soc Am 88: 680–697.
Assmann PF, Summerfield Q (1994) The contribution of waveform interactions to the perception of concurrent vowels. J Acoust Soc Am 95:471–484.
Auer ET Jr, Bernstein LE (1997) Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. J Acoust Soc Am 102:3704–3710.
Baer T, Moore BCJ (1993) Effects of spectral smearing on the intelligibility of sentences in noise. J Acoust Soc Am 94:1229–1241.
Baer T, Moore BCJ (1994) Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech [letter]. J Acoust Soc Am 95: 2277–2280.
Baer T, Moore BCJ, Gatehouse S (1993) Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. J Rehabil Res Dev 30:49–72.
Bakkum MJ, Plomp R, Pols LCW (1993) Objective analysis versus subjective assessment of vowels pronounced by native, non-native, and deaf male speakers of Dutch. J Acoust Soc Am 94:1983–1988.
Bashford JA, Reiner KR, Warren RM (1992) Increasing the intelligibility of speech through multiple phonemic restorations. Percept Psychophys 51:211–217.
Beddor PS, Hawkins S (1990) The influence of spectral prominence on perceived vowel quality. J Acoust Soc Am 87:2684–2704.
Beranek LL (1947) The design of speech communication systems. Proc Inst Radio Engineers 35:880–890.
Bird J, Darwin CJ (1999) Effects of a difference in fundamental frequency in separating two sentences. In: Palmer A, Rees A, Summerfield Q, Meddis R (eds) Psychophysical and physiological advances in hearing. London: Whurr.
Bladon RAW (1982) Arguments against formants in the auditory representation of speech In: Carlson R, Granstrom B (eds) The Representation of Speech in the Peripheral Auditory System. Elsevier Biomedical Press.
Blauert J (1996) Spatial Hearing: The Psychophysics of Human Sound Localization, 2nd ed. Cambridge, MA: MIT Press.
Blesser B (1972) Speech perception under conditions of spectral transformation. I. Phonetic characteristics. J Speech Hear Res 15:5–41.

Boothroyd A, Nittrouer S (1988) Mathematical treatment of context effects in phoneme and word recognition. J Acoust Soc Am 84:101–114.
Bradlow AR, Pisoni DB (1999). Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. J Acoust Soc Am 106:2074–2085.
Bregman AS (1990) Auditory Scene Analysis. Cambridge, MA: MIT Press.
Broadbent DE (1958) Perception and Communication. Oxford: Pergamon Press.
Brokx JPL, Nooteboom SG (1982) Intonation and the perception of simultaneous voices. J Phonetics 10:23–26.
Bronkhorst AW (2000) The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acustica 86:117–128.
Bronkhorst AW, Plomp R (1988) The effect of head-induced interaural time and level differences on speech intelligibility in noise. J Acoust Soc Am 83:1508–1516.
Brungart DS (2001a) Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am 109:1101–1109.
Brungart DS (2001b) Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am 110:2527–2538.
Buuren RA van, Festen JM, Houtgast T (1996) Peaks in the frequency response of hearing aids: evaluation of the effects on speech intelligibility and sound quality. J Speech Hear Res 39:239–250.
Buus S (1985) Release from masking caused by envelope fluctuations. J Acoust Soc Am 78:1958–1965.
Byrne D, Dillon H, Tran K, et al. (1994) An international comparison of long-term average speech spectra. J Acoust Soc Am 96:2108–2120.
Carhart R (1965) Monaural and binaural discrimination against competing sentences. Int Audiol 4:5–10.
Carhart R, Tillman TW, Greetis ES (1969) Perceptual masking in multiple sound background. J Acoust Soc Am 45:411–418.
Cariani PA, Delgutte B (1996a) Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. J Neurophys 76:1698–1716.
Cariani PA, Delgutte B (1996b) Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. J Neurophys 76:1717–1734.
Carlson R, Fant G, Granstrom B (1974) Two-formant models, pitch, and vowel perception. Acustica 31:360–362.
Carlson R, Granstrom B, Klatt D (1979) Vowel perception: the relative perceptual salience of selected acoustic manipulations. Speech Transmission Laboratories (Stockholm) Quarterly Progress Report SR 3–4, pp. 73–83.
Carlyon RP (1989) Changes in the masked thresholds of brief tones produced by prior bursts of noise. Hear Res 41:223–236.
Carlyon RP (1994) Further evidence against an across-frequency mechanism specific to the detection of FM incoherence between resolved frequency components. J Acoust Soc Am 95:949–961.
Carney LH, Yin TCT (1988) Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. J Neurophys 60:1653–1677.
Carrell TD, Opie JM (1992) The effect of amplitude comodulation on auditory object formation in sentence perception. Percept Psychophys 52:437–445.

Carterette EC, Møller A (1962) The perception of real and synthetic vowels after very sharp filtering. Speech Transmission Laboratories (Stockholm) Quarterly Progress Report SR 3, pp. 30–35.

Castle WE (1964) The Effect of Narrow Band Filtering on the Perception of Certain English Vowels. The Hague: Mouton.

Chalikia M, Bregman A (1989) The perceptual segregation of simultaneous auditory signals: pulse train segregation and vowel segregation. Percept Psychophys 46:487–496.

Cherry C (1953) Some experiments on the recognition of speech, with one and two ears. J Acoust Soc Am 25:975–979.

Cherry C, Wiley R (1967) Speech communication in very noisy environments. Nature 214:1164.

Cheveigné A de (1997) Concurrent vowel identification. III: A neural model of harmonic interference cancellation. J Acoust Soc Am 101:2857–2865.

Cheveigné A de, McAdams S, Laroche J, Rosenberg M (1995) Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement. J Acoust Soc Am 97:3736–3748.

Chistovich LA (1984) Central auditory processing of peripheral vowel spectra. J Acoust Soc Am 77:789–805.

Chistovich LA, Lublinskaya VV (1979) The "center of gravity" effect in vowel spectra and critical distance between the formants: psychoacoustic study of the perception of vowel-like stimuli. Hear Res 1:185–195.

Ciocca V, Bregman AS (1987) Perceived continuity of gliding and steady-state tones through interrupting noise. Percept Psychophys 42:476–484.

Coker CH, Umeda N (1974) Speech as an error correcting process. Speech Communication Seminar, SCS-74, Stockholm, Aug. 1–3, pp. 349–364.

Cooke MP, Ellis DPW (2001) The auditory organization of speech and other sources in listeners and computational models. Speech Commun 35:141–177.

Cooke MP, Morris A, Green PD (1996) Recognising occluded speech. In: Greenberg S, Ainsworth WA (eds) Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, pp. 297–300.

Culling JE, Darwin CJ (1993a) Perceptual separation of simultaneous vowels: within and across-formant grouping by f_0. J Acoust Soc Am 93:3454–3467.

Culling JE, Darwin CJ (1993b) The role of timbre in the segregation of simultaneous voices with intersecting f_0 contours. Percept Psychophys 34:303–309.

Culling JE, Darwin CJ (1994) Perceptual and computational separation of simultaneous vowels: cues arising from low frequency beating. J Acoust Soc Am 95: 1559–1569.

Culling JF, Summerfield Q (1995a) Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. J Acoust Soc Am 98:785–797.

Culling JF, Summerfield Q (1995b) The role of frequency modulation in the perceptual segregation of concurrent vowels. J Acoust Soc Am 98:837–846.

Culling JF, Summerfield Q, Marshall DH (1994) Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. Speech Commun 14:71–95.

Darwin CJ (1984) Perceiving vowels in the presence of another sound: constraints on formant perception. J Acoust Soc Am 76:1636–1647.

Darwin CJ (1990) Environmental influences on speech perception. In: Advances in Speech, Hearing and Language Processing, vol. 1. London: JAI Press, pp. 219–241.

Darwin CJ (1992) Listening to two things at once. In: Schouten MEH (ed) The Auditory Processing of Speech: From Sounds to Words. Berlin: Mouton de Gruyter, pp. 133–147.

Darwin CJ, Carlyon RP (1995) Auditory Grouping. In: Moore BCJ (ed) The Handbook of Perception and Cognition, vol. 6, Hearing. London: Academic Press.

Darwin CJ, Gardner RB (1986) Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. J Acoust Soc Am 79:838–845.

Darwin CJ, Hukin RW (1997a) Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. J Acoust Soc Am 102: 2316–2324.

Darwin CJ, Hukin RW (1997b) Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. J Acoust Soc Am 103:1080–1084.

Darwin CJ, McKeown JD, Kirby D (1989) Compensation for transmission channel and speaker effects on vowel quality. Speech Commun 8:221–234.

Delattre P, 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 8:195–201.

Delgutte B (1980) Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J Acoust Soc Am 68:843–857.

Delgutte B (1996) Auditory neural processing of speech. In: Hardcastle WJ, Laver J (eds) The Handbook of Phonetic Sciences. Oxford: Blackwell.

Delgutte B, Kiang NYS (1984) Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. J Acoust Soc Am 75:897–907.

Deng L, Kheirallah I (1993) Dynamic formant tracking of noisy speech using temporal analysis on outputs from a nonlinear cochlear model. IEEE Trans Biomed Eng 40:456–467.

Dirks DD, Bower DR (1969) Masking effects of speech competing messages. J Speech Hear Res 12:229–245.

Dirks DD, Wilson RH (1969) The effect of spatially separated sound sources on speech intelligibility. J Speech Hear Res 12:5–38.

Dirks DD, Wilson RH, Bower DR (1969) Effects of pulsed masking on selected speech materials. J Acoust Soc Am 46:898–906.

Dissard P, Darwin CJ (2000) Extracting spectral envelopes: formant frequency matching between sounds on different and modulated fundamental frequencies. J Acoust Soc Am 107:960–969.

Dorman MF, Loizou PC, Rainey D (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise outputs. J Acoust Soc Am 102:2403–2411.

Dorman MF, Loizou PC, Fitzke J, Tu Z (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. J Acoust Soc Am 104:3583–3585.

Dreher JJ, O'Neill JJ (1957) Effects of ambient noise on speaker intelligibility for words and phrases. J Acoust Soc Am 29:1320–1323.

Drullman R (1995a) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97:585–592.

Drullman R (1995b) Speech intelligibility in noise: relative contribution of speech elements above and below the noise level. J Acoust Soc Am 98:1796–1798.
Drullman R, Festen JM, Plomp R (1994a) Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95:2670–2680.
Drullman R, Festen JM, Plomp R (1994b) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95:1053–1064.
Dubno J, Ahlstrom JB (1995) Growth of low-pass masking of pure tones and speech for hearing-impaired and normal-hearing listeners. J Acoust Soc Am 98:3113–3124.
Duifhuis H, Willems LF, Sluyter RJ (1982) Measurement of pitch on speech: an implementation of Goldstein's theory of pitch perception. J Acoust Soc Am 71:1568–1580.
Dunn HK, White SD (1940) Statistical measurements on conversational speech. J Acoust Soc Am 11:278–288.
Duquesnoy AJ (1983) Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. J Acoust Soc Am 74:739–743.
Duquesnoy AJ, Plomp R (1983) The effect of a hearing aid on the speech-reception threshold of a hearing-impaired listener in quiet and in noise. J Acoust Soc Am 73:2166–2173.
Egan JP, Wiener FM (1946) On the intelligibility of bands of speech in noise. J Acoust Soc Am 18:435–441.
Egan JP, Carterette EC, Thwing EJ (1954) Some factors affecting multi-channel listening. J Acoust Soc Am 26:774–782.
Elliot LL (1995) Verbal auditory closure and the Speech Perception in Noise (SPIN) test. J Speech Hear Res 38:1363–1376.
Fahey RP, Diehl RL, Traunmuller H (1996) Perception of back vowels: effects of varying F_1–f_0 Bark distance. J Acoust Soc Am 99:2350–2357.
Fant G (1960) Acoustic Theory of Speech Production. Mouton: The Hague.
Festen JM (1993) Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice. J Acoust Soc Am 94:1295–1300.
Festen JM, Plomp R (1981) Relations between auditory functions in normal hearing. J Acoust Soc Am 70:356–369.
Festen JM, Plomp R (1990) Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J Acoust Soc Am 88:1725–1736.
Finitzo-Hieber T, Tillman TW (1978) Room acoustics effects on monosyllabic word discrimination ability for normal and hearing impaired children. J Speech Hear Res 21:440–458.
Fletcher H (1952) The perception of sounds by deafened persons. J Acoust Soc Am 24:490–497.
Fletcher H (1953) Speech and Hearing in Communication. New York: Van Nostrand (reprinted by the Acoustical Society of America, 1995).
Fletcher H, Galt RH (1950) The perception of speech and its relation to telephony. J Acoust Soc Am 22:89–151.
French NR, Steinberg JC (1947) Factors governing the intelligibility of speech sounds. J Acoust Soc Am 19:90–119.
Fu Q-J, Shannon RV, Wang X (1998) Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. J Acoust Soc Am 104:3586–3596.

Gardner RB, Gaskill SA, Darwin CJ (1989) Perceptual grouping of formants with static and dynamic differences in fundamental frequency. J Acoust Soc Am 85:1329–1337.

Gat IB, Keith RW (1978) An effect of linguistic experience. Auditory word discrimination by native and non-native speakers of English. Audiology 17:339–345.

Gatehouse S (1992) The time course and magnitude of perceptual acclimitization to frequency responses: evidence from monaural fitting of hearing aids. J Acoust Soc Am 92:1258–1268.

Gatehouse S (1993) Role of perceptual acclimitization to frequency responses: evidence from monaural fitting of hearing aids. J Am Acad Audiol 4:296–306.

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

Glasberg BR, Moore BCJ (1986) Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. J Acoust Soc Am 79:1020–1033.

Gong Y (1994) Speech recognition in noisy environments: a survey. Speech Commun 16:261–291.

42 Gordon-Salant S, Fitzgibbons (1995) Recognition of multiply degraded speech by young and elderly listeners. J Speech Hear Res 38:1150–1156.

Grant KW, Ardell LH, Kuhl PK, Sparks DW (1985) The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. J Acoust Soc Am 77:671–677.

Grant KW, Braida LD, Renn RJ (1991) Single band amplitude envelope cues as an aid to speechreading. Q J Exp Psychol 43A:621–645.

Grant KW, Braida LD, Renn RJ (1994) Auditory supplements to speechreading: combining amplitude envelope cues from different spectral regions of speech. J Acoust Soc Am 95:1065–1073.

Greenberg S (1995) Auditory processing of speech. In: Lass NJ (ed) Principles of Experimental Phonetics. St. Louis: Mosby-Year Book, pp. 362–407.

Greenberg S (1996) Understanding speech understanding: Towards a unified theory of speech perception. In: Greenberg S, Ainsworth WA (eds) Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, pp. 1–8.

Greenberg S, Arai T (1998) Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. Proceedings of the Joint Meeting of the Acoustical Society of America and the International Congress on Acoustics, pp. 2677–2678.

Greenberg S, Arai T, Silipo R (1998) Speech intelligibility derived from exceedingly sparse spectral information. Proceedings of the International Conference on Spoken Language Processing, Sydney, pp. 74–77.

Gustafsson HA, Arlinger SD (1994) Masking of speech by amplitude-modulated noise. J Acoust Soc Am 95:518–529.

Haggard MP (1985) Temporal patterning in speech: the implications of temporal resolution and signal processing. In: Michelson A (ed) Time Resolution in Auditory Systems. Berlin: Springer-Verlag, pp. 217–237.

Hall JW, Grose JH (1991) Relative contributions of envelope maxima and minima to comodulation masking release. Q J Exp Psychol 43A:349–372.

Hall JW, Haggard MP, Fernandez MA (1984) Detection in noise by spectro-temporal analysis. J Acoust Soc Am 76:50–56.

Hanson BA, Applebaum TH (1990) Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech. Proc Int Conf Acoust Speech Signal Processing 90:857–860.

Hartmann WM (1996) Pitch, periodicity, and auditory organization. J Acoust Soc Am 100:3491–3502.
Hawkins JE Jr, Stevens SS (1950) The masking of pure tones and of speech by white noise. J Acoust Soc Am 22:6–13.
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.
Hicks ML, Bacon SP (1992) Factors influencing temporal effects with notched-noise maskers. Hear Res 64:123–132.
Hillenbrand JM, Nearey TM (1999) Identification of resynthesized /hVd/ utterances: effects of formant contour. J Acoust Soc Am 105:3509–3523.
Hillenbrand JM, Getty LA, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. J Acoust Soc Am 97:3099–3111.
Hockett CF (1955) A Manual of Phonology. Bloomington, IN: Indiana University Press.
Horwitz AR, Turner CW (1997) The time course of hearing aid benefit. Ear Hear 18:1–11.
Houtgast T (1972) Psychophysical evidence for lateral inhibition in hearing. J Acoust Soc Am 51:1885–1894.
Houtgast T, Steeneken HJM (1973) The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acustica 28:66–73.
Houtgast T, Steeneken HJM (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am 77:1069–1077.
Howard-Jones PA, Rosen S (1993) Uncomodulated glimpsing in "checkerboard" noise. J Acoust Soc Am 93:2915–2922.
Howes D (1957) On the relation between the intelligibility and frequency of occurrence of English words. J Acoust Soc Am 29:296–303.
Huggins AWF (1975) Temporally segmented speech. Percept Psychophys 18:149–157.
Hukin RW, Darwin CJ (1995) Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. Percept Psychophys 57:191–196.
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.
Humes LE, Boney S, Loven F (1987) Further validation of the speech transmission index (STI). J Speech Hear Res 30:403–410.
Hygge S, Rönnberg J, Larsby B, Arlinger S (1992) Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. J Speech Hear Res 35:208–215.
Irwin RJ, McAuley SF (1987) Relations among temporal acuity, hearing loss, and the perception of speech distorted by noise and reverberation. J Acoust Soc Am 81:1557–1565.
Joris PX, Yin TC (1995) Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences. J Neurophys 73:1043–1062.
Junqua JC, Anglade Y (1990) Acoustic and perceptual studies of Lombard speech: application to isolated words automatic speech recognition. Proc Int Conf Acoust Speech Signal Processing 90:841–844.

Kalikow DN, Stevens KN, Elliot LL (1977) Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J Acoust Soc Am 61:1337–1351.

Kates JM (1987) The short-time articulation index. J Rehabil Res Dev 24:271–276.

Keurs M ter, Festen JM, Plomp R (1992) Effect of spectral envelope smearing on speech reception. I. J Acoust Soc Am 91:2872–2880.

Keurs M ter, Festen JM, Plomp R (1993a) Effect of spectral envelope smearing on speech reception. II. J Acoust Soc Am 93:1547–1552.

Keurs M ter, Festen JM, Plomp R (1993b) Limited resolution of spectral contrast and hearing loss for speech in noise. J Acoust Soc Am 94:1307–1314.

Kewley-Port D, Zheng Y (1998) Auditory models of formant frequency discrimination for isolated vowels. J Acoust Soc Am 103:1654–1666

Klatt DH (1982) Speech processing strategies based on auditory models. In: Carlson R, Granstrom B (eds) The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier.

Klatt DH (1989) Review of selected models of speech perception. In: Marslen-Wilson W (ed) Lexical Representation and Process. Cambridge, MA: MIT Press, pp.169–226.

Kluender KR, Jenison RL (1992) Effects of glide slope, noise intensity, and noise duration in the extrapolation of FM glides through noise. Percept Psychophys 51:231–238.

Kreiman J (1997) Listening to voices: theory and practice in voice perception research. In: Johnson K, Mullenix J (eds) Talker Variability in Speech Processing. San Diego: Academic Press.

Kryter KD (1946) Effects of ear protective devices on the intelligibility of speech in noise. J Acoust Soc Am 18:413–417.

Kryter KD (1962) Methods for the calculation and use of the articulation index. J Acoust Soc Am 34:1689–1697.

Kryter D (1985) The Effects of Noise on Man, 2nd ed. London: Academic Press.

Kuhn GF (1977) Model for the interaural time differences in the azimuthal plane. J Acoust Soc Am 62:157–167.

Ladefoged P (1967) Three Areas of Experimental Phonetics. Oxford: Oxford University Press, pp. 162–165.

Lane H (1967) Psychophysical parameters of vowel perception. Psychol Monogr 76(44).

Lane H, Tranel B (1971) The Lombard sign and the role of hearing in speech. J Speech Hear Res 14:677–709.

Langner G (1992) Periodicity coding in the auditory system. Hear Res 60:115–142.

Lea AP (1992) Auditory modeling of vowel perception. PhD thesis, University of Nottingham.

Lea AP, Summerfield Q (1994) Minimal spectral contrast of formant peaks for vowel recognition as a function of spectral slope. Percept Psychophys 56:379–391.

Leek MR, Dorman MF, Summerfield, Q (1987) Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners. J Acoust Soc Am 81:148–154.

Lehiste I, Peterson GE (1959) The identification of filtered vowels. Phonetica 4:161–177.

Levitt H, Rabiner LR (1967) Predicting binaural gain in intelligibility and release from masking for speech. J Acoust Soc Am 42:820–829.

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

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461.

Licklider JCR, Guttman N (1957) Masking of speech by line-spectrum interference. J Acoust Soc Am 29:287–296.

Licklider JCR, Miller GA (1951) The perception of speech. In: Stevens SS (ed) Handbook of Experimental Psychology. New York: John Wiley, pp. 1040–1074.

Lindblom B (1990) Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle WJ, Marshall A (eds) Speech Production and Speech Modelling. Dordrecht: Kluwer Academic, pp. 403–439.

Lippmann R (1996a) Speech perception by humans and machines. In: Greenberg S, Ainsworth WA (eds) Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception. pp. 309–316.

Lippmann R (1996b) Accurate consonant perception without mid-frequency speech energy. IEEE Trans Speech Audio Proc 4:66–69.

Liu SA (1996) Landmark detection for distinctive feature-based speech recognition. J Acoust Soc Am 100:3417–3426.

Lively SE, Pisoni DB, Van Summers W, Bernacki RH (1993) Effects of cognitive workload on speech production: acoustic analyses and perceptual consequences. J Acoust Soc Am 93:2962–2973.

Lombard E (1911) Le signe de l'élévation de la voix. Ann Malad l'Oreille Larynx Nez Pharynx 37:101–119.

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

Luce PA, Pisoni DB, Goldinger SD (1990) Similarity neighborhoods of spoken words. In: Altmann GTM (ed) Cognitive Models of Speech Processing. Cambridge: MIT Press, pp. 122–147.

Ludvigsen C (1987) Prediction of speech intelligibility for normal-hearing and cochlearly hearing impaired listeners. J Acoust Soc Am 82:1162–1171.

Ludvigsen C, Elberling C, Keidser G, Poulsen T (1990) Prediction of intelligibility for nonlinearly processed speech. Acta Otolaryngol Suppl 469:190–195.

MacLeod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise. Br J Audiol 21:131–141.

Marin CMH, McAdams SE (1991) Segregation of concurrent sounds. II: Effects of spectral-envelope tracing, frequency modulation coherence and frequency modulation width. J Acoust Soc Am 89:341–351.

Markel JD, Gray AH (1976) Linear Prediction of Speech. New York: Springer-Verlag.

Marslen-Wilson W (1989) Access and integration: projecting sound onto meaning. In: Marslen-Wilson W (ed) Lexical Representation and Process. Cambridge: MIT Press, pp. 3–24.

Mayo LH, Florentine M, Buus S (1997) Age of second-language acquisition and perception of speech in noise. J Speech Lang Hear Res 40:686–693.

McAdams SE (1989) Segregation of concurrent sounds: effects of frequency-modulation coherence and a fixed resonance structure. J Acoust Soc Am 85:2148–2159.

McKay CM, Vandali AE, McDermott HJ, Clark GM (1994) Speech processing for multichannel cochlear implants: variations of the Spectral Maxima Sound Processor strategy. Acta Otolaryngol 114:52–58.

Meddis R, Hewitt M (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. J Acoust Soc Am 89: 2866–2882.

Meddis R, Hewitt M (1992) Modelling the identification of concurrent vowels with different fundamental frequencies. J Acoust Soc Am 91:233–245.

Miller GA (1947) The masking of speech. Psychol Bull 44:105–129.

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

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. J Acoust Soc Am 27:338–352.

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

Moncur JP, Dirks D (1967) Binaural and monaural speech intelligibility in reverberation. J Speech Hear Res 10:186–195.

Moore BCJ (1995) Perceptual Consequences of Cochlear Hearing Impairment. London: Academic Press.

Moore BCJ, Glasberg BR (1983a) Masking patterns for synthetic vowels in simultaneous and forward masking. J Acoust Soc Am 73:906–917.

Moore BCJ, Glasberg BR (1983b) Suggested formulae for calculating auditory-filter shapes and excitation patterns. J Acoust Soc Am 74:750–753.

Moore BCJ, Glasberg BR (1987) Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. Hear Res 28:209–225.

Moore BCJ, Glasberg BR, Peters RW (1985) Relative dominance of individual partials in determining the pitch of complex tones. J Acoust Soc Am 77:1861–1867.

Moore BCJ, Glasberg BR, Simpson, AM (1992) Evaluation of a method of simulating reduced frequency selectivity. J Acoust Soc Am 91:3402–3423.

Müsch H, Buus S (2001a). Using statistical decision theory to predict speech intelligibility. I. Model structure. J Acoust Soc Am 109:2896–2909.

Müsch H, Buus S (2001b). Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance. J Acoust Soc Am 109:2910–2920.

Nábelek AK (1988) Identification of vowels in quiet, noise, and reverberation: relationships with age and hearing loss. J Acoust Soc Am 84:476–484.

Nábelek AK, Dagenais PA (1986) Vowel errors in noise and in reverberation by hearing-impaired listeners. J Acoust Soc Am 80:741–748.

Nábelek AK, Letowski TR (1988) Similarities of vowels in nonreverberant and reverberant fields. J Acoust Soc Am 83:1891–1899.

Nábelek AK, Pickett JM (1974) Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners. J Speech Hear Res 17:724–739.

Nábelek AK, Robinson PK (1982) Monaural and binaural speech perception in reverberation in listeners of various ages. J Acoust Soc Am 71:1242–1248.

Nábelek AK, Letwoski TR, Tucker FM (1989) Reverberant overlap- and self-masking in consonant identification. J Acoust Soc Am 86:1259–1265.

Nábelek AK, Czyzewski Z, Crowley H (1994) Cues for perception of the diphthong [ai] in either noise or reverberation: I. Duration of the transition. J Acoust Soc Am 95:2681–2693.
Nearey TM (1989) Static, dynamic, and relational properties in vowel perception. J Acoust Soc Am 85:2088–2113.
Neuman AC, Hochberg I (1983) Children's perception of speech in reverberation. J Acoust Soc Am 73:2145–2149.
Nocerino N, Soong FK, Rabiner LR, Klatt DH (1985) Comparative study of several distortion measures for speech recognition. Speech Commun 4:317–331.
Noordhoek IM, Drullman R (1997) Effect of reducing temporal intensity modulations on sentence intelligibility. J Acoust Soc Am 101:498–502.
Nooteboom SG (1968) Perceptual confusions among Dutch vowels presented in noise. IPO Ann Prog Rep 3:68–71.
Palmer AR (1995) Neural signal processing. In: Moore BCJ (ed) The Handbook of Perception and Cognition, vol. 6, Hearing. London: Academic Press.
Palmer AR, Summerfield Q, Fantini DA (1995) Responses of auditory-nerve fibers to stimuli producing psychophysical enhancement. J Acoust Soc Am 97: 1786–1799.
Patterson RD, Moore BCJ (1986) Auditory filters and excitation patterns as representations of auditory frequency selectivity. In: Moore BCJ (ed) Frequency Selectivity in Hearing. London: Academic Press.
Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand MH (1992) Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K (eds) Auditory Physiology and Perception. Oxford: Pergamon Press, pp. 429–446.
Pavlovic CV (1987) Derivation of primary parameters and procedures for use in speech intelligibility predictions. J Acoust Soc Am 82:413–422.
Pavlovic CV, Studebaker GA (1984) An evaluation of some assumptions underlying the articulation index. J Acoust Soc Am 75:1606–1612.
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 95:1581–1592.
Peters RW, Moore BCJ, Baer T (1998) Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. J Acoust Soc Am 103:577–587.
Peterson GE, Barney HL (1952) Control methods used in a study of vowels. J Acoust Soc Am 24:175–184.
Picheny M, Durlach N, Braida L (1985) Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. J Speech Hear Res 28:96–103.
Picheny M, Durlach N, Braida L (1986) Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. J Speech Hear Res 29:434–446.
Pickett JM (1956) Effects of vocal force on the intelligibility of speech sounds. J Acoust Soc Am 28:902–905.
Pickett JM (1957) Perception of vowels heard in noises of various spectra. J Acoust Soc Am 29:613–620.

Pisoni DB, Bernacki RH, Nusbaum HC, Yuchtman M (1985) Some acoustic-phonetic correlates of speech produced in noise. Proc Int Conf Acoust Speech Signal Proc, pp. 1581–1584.

Plomp R (1976) Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise). Acustica 24:200–211.

Plomp R (1983) The role of modulation in hearing. In: Klinke R (ed) Hearing: Physiological Bases and Psychophysics. Heidelberg: Springer-Verlag, pp. 270–275.

Plomp R, Mimpen AM (1979) Improving the reliability of testing the speech reception threshold for sentences. Audiology 18:43–52.

Plomp R, Mimpen AM (1981) Effect of the orientation of the speaker's head and the azimuth of a sound source on the speech reception threshold for sentences. Acustica 48:325–328.

Plomp R, Steeneken HJM (1978) Place dependence of timbre in reverberant sound fields. Acustica 28:50–59.

Pollack I, Pickett JM (1958) Masking of speech by noise at high sound levels. J Acoust Soc Am 30:127–130.

Pollack I, Rubenstein H, Decker L (1959) Intelligibility of known and unknown message sets. J Acoust Soc Am 31:273–279.

Pols L, Kamp L van der, Plomp R (1969) Perceptual and physical space of vowel sounds. J Acoust Soc Am 46:458–467.

Powers GL, Wilcox JC (1977) Intelligibility of temporally interrupted speech with and without intervening noise. J Acoust Soc Am 61:195–199.

Rankovic CM (1995) An application of the articulation index to hearing aid fitting. J Speech Hear Res 34:391–402.

Rankovic CM (1998) Factors governing speech reception benefits of adaptive linear filtering for listeners with sensorineural hearing loss. J Acoust Soc Am 103: 1043–1057.

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

Roberts B Moore BCJ (1990) The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels. J Acoust Soc Am 88:2571–2583.

Roberts B, Moore BCJ (1991a) The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony. J Acoust Soc Am 89:2922–2932.

Roberts B, Moore BCJ (1991b) Modeling the effects of extraneous sounds on the perceptual estimation of first-formant frequency in vowels. J Acoust Soc Am 89:2933–2951.

Rooij JC van, Plomp R (1991) The effect of linguistic entropy on speech perception in noise in young and elderly listeners. J Acoust Soc Am 90:2985–2991.

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. In: Carlyon RP, Darwin CJ, Russell IJ (eds) Processing of Complex Sounds by the Auditory System. Oxford: Oxford University Press, pp. 73–80.

Rosen S, Faulkner A, Wilkinson L (1998) Perceptual adaptation by normal listeners to upward shifts of spectral information in speech and its relevance for users of cochlear implants. Abstracts of the 1998 Midwinter Meeting of the Association for Research in Otolaryngology.

Rosner BS, Pickering JB (1994) Vowel Perception and Production. Oxford: Oxford University Press.

Rostolland D (1982) Acoustic features of shouted voice. Acustica 50:118–125.
Rostolland D (1985) Intelligibility of shouted voice. Acustica 57:103–121.
Scheffers MTM (1983) Sifting Vowels: Auditory Pitch Analysis and Sound Segregation. PhD thesis, Rijksuniversiteit te Groningen, The Netherlands.
Shannon CE (1951) Prediction and entropy of printed English. Bell Sys Tech J 30:50–64.
Shannon CE, Weaver W (1949) A Mathematical Theory of Communication. Urbana, IL: University of Illinois Press.
Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304.
Shannon RV, Zeng F-G, Wygonski J (1998). Speech recognition with altered spectral distribution of envelope cues. J Acoust Soc Am 104:2467–2476.
Simpson AM, Moore BCJ, Glasberg BR (1990) Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners. Acta Otolaryngol Suppl 469:101–107.
Skinner MW, Clark GM, Whitford LA, et al. (1994) Evaluation of a new spectral peak coding strategy for the Nucleus 22 Channel Cochlear Implant System. Am J Otol 15 (suppl 2):15–27.
Smith RL (1979) Adaptation, saturation, and physiological masking in single auditory-nerve fibers. J Acoust Soc Am 65:166–178.
Sommers M, Kewley-Port D (1996) Modeling formant frequency discrimination of female vowels J Acoust Soc Am 99:3770–3781.
Speaks C, Karmen JL, Benitez L (1967) Effect of a competing message on synthetic sentence identification. J Speech Hear Res 10:390–395.
Spieth W, Webster JC (1955) Listening to differentially filtered competing messages. J Acoust Soc Am 27:866–871.
Spieth W, Curtis JF, Webster JC (1954) Cues that aid in listening to one of two simultaneous voice messages. J Acoust Soc Am 26:391–396.
Steeneken HJM, Houtgast T (1980) A physical method for measuring speech-transmission quality. J Acoust Soc Am 67:318–326.
Steeneken HJM, Houtgast T (2002) Validation of the revised STI_r method. Speech Commun 38:413–425.
Stevens KN (1980) Acoustic correlates of some phonetic categories. J Acoust Soc Am 68:836–842.
Stevens KN (1983) Acoustic properties used for the identification of speech sounds. In: Parkins CW, Anderson SW (eds) Cochlear Prostheses: An International Symposium Ann NY Acad Sci 403:2–17.
Stevens SS, Miller GA, Truscott I (1946) The masking of speech by sine waves, square waves, and regular and modulated pulses. J Acoust Soc Am 18:418–424.
Stickney GS, Assmann PF (2001) Acoustic and linguistic factors in the perception of bandpass-filtered speech. J Acoust Soc Am 109:1157–1165.
Stubbs RJ, Summerfield AQ (1991) Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms. J Acoust Soc Am 89:1383–1393.
Studebaker GA, Sherbecoe RL (2002) Intensity-importance functions for band-limited monosyllabic words. J Acoust Soc Am 111:1422–1436.
Studebaker GA, Pavlovic CV, Sherbecoe RL (1987) A frequency importance function for continuous discourse. J Acoust Soc Am 81:1130–1138.

Summerfield Q (1983) Audio-visual speech perception, lipreading, and artificial stimulation. In: Lutman ME, Haggard MP (eds) Hearing Science and Hearing Disorders. London: Academic Press, pp. 131–182.

Summerfield Q (1987) Speech perception in normal and impaired hearing. Br Med Bull 43:909–925.

Summerfield Q (1992) Role of harmonicity and coherent frequency modulation in auditory grouping. In: Schouten, MEH (ed) The Auditory Processing of Speech. Berlin: Mouton de Gruyter.

Summerfield Q, Assmann PF (1987) Auditory enhancement in speech perception. In: Schouten MEH (ed) The Psychophysics of Speech Perception. Dordrecht: Martinus Nijhoff, pp. 140–150.

Summerfield Q, Assmann PF (1989) Auditory enhancement and the perception of concurrent vowels. Percept Psychophys 45:529–536.

Summerfield Q, Culling JF (1992) Auditory segregation of competing voices: absence of effects of FM or AM coherence. Philos Trans R Soc Lond B 336:357–366.

Summerfield Q, Culling JF (1995) Auditory computations which separate speech from competing sounds: a comparison of binarual and monaural processes. In: Keller E (ed) Speech Synthesis and Speech Recognition. London: John Wiley.

Summerfield Q, Haggard MP, Foster JR, Gray S (1984) Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect. Percept Psychophys 35:203–213.

Summerfield Q, Sidwell A, Nelson T (1987) Auditory enhancement of changes in spectral amplitude. J Acoust Soc Am 81:700–708.

Summers WV, Pisoni DB, Bernacki RH, Pedlow RI, Stokes MA (1988) Effects of noise on speech production: acoustic and perceptual analyses. J Acoust Soc Am 84:917–928.

Sussman HM, McCaffrey HA, Matthews SA (1991) An investigation of locus equations as a source of relational invariance for stop place categorization. J Acoust Soc Am 90:1309–1325.

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

Tartter VC (1991) Identifiability of vowels and speakers from whispered syllables. Percept Psychophys 49:365–372.

Trees DA, Turner CC (1986) Spread of masking in normal and high-frequency hearing-loss subjects. Audiology 25:70–83.

Treisman AM (1960) Contextual cues in selective listening. Q J Exp Psychol 12:242–248.

Treisman AM (1964) Verbal cues, language, and meaning in selective attention. Am J Psychol 77:206–219.

Turner CW, Bentler RA (1998) Does hearing aid benefit increase over time? J Acoust Soc Am 104:3673–3674.

Turner CW, Henn CC (1989) The relation between frequency selectivity and the recognition of vowels. J Speech Hear Res 32:49–58.

Turner CW, Souza PE, Forget LN (1995) Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners. J Acoust Soc Am 97:2568–2576.

Uchanski RM, Choi SS, Braida LD, Reed CM, Durlach NI (1994) Speaking clearly for the hard of hearing. IV: Further studies on speaking rate. J Speech Hear Res 39:494–509.

Van Tasell DJ, Fabry DA, Thibodeau LM (1987a) Vowel identification and vowel masking patterns of hearing-impaired listeners. J Acoust Soc Am 81:1586–1597.
Van Tasell DJ, Soli SD, Kirby VM, Widin GP (1987b) Temporal cues for consonant recognition: training, talker generalization, and use in evaluation in cochlear implants. J Acoust Soc Am 82:1247–1257.
Van Wijngaarden SJ, Steeneken HJM, Houtgast T (2002) Quantifying the intelligibility of speech in noise for non-native listeners. J Acoust Soc Am 111:1906–1916.
Veen TM, Houtgast T (1985) Spectral sharpness and vowel dissimilarity. J Acoust Soc Am 77:628–634.
Vershuure J, Brocaar MP (1983) Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. Percept Psychophys 33:232–240.
Viemeister N (1979) Temporal modulation transfer functions based upon modulation transfer functions. J Acoust Soc Am 66:1364–1380.
Viemeister NF (1980) Adaptation of masking. In: Brink G van der, Bilsen FA (eds) Psychophysical, Physiological and Behavioural Studies in Hearing. Delft: Delft University Press.
Viemeister NF, Bacon S (1982) Forward masking by enhanced components in harmonic complexes. J Acoust Soc Am 71:1502–1507.
Walden BE, Schwartz DM, Montgomery AA, Prosek RA (1981) A comparison of the effects of hearing impairment and acoustic filtering on consonant recognition. J Speech Hear Res 24:32–43.
Wang MD, Bilger RC (1973) Consonant confusions in noise: a study of perceptual features. J Acoust Soc Am 54:1248–1266.
Warren RM (1970) Perceptual restoration of missing speech sounds. Science 167:392–393.
Warren RM (1996) Auditory illusions and the perceptual processing of speech. In: Lass NJ (ed) Principles of Experimental Phonetics. St Louis: Mosby-Year Book.
Warren RM, Obusek CJ (1971) Speech perception and perceptual restorations. Percept Psychophys 9:358–362.
Warren RM, Obusek CJ, Ackroff JM (1972) Auditory induction: perceptual synthesis of absent sounds. Science 176:1149–1151.
Warren RM, Riener KR, Bashford Jr JA, Brubaker BS (1995) Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. Percept Psychophys 57:175–182.
Warren RM, Hainsworth KR, Brubaker BS, Bashford A Jr, Healy EW (1997) Spectral restoration of speech: intelligibility is increased by inserting noise in spectral gaps. Percept Psychophys 59:275–283.
Watkins AJ (1988) Spectral transitions and perceptual compensation for effects of transmission channels. Proceedings of Speech '88: 7th FASE Symposium, Institute of Acoustics, pp. 711–718.
Watkins AJ (1991) Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. J Acoust Soc Am 90:2942–2955.
Watkins AJ, Makin SJ (1994) Perceptual compensation for speaker differences and for spectral-envelope distortion. J Acoust Soc Am 96:1263–1282.
Watkins AJ, Makin SJ (1996) Effects of spectral contrast on perceptual compensation for spectral-envelope distortion J Acoust Soc Am 99:3749–3757.
Webster JC (1983) Applied research on competing messages. In: Tobias JV, Schubert ED (eds) Hearing Research and Theory, vol. 2. New York: Academic Press, pp. 93–123.

Wegel RL, Lane CL (1924) The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. Phys Rev 23:266–285.

Young K, Sackin S, Howell P (1993) The effects of noise on connected speech: a consideration for automatic processing. In: Cooke M, Beet S, Crawford M (eds) Visual Representations of Speech. Chichester: John Wiley.

Yost WA, Dye RH, Sheft S (1996) A simulated "cocktail party" with up to three sound sources. Percept Psychophys 58:1026–1036.

Zahorian SA, Jagharghi AJ (1993) Spectral-shape features versus formants as acoustic correlates for vowels. J Acoust Soc Am 94:1966–1982.