SPEECH PERCEPTION

The study of speech perception is concerned with the process by which the human listener, as a participant in a communicative act, derives meaning from spoken utterances. Modern speech research began in the late 1940s, and the problems that researchers in speech perception have focused on have remained relatively unchanged since. They are (1) variability in the physical signal and the search for acoustic invariants, (2) human perceptual constancy in the face of diverse physical stimulation, and (3) the neural representation of the speech signal. The goal of this article is to examine how these problems have been addressed by various theories of speech perception and to describe how basic assumptions about the nature of the problem have shaped the course of research. Because of the breadth of information to be covered, this article will not examine the specifics of experimental methodology or survey the empirical literature in the field. Detailed reviews of speech perception can supply further background on these topics (1–6).

The process of speech perception may be limited to the auditory channel alone as in the case of a telephone conversation. However, in everyday spoken language, the visual channel is also involved as is the study of multimodal speech perception, and spoken language processing is one of the central areas of current research. Even though stimulus variability, perceptual constancy, and neural representation are core problems in all areas of perception research, speech perception is unlike other perceptual processes because the per-

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*.

ceiver also produces spoken language and therefore has intimate knowledge of the signal source. This relationship, combined with the high communicative load of speech constrains the signal significantly and affects both perception and production strategies (7–9). Speech perception is also unique in its remarkable robustness in the face of a wide range of environmental and communicative conditions. The listeners remains remarkably constant in the face of a significant amount of production-related variation in the signal. Furthermore, even in the worst of environmental conditions in which large portions of the signal are distorted or masked, the spoken message is recovered with little or no error. As we shall see, part of this perceptual robustness derives from the richness and redundancy of information in the signal, part of it lies in the highly structured nature of language, and part comes from the context dependent nature of spoken language.

Extracting meaning from the acoustic signal may at first glance seem like a relatively straightforward task. It would seem to be simply a matter of identifying the acoustically invariant characteristics in the frequency and time domains of the signal that correspond to the appropriate serially ordered linguistic units (i.e., reversing the encoding of those mental units by the production process). From those units, the hearer can then retrieve the appropriate lexical entries from memory. Although stated rather simply here, this approach is based on an assumption about the process of speech perception that has been at the core of most symbolic processing approaches (1). That is, the process involves the segmentation of the signal into discrete and abstract linguistic units such as features, phonemes, or syllables. Before or during segmentation the extralinguistic information is segregated from the intended message and is processed separately or discarded. For this process to succeed, the spoken signal must meet two conditions The first condition, known as the *invariance condition,* is that there is invariant information in the signal that is present in all instances that correspond to the perceived linguistic unit. The second condition, known as the *linearity condition,* is that the information in the signal is serially ordered so that information about the first linguistic unit precedes and does not completely overlap or follow information about the next linguistic unit and so forth.

It has become apparent to speech researchers over the last 40 years that the invariance and linearity conditions are almost never met in the actual speech signal (10,11). This has led to several innovations that have achieved varying degrees of success in accommodating some of the variability and much of the nonlinearity inherent in the speech signal (12,13). However, inter- and intra-talker variability remains an intractable problem within these conceptual/theoretical frameworks. Recent approaches that treat the signal holistically have proven promising alternatives. Much of the variability that researchers sought to strip away in traditional approaches contains important information about the talker and about the intended message. Recent approaches, while differing significantly in their view of perception, treat the signal as information-rich. The information in the speech signal is both linguistic, the traditional message of the signal, and nonlinguistic or indexical (14), supplying information about the talker's immediate physical and emotional state, about the talker's relationship to the environment, about the social context, and the like. Much of the variability and redundancy in the signal can be used to enhance the perceptual process rather than being discarded as noise (2,15,16).

THE ABSTRACTIONIST/SYMBOLIC APPROACH TO SPEECH PERCEPTION

Traditional approaches to speech perception are based on ideas that originated in information theory and have treated the process of speech perception as distinct from word recognition, sentence understanding, and speaker recognition. In this view, the decoding of the speech signal into abstract symbolic units (i.e., features, phonemes, syllables) is the goal of speech perception, and the discrete units are then passed along to be used by higher-level parsers that identify lexical items such as morphemes or words. Listeners are hypothesized to extract abstract, invariant properties of the acoustic signal to be matched to prototypical representations stored in long-term memory (17,18). In fact, most models of word recognition have been implemented using either the segment or the syllable as the fundamental unit of processing [e.g., Refs. (19) and (20)]. Although they could in theory be implemented to work directly off of acoustic input, in current forms they implicitly assume some type of low-level recoding process.

The assumption that speech is perceived in abstract idealized units has led researchers to search for simple first-order physical invariants and to ignore the problem of stimulus variability in the listener's environment (12). In this view, variability is treated as noise. This means that much of the talker-specific characteristics, or indexical information, that a listener uses to identify a particular talker, or a talkers state, is removed through a process of normalization, leaving behind the intended linguistic message (1). In this view, normalization converts the physical signal to a set of abstract units that represent the linguistic message symbolically.

The dissociation of form from content in speech perception has persisted in large part despite the fact that the both sources of information are carried simultaneously and in parallel in the acoustic signal and despite the potential gain that a listener may get from simultaneously receiving contextual information such as the rate of an utterance, or the gender, socioeconomic status, and mood of the talker. Following models of concept learning and memory, this view of speech perception has been termed the *abstractionist* approach (4). Because the abstractionist approach relies on a set of idealized linguistic units, it is useful to review the types of perceptual units that are commonly used and the motivations for abstract units in the first place.

The use of abstract symbolic units in almost all traditional models of speech perception came about for several reasons, one being that linguistic theory has had a great impact on speech research. The abstract units that had been proposed as tools for describing patterns of spoken language, themselves a reflection of the influence of information theory on linguistics (21), were adopted by many speech researchers (10). This view can be summed up by a quote from Halle (22):

> when we learn a new word we practically never remember most of the salient acoustic properties that must have been present when the acoustic signal struck our ears; for example, we do not remember the voice quality, the speed of utterance, and other

> properties directly linked to the unique circumstances directly surrounding every utterance (p. 101).

Even though linguistic theory has moved away from the phoneme as a unit of linguistic description to a temporally distributed featural or gestural array (23,24), many researchers in speech perception continue to use the phoneme as a unit of perception.

Another reason for the use of abstract units lies in the nature of the speech signal. Because of the way speech is produced in the vocal tract, the resulting acoustic signal is continuously changing, making all information in the signal highly variable and transient. This variability, combined with the constraints on auditory memory, led many researchers to assume that the analog signal must be rapidly recoded into discrete and progressively more abstract units (25). This process achieves a large reduction of data that was thought to be redundant or extraneous into a few predefined and timeless dimensions. However, even though reduction of redundancy potentially reduces the memory load, it increases the processing load and greatly increases the potential for an unrecoverable error on the part of the hearer (2). Furthermore, there is evidence that much of the information in the signal that was deemed extraneous is encoded and stored by the memory system and subsequently used by the hearer in extracting meaning from the spoken signal (4,26).

An additional motivation for postulating abstract units comes from the phenomenon of perceptual constancy. Although there is substantial contextual variation in the acoustic signal, the hearer appears to perceive a single unit of sound. For example, a voiceless stop consonant such as /t/ that is at the beginning of a word, as in the word top, is accompanied by a brief puff of air at its release and a period of voicelessness in the following vowel which together are generally referred to as *aspiration.* When that same stop is preceded by the fricative /s/, as in the word stop, the aspiration is largely absent. Yet the hearer perceives the two very different acoustic signals as being the same sound category /t/. This particular example of perceptual constancy may be explained in terms of the possible lexical contrasts of English. Even though there are lexical distinctions that are based on the voicing contrast, cat versus cad for example, no lexical distinction in English is based on an aspiration contrast. It should be remembered that most contextual variation that is noncontrastive in one language is often the foundation of a lexical contrast in another language (27). Rather than being hardwired into the brain at birth or being imposed on the hearer by transformations of the peripheral auditory system, these contrastive characteristics of a particular language must be learned. Thus, the complex process by which perceptual normalization takes place in a particular language is almost entirely the result of perceptual learning and categorization.

Finally, segmenting the speech signal into units that are hierarchically organized permits a duality of patterning of sound and meaning (28) that is thought to give language its communicative power. That is, smaller units such as phonemes may be combined according to language-specific phonotactic constraints into morphemes and words, and words may be organized according to grammatical constraints into sentences. This means that with a small set of canonical sound units, and the possibility of recursiveness, the talker may produce and the hearer may decode and parse a virtually unbounded number of utterances in the language. There are many types of proposed abstract linguistic units that are related in a nested structure with features at the terminal nodes and other types of units as branching nodes that dominate them [e.g., Refs. (24) and (29)]. The higher level units include, in ascending order, phonemes, syllables, morphemes, words, syntactic phrases, and intonation phrases.

Different approaches to speech perception employ different units and different assumptions about levels of processing. Yet, there is no evidence for the primacy of any particular unit in perception. In fact, the perceptual task itself may determine the units that hearers use to analyze the speech signal (30). Many behavioral studies have found that human listeners appear to segment the signal into phoneme-sized units. For example, it has been found that reaction times of English listeners to phonotactically permissible consonant-vowel-consonant (CVC) syllables (where all the sounds in isolation are permissible) were no faster than reaction times to phonotactically impermissible CV syllables indicating that the syllable plays no role in spoken language processing (31). However, the same experiments conducted with native speakers of French have found that listeners' response times are significantly more rapid to the phonotactically permissible CVC syllable than to the CV syllables, whereas responses of Japanese listeners to the CVC were significantly slower than to the CV. Taken together, findings of task-specific and language-specific biases in the preferred units of segmentation indicate that a particular processing unit is contingent on a number of factors. Moreover, there is a great deal of evidence that smaller units like the phoneme or syllable are perceptually contingent on larger units such as the word or phrase (32,33). This interdependence argues against the strict hierarchical view of speech perception in which the smallest units are extracted as a precursor to the next higher level of processing. Rather, the listeners' responses appear to be sensitive to attentional demands, processing contingencies, and available (34).

BASIC STIMULUS PROPERTIES

Understanding the nature of the stimulus is an important step in approaching the basic problems in speech perception. This section will review some of the crucial findings that are relevant to models of speech perception. A large portion of the research on speech perception has been devoted to the investigation of speech cues and some of the better known findings are discussed in four subsections: vowels, consonant place, consonant manner, and consonant voicing. In addition to the auditory channel, the visual channel is known to affect speech perception, and we discuss some of the key findings. Because the speech signal is produced by largely overlapping articulatory gestures, information in the signal is distributed in an overlapping fashion. Nonlinearity of information in the speech signal is reviewed and implications for speech perception are touched upon. Finally, although it was largely ignored in the past, variability is arguably the most important issue in speech perception research. This problem comes from many sources; some of the most important sources of variability in speech and the perceptual consequences are reviewed in the last part of this section.

Speech Cues

Since the advent of modern speech research at the end of the Second World War, much of the work on speech perception has focused on identifying aspects of the speech signal that contain the minimal information necessary to convey a speech contrast. These components of the signal are referred to as speech cues. The assumption that a small set of acoustic features or attributes in the acoustic signal provide cues to linguistic contrasts was the motivation for the search for invariant cues which was central to speech perception research from the mid-1950s up until the present time. The more the signal was explored for invariant acoustic cues, the more the problems of variability and nonlinearity became evident. One solution to this problem was to study speech using highly controlled stimuli to minimize the variability. Stimuli for perception experiments were usually constructed using a single synthetic voice, producing words in isolation with a single contrast in one segmental context and syllable position. This approach produced empirical evidence about a very limited set of circumstances, thereby missing much of the systematic variability and redundancy of information that plays a part in speech perception. It also artificially removed talker-specific information and other extra-linguistic contextual information that was later shown to be used by listeners during speech perception. Despite these shortcomings, early work on speech perception has provided valuable empirical data that must be considered when evaluating the relative merits of current speech perception models.

The acoustic signal is produced by articulatory gestures that are continuous and overlapping to various degrees; thus, the resulting acoustic cues vary greatly with context, speaking rate, and talker. Contextual variation is a factor that contributes to redundancy in the signal. Although the early study of speech cues sought to identify a single primary cue to a particular linguistic contrast, it is improbable that the human perceptual system fails to take advantage of the redundancy of information in the signal. It should also be noted that many factors may contribute to the salience of a particular acoustic cue in the identification of words, syllables, or speech sounds. These include factors that are part of the signal such as coarticulation or positional allophony and semantic factors such as the predictability of a word that the listener is trying to recover and whether or not the listener has access to visual information generated by the talker. The extent to which a listener attends to particular information in the speech signal is also dependent on the particular system of sound contrasts in his/her language. For example, while speakers of English can use a lateral/rhotic contrast (primarily in F3) to distinguish words like light from words like right, many languages lack this subtle contrast. Speakers of those languages (e.g., Japanese) have great difficulty attending to the relevant speech cues when trying to distinguish /l/ from /r/. Thus, the speech cues that are discussed in this section should not be considered invariant or universal, instead they are context-sensitive and highly interactive.

Vocalic Contrasts. The vocal tract acts as a time-varying filter with resonant properties that transform frequency spectra of the sound sources generated in the vocal tract (35). Movements of the tongue, lips, and jaw cause changes in the resonating characteristics of the vocal tract that are important in distinguishing one vowel from another. Vowels distinctions are generally thought to be based in part on the relative spacing of the fundamental frequency (f0) and the first three vocal tract resonances or formants (F1, F2, F3) (36). In general, there is an inverse relationship between the degree of constriction (vowel height) and the height of the first formant (35). That is, as the degree of constriction in the vocal tract increases, increasing the vowel height, F1 lowers in frequency. The second formant is generally correlated with the backness of the vowel: the further back in the vocal tract the vowel constriction is made, the lower the second formant (F2). F2 is also lowered by lip rounding and protrusion. Thus, the formant frequencies of vowels and vowellike sounds are produced by changes in the length and shape of the resonant cavities of the vocal tract above the laryngeal sound source.

In very clear speech, vowels contain steady state portions where the relative spacing between the formants remains fixed, and the f0 remains relatively constant. Words spoken with care and speech samples from read sentences may contain steady state vowels. Early work on vowel perception that was modeled on this form of carefully articulated speech found the primary cue for vowel perception was the steady state formant values (37). Figure 1 is a spectrogram illustrating the formant structure of five carefully pronounced nonsense words with representative vowels dVd contexts: /did/ (sounds like deed), /ded/ (sounds like dayed), /dad/ (sounds like dodd), /dod/ (sounds like doad), and /dud/ (sounds like dood). The first two formants, the lowest two dark bars, are clearly present and have relatively steady state portions near the center of each word. The words /did/ and /dad/ have the clearest steady state portions.

In naturally spoken language, however, formants rarely achieve a steady state and are usually flanked by other speech sounds that shape the formant structure into a dynamic time-varying pattern. For the same reasons, vowels often fall short of the formant values observed in careful speech resulting in undershoot (35,38). The degree of undershoot is a complex function of the flanking articulations, speaking rate, prosody, sentence structure, dialect, and individual speaking style (9). These observations have led researchers to question the assumption that vowel perception relies on the perception of steady state formant relationships, and subsequent experimentation has revealed that dynamic spectral information in the formant transitions into and out of the vowel are sufficient to identify vowels even in the absence of any steady state information (39).

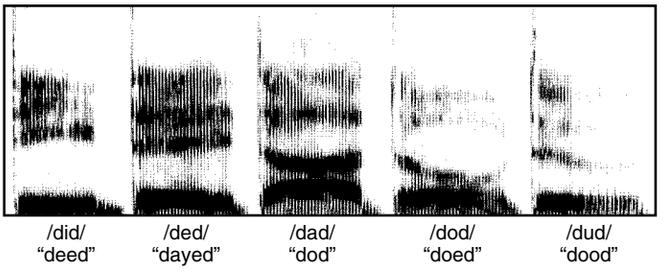

Figure 1. A spectrogram illustrating the formant structure of five representative vowels in dVd contexts. The lowest two dark bands are the first and second formants.

Although secondary vocalic contrasts are not found in English and are not well studied in the perception literature, they are found in many of the world's languages in addition to vowel height and backness contrasts. A few of the more common secondary contrasts are briefly described here; for a complete review, see Ladefoged and Maddieson (27). A secondary contrast may effectively double or, in concert with other secondary contrasts, triple or quadruple the vowel inventory of a language. The most common type of secondary contrast, found in 20% of the world's languages, is nasalization (40). Nasalization in speech is marked by a partial attenuation of energy in the higher frequencies, by a broadening of formant bandwidths, and by an additional weak nasal formant around 300 Hz (35). Vowel length contrasts are also commonly observed and are found in such diverse languages as Estonian (41), Thai (42) and Japanese (43). *Source characteristics,* changes in the vibration characteristics of the vocal folds, may also serve as secondary contrasts. These include *creaky* and *breathy* vowels. In a creaky vowel, the vocal fold vibration is characterized by a smaller open-to-closed ratio resulting in more energy in the harmonics of the first and second formants, narrower formant bandwidths, and often more *jitter* (irregular vocal cord pulse rate) (44). In breathy vowels, the vocal fold vibration is characterized by a greater open-to-closed ratio resulting in more energy in the fundamental frequency, broader formant bandwidths, and often more random energy (noise component).

Consonant Place of Articulation

There are several potential sources of cues to the place of articulation of a consonant, including second formant transitions, stop release bursts, nasal pole-zero patterns, and the generation of fricative noise. The strongest place of articulation cues are found in the brief transitional period between a consonant and an adjacent vowel. Some speech cues are internal, as is the case in fricatives such as /s/ or nasals such as /n/. Other speech cues are distributed over an entire syllable as in vowel coloring by laterals such as /l/, rhotics such as /r/, and retroflex consonants such as those found in Malayam and Hindi. We briefly review several of the most important cues to place of articulation here.

Formant Transitions. The second formant (F2), and to a lesser degree the third formant (F3), provide the listener with perceptual cues to the place of articulation of consonants with oral constrictions, particularly the stops, affricates, nasals, and fricatives (45). Transitions are the deformation of the vowels formants resulting from the closure or aperture phase of a consonant's articulation (i.e., a rapid change in the resonating cavity, overlapping with the relatively open articulation of a flanking vowel). Because they are the result of very fast movements of the articulators from one position to another, formant transitions are transient and dynamic, with the speed of the transitions depending on the manner, the place of articulation (to a lesser degree), and such factors as the individual talker's motor coordination, the speaking rate and style, and the novelty of the utterance.

Unlike other consonants, glides and liquids have clear formant structure throughout their durations. Glides are distinguished from each other by the distance between the first and second formant values at the peak of constriction, whereas the English /l/ is distinguished from /r/ by the relative frequency of the third formant (46). English /l/ and /r/ cause *vowel coloring,* a change in the formant structure of the adjacent vowels, particularly the preceding one, that may last for much of the vowel's duration.

The transitions into and out of the period of consonant constriction provide place cues for consonants that are between vowels. In other positions, there is at most only a single set of formant transitions to cue place: the formant transitions out of the consonant constriction (C to V) in word onset and postconsonantal positions, and the formant transitions into the consonant constriction (V to C) in word final and preconsonantal positions. For stops in the VC (postvoculic) position with no *audible* release, formant transitions may provide the only place cues. Following the release of voiceless stops, there is a brief period of voicelessness during which energy in the formants is weakened. Following the release of aspirated stops, a longer portion or all the transition may be present in a much weaker form in the aspiration noise. It is widely thought that CV formant transitions provide more salient information about place than VC transitions (see Ref. 47 for discussion). When formant transitions into a stop closure (V to C) conflict with the transitions out of the closure (C to V), listeners identify the stop as having the place of articulation that corresponds with the C to V transitions (48). The relative prominence of CV transitions over VC transitions is also influenced by the language of the listener. For example, native Japanese-speaking listeners have been shown to be very poor at distinguishing place from VC transitions alone, whereas native Dutch and English speakers are good at distinguishing place with VC transitions (49). In this case, the difference in performance can be attributed to differences in syllable structure between the languages. English and Dutch allow postvocalic stops with contrasting place of articulation (e.g., a$\underline{c}$tor or ba$\underline{d}$), but Japanese does not; experience with Japanese syllable structure has biased Japanese speakers toward relying more on the CV transitions than VC transitions.

Fricative Noise. Fricatives are characterized by a narrow constriction in the vocal tract that results in turbulent noise either at the place of the constriction or at an obstruction downstream from the constriction (50). Frication noise is aperiodic with a relatively long duration. Its spectrum is shaped primarily by the cavity in front of the noise source (51). The spectrum of the frication noise is sufficient for listeners to recover the place of articulation reliably in sibilant fricatives such as /s/ and /z/. However, in other fricatives with lower amplitude and more diffuse spectra, such as /f/ and /v/, the F2 transition has been found to be necessary for listeners to distinguish place of articulation reliably (52). Of these, the voiced fricatives, as in the words $\underline{th}$at and $\underline{v}$at are the least reliably distinguished (53). It shou$\underline{l}\underline{d}$ be note$\underline{d}$ that this labiodental versus interdental contrast in fricatives in English is very rare in the worlds' languages (40). The intensity of frication noise and the degree of front cavity shaping is expected to affect the relative importance of the fricative noise as a source of information for other fricatives as well.

Because fricatives have continuous noise that is shaped by the cavity in front of the constriction, they also convey information about adjacent consonants in a fashion that is similar to vowels. Overlap with other consonant constrictions results in changes in the spectral shape of a portion of the frication

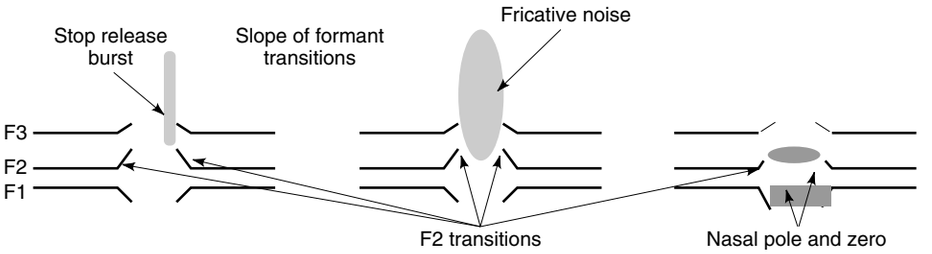

Figure 2. Schematic illustration of place cues in three VCV sequences where V is the vowel /a/ and C is a voiceless alveolar stop, a voiceless alveolar fricative, and an alveolar nasal.

noise, most markedly when the constriction is in front of the noise source. The offset frequency of the fricative spectrum in fricative-stop clusters serves as a cue to place of articulation of the stop (54).

Stop Release Bursts. In oral stop articulations, there is complete occlusion of the vocal tract and a resulting build-up of pressure behind the closure. The sudden movement away from complete stricture results in brief high-amplitude noise known as the *release burst* or *release transient.* Release bursts are aperiodic with a duration of approximately 5 to 10 ms. The bursts duration depends on both the place of articulation of the stop and the quality of the following vowel. Velar stop releases (/k/ and /g/) are longer and noisier than labial and dental stops, and both dental and velar stops show an increased noisiness and duration of release before high vowels. Release bursts have been shown to play an important role in the perception of place of articulation of stop consonants (55). Although the release burst or the formant transitions alone are sufficient cues to place, the formant transitions have been shown to dominate place perception (i.e., if the release burst spectrum and the F2 transition provide conflicting place cues, listeners perceive place according to the F2 transition) (56). Listeners show the greatest reliance on the transition in identifying velar place in stops (55). Although less studied as a source of cues, there are many other subtler place-dependent differences among stops that are a potential source of information to the listener. For example, velar stops (/k/ and /g/) tend to have shorter closure durations than labial stops (/p/ and /b/), and amplitude differences may help in distinguishing among fricatives.

An additional class of sounds known as *affricates* are similar in some respects to stops and in other aspects to fricatives; they have a stop portion followed by a release into a fricative portion. In their stop portion, they have a complete closure, a build-up of pressure, and the resultant burst at release. The release is followed by a period of frication longer than stop aspiration but shorter than a full fricative. Both the burst and the frication provide place cues. In English, all affricates are palatoalveolar, but there is a voicing contrast (chug versus jug). The palatoalveolar affricate found in English is the most commonly found, 45% of the worlds languages have it (40), but many other places of articulation are common, and many languages have an affricate place contrast (e.g., /pf/ versus /ts/ in German).

Nasal Cues. Like the oral stops, nasal consonants have an oral constriction that results in formant transitions in the adjacent vowels. In addition, nasals show a marked weakening in the upper formants resulting from the antiresonance (zero) and a low-frequency resonance (pole) below 500 Hz. The nasal pole-zero pattern serves as a place cue (57). This cue is most reliable in distinguishing /n/ and /m/, and less so for other nasals (58). Listeners identify the place of articulation more reliably from external formant transitions than from the internal nasal portion of the signal (59). Figure 2 schematically illustrates some of the most frequently cited cues to consonant place of articulation for three types of consonants: a voiceless alveolar stop /t/, a voiceless alveolar fricative /s/, and an alveolar nasal /n/. The horizontal bars represent the first three formants of the vowel /a/, the deformation of the bars represents formant transitions, and the hatched areas represent fricative and stop release noises. Table 1 summarizes the consonant place cues discussed above. Although it lists some of the more frequently discussed cues, Table 1 should not be seen as exhaustive because there are many secondary and contextual cues that contribute to a consonant percept that are not listed here.

Consonants of all types have much narrower constrictions than vowels. They can be viewed as the layering of a series of rapidly constricting movements onto a series of more slowly moving transitions from one vowel to the next (23). For all types of consonants, the changes in the vowels' formants that result from the influence of the consonants narrower constriction are the most robust types of cues. However, we have seen that there are a number of other sources of information about the place of articulation of a consonant that the listener may use in identifying consonants. This article has touched on a few of the better known such as stop release bursts, nasal pole-zero patterns, and fricative noise. These are often referred to as secondary cues because perceptual tests have shown that when paired with formant transitions that provide conflicting information about the consonant place of articulation, the perceived place is that appropriate for the formant transitions. However, depending on the listening conditions, the linguistic context, and the perceptual task, these so-called secondary cues may serve as the primary source of information about a consonants place of articulation.

Table 1. Summary of Place Cues

Cue	Applies to	Distribution
F2 transition	All	VC, CV transitions
Burst spectrum	Stops	C-release
Frication spectrum	Fricatives, affricates (esp. sibilants)	Internal
Frication amplitude	Fricatives	Internal
Nasal pole, zero	Nasals	Internal
Fricative noise transition	Stops	Fricative edge
F3 height	Liquids and glides	Internal

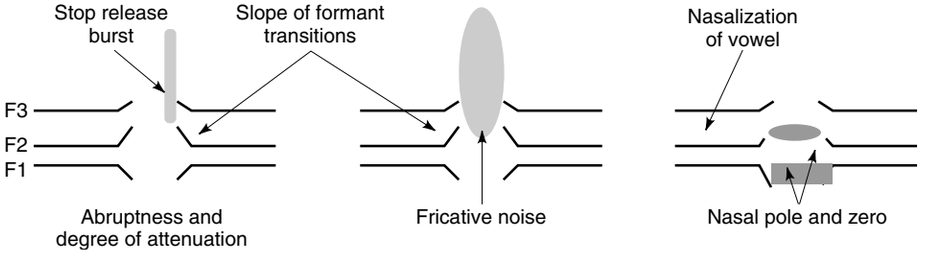

Figure 3. Schematic illustration of manner cues in three VCV sequence where V is the vowel /a/ and C is a voiceless alveolar stop, a voiceless alveolar fricative, and an alveolar nasal.

For example, in word initial fricative-stop clusters, the fricative noise may provide the sole source of information about the fricatives place of articulation. Even though in English only /s/ appears in word initial fricative-stop clusters, many languages contrast fricative place in such clusters and many others also have stop-stop clusters or nasal-stop clusters (see Ref. 47 for a description of a language that has all three types of clusters). Thus, it is likely that phonotactic constraints, position within sentence, position within word, position within syllable, background noise, and so on, must be taken into consideration before a relative prominence or salience is assigned to any particular acoustic cue in the signal.

Consonant Manner Contrasts

All oral constrictions result in an attenuation of the signal, particularly in the higher frequencies. The relative degree of attenuation is a strong cue to the manner of a consonant. An abrupt attenuation of the signal in all frequencies is a cue to the presence of a stop. Insertion of a period of silence in a signal, either between vowels or between a fricative and a vowel can result in the listener perceiving a stop (54). A complete attenuation of the harmonic signal together with fricative noise provides the listener with cues to the presence of a fricative. A less severe drop in amplitude accompanied by nasal murmur and a nasal pole and zero are cues to the presence of a nasal. Nasalization of the preceding vowel provides look-ahead cues to postvocalic nasal consonants (60).

Glides and liquids maintain formant structure throughout their peak of stricture, but both attenuate the signal more than vowels. Glides are additionally differentiated from other consonants by the relative gradualness of the transitions into and out of the peak of stricture. Lengthening the duration of synthesized formant transitions has been shown to change the listener's percept of manner from stop to glide (61). A similar cue is found in the amplitude envelope at the point of transition between consonant and vowel: stops have the most abrupt and glides have the most gradual amplitude rise time (62).

Manner cues in general tend to be more robust than place cues because they result in more salient changes in the signal, although distinguishing stop from fricative manner is less reliable with the weaker fricatives (53). Figure 3 schematically illustrates some of the most frequently cited cues to consonant manner of articulation for three types of consonants: a voiceless alveolar stop /t/, a voiceless alveolar fricative /s/, and an alveolar nasal /n/. The horizontal bars represent the first three formants of the vowel /a/, the deformation of the bars represents formant transitions, the hatched areas represents fricative and stop release noises. Table 2 summarizes the consonant manner cues discussed earlier. Again, Table 2 should not be seen as exhaustive there are many secondary and contextual cues that contribute to a consonant percept that are not listed here.

Cues to Voicing Contrasts

Vocal fold vibration, resulting in periodicity in the signal, is the primary cue to voicing; however, tight oral constriction inhibits the airflow necessary for vocal fold vibration. In English and many other languages, voiced obstruents, especially stops, may have little or no vocal fold activity. This is more common for stops in syllable final position. In this situation, the listener must rely on other cues to voicing. There are several other important cues such as voicing onset time (VOT), the presence and the amplitude of aspiration noise, and durational cues. For syllable initial stops in word onset position, the primary cue appears to be VOT. This is not really a single cue in the traditional sense but a dynamic complex that includes the time between the release burst and the onset of vocal fold vibration together with aspiration noise (i.e., low-amplitude noise with spectral peaks in the regions of the following vowels formants). VOT appears to be important even in languages like French that maintain voicing during stop closure. The relationship between VOT and voicing is, in part, dependent on how contrasts are realized in a particular language. For example, for the same synthetic VOT continuum, Spanish and English speakers have different category boundaries despite the fact that both languages have a single voiced-voiceless contrast (63). In Thai, there are two boundaries, one similar to English and one similar to Spanish, because there is a three-way voiced–voiceless–aspirated contrast in the language.

Generally, a short or negative VOT is a cue to voicing, a long VOT is a cue to voicelessness, and a very long VOT is a cue to aspiration (in languages with an aspiration contrast).

Table 2. Summary of Manner Cues

Cue	Applies to	Distribution
Silence/near silence	Stops, affricates	Internal
Frication noise	Fricatives, affricates	Internal
Nasal pole and zero	Nasals	Internal
Vowel nasalization	Nasals	Adjacent vowel
Formant structure	Liquids, glides (vowels)	Internal
Release burst	Stops	C-release
Noise duration	Stop, affricate, fricative	Internal
Noise onset rise-time	Stop/affricate, fricative	Internal
Transition duration	Stop, glide	VC, CV transitions

For English, and presumably other languages, the relative amplitude and the presence or absence of aspiration noise is a contributing cue to voicing for word initial stops. An additional cue to voicing in syllable onset stops is the relative amplitude of the release burst: a low-amplitude burst cues voiced stops, whereas a high-amplitude burst cues voiceless stops (64).

The duration and spectral properties of the preceding vowel also provide cues to voicing in postvocalic stops and fricatives (65). When the vowel is short, with a shorter steady state relative to its offset transitions, voicelessness is perceived. The duration of the consonant stricture is also a cue to both fricative and stop voicing: longer duration cues voicelessness (65a). Figure 4 schematically illustrates some of the most frequently cited cues to consonant voicing for two types of consonants: a voiceless alveolar stop /t/, and a voiced alveolar fricative /z/. The horizontal bars represent the first three formants of the vowel /a/, the deformation of the bars represents formant transitions, the hatched areas represents fricative and stop release noises, and the dark bar at the base of the /z/ represents the voicing bar. Table 2 summarizes the consonant manner cues discussed earlier. Again, Table 2 should not be seen as exhaustive because there are many secondary and contextual cues that contribute to a consonant percept that are not listed here.

Visual Information: Multimodal Speech Perception

Much of the research on speech perception focuses on the acoustic channel alone. In part, the concentration on auditory perception is related to the fact that the acoustic signal is richer in information about spoken language than the visual signal. However, the visual signal may have a large impact on the perception of the auditory signal under degraded conditions. When a hearer can see a talker's face, the gain in speech intelligibility in a noisy environment is equivalent to a 15-dB gain in the acoustic signal alone (66). This is a dramatic difference, superior to that of even the most sophisticated hearing aids. The relative importance of the visual signal increases as the auditory channel is degraded through noise, distortion, filtering, hearing loss, and potentially through unfamiliarity with a particular talker, stimulus set, or listening condition.

When information in the visual channel is in disagreement with the information in the auditory channel, the visual channel may change or even override the percept of the auditory channel alone. McGurk and MacDonald (67) produced stunning evidence, now known as the McGurk effect, of the strength of the visual signal in the perception of speech in an experiment that has since been replicated under a variety of conditions. They prepared a videotape of a talker producing two-syllable utterances with the same vowel but varying in the onset consonants such as baba, mama, or tata. The audio and video channels were separated and the audio tracks of one utterance were dubbed onto video tracks of different utterances. With their eyes open, subjects perceptions were strongly influenced by the video channel. For example, when presented with a video of a talker saying tata together with the audio of the utterance mama the subjects perceived nana. But with their eyes closed, subjects perceived mama. This effect of cross-modal integration is strong and immediate, there is no hesitation or contemplation on the part of the subjects who are completely unaware of the conflict between the two channels. The McGurk effect is considered by many theorists as evidence that the auditory and visual integration occurs at a low level because of its automatic nature. It also reflects limitations on the information that can be obtained through the visual channel. Many aspects of the speech production process are hidden from view. These include voicing, nasalization, and many vowel and consonant contrasts.

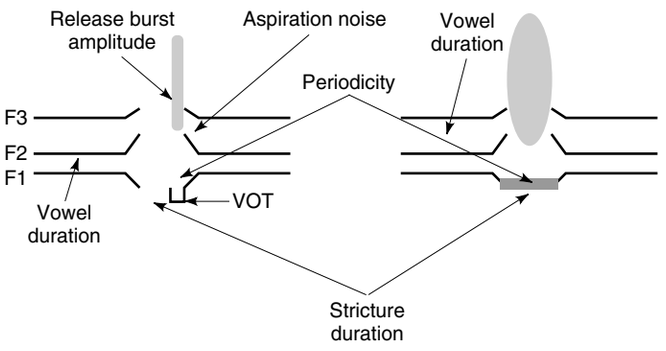

Figure 4. Schematic illustration of the voicing cues in two VCV sequences where V is the vowel /a/ and C is a voiceless alveolar stop /t/ and a voiced alveolar fricative /z/.

Nonlinearity of the Speech Signal

As a result of the way in which speech is produced, much of the information in the signal is distributed, overlapping, and contextually varying. In producing speech, the articulatory organs of the human vocal tract move continuously with sets of complex gestures that are partially or wholly coextensive and covarying (see SPEECH PRODUCTION). The resulting acoustic and visual signals are continuous and the information that can be identified with a particular linguistic unit shows a high degree of overlap and covariance with information about adjacent units (45). This is not to say that segmentation of the signal is impossible; acoustic analysis reveals portions of the signal that can act as reliable acoustic markers for points at which the influence of one segment ends or begins. However, the number of segments determined in this way and their acoustic characteristics are themselves highly dependent on the context (67a).

As noted earlier, because of the distributed and overlapping nature of phonetic/linguistic information, the speech signal fails to meet the *linearity* condition (11). This poses great problems for phoneme based-speech recognizers and, if discrete units do play a part in perception, they should also be problematic for the human listener. Yet, the listener appears to segment the signal into discrete and linear units such as words, syllables, and phonemes with little effort. In the act of writing, much of the world's population can translate a heard or internally generated signal into wordlike units, syllablelike units, or phonemelike units. Although this is often cited as an argument for a segmentation process in speech perception, the relation between the discrete representation of speech seen in the world's writing systems and the continuous signal is complex and may play little role in the perceptual process. Segmentation may be imposed on an utterance after the perceptual process has been completed. It is not clear that a signal of discrete units would be preferable; the distributed na-

ture of information in the signal contributes to robustness by providing redundant look-ahead and look-back information. Reducing speech to discrete segments could result in a system that, unlike human speech perception, cannot recover gracefully from errorful labeling (2).

Informational Burden of Consonants and Vowels

From an abstract phonemic point of view, consonant phonemes bear a much greater informational burden than vowel phonemes. That is, there are far more consonant phonemes in English than there are vowel phonemes, and English syllables permit more consonant phonemes per syllable than vowel phonemes. Thus, many more lexical contrasts depend on differences in consonant phonemes than in vowel phonemes. However, the complex overlapping and redundant nature of the speech signal means that the simple information theoretic analysis fails in its predictions about the relative importance of consonant and vowel portions of the signal in speech perception.

The importance of vowels is a result of the gross differences in the ways consonants and vowels are produced by the vocal tract. Consonants are produced with a complete or partial occlusion of the vocal tract, causing a rapid attenuation of the signal, particularly in the higher frequencies. In the case of oral stops, all but the lowest frequencies (which can emanate through the fleshy walls of the vocal tract) are absent from the signal. In contrast, vowels are produced with a relatively open vocal tract; therefore, there is little overall attenuation, and formant transitions are saliently present in the signal (35). This dichotomy means that the vowels are more robust in noise and that vowel portions of the signal carry more information about the identity of the consonant phonemes than the consonant portions of the signal carry about the vowel phonemes.

Although they are partially the result of articulator movement associated with consonants, the formant transitions are considered part of the vowel because of their acoustic characteristics. Generally speaking, the transitions have a relatively high intensity and long duration compared to other types of consonantal cues in the signal. The intensity, duration, and periodic structure of the transitions make them more resistant to many types of environmental masking than release bursts, nasal pole-zero patterns, or frication noise. Formant transitions bear a dual burden of simultaneously carrying information about both consonant and vowel phonemes. In addition, information about whether or not a consonant phoneme is a nasal, a lateral, or a rhotic is carried in the vowel more effectively than during the consonant portion of the signal. Figure 5 is a speech spectrogram of the word formant illustrating the informational burden of the vowel. What little information consonants carry about the flanking vowel phonemes is found in portions of the signal that are low-intensity, aperiodic, or transient. Therefore, it is more easily masked by environmental noise.

It is well known that spoken utterances are made up of more than the segmental distinctions represented by consonant and vowel phonemes. Languages like English rely on lexical stress to distinguish words. The majority of the world's languages have some form of *tone* contrast, whether fixed on a single syllable as in the Chinese languages, or mobile across several syllables as in Kikuyu (68). *Pitch-accent,* like that

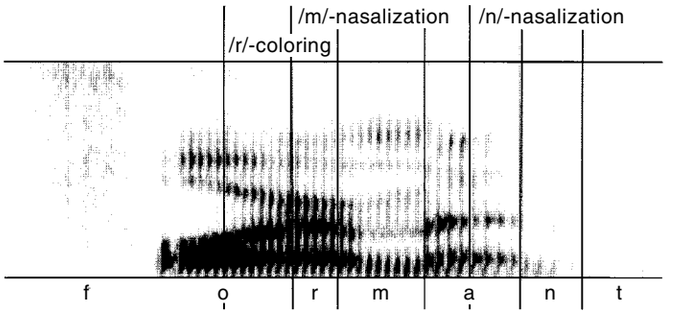

Figure 5. A spectrogram of the word formant illustrating the information that adjacent vowels carry about rhotics and nasals.

seen in Japanese, is another form of tone-based lexical distinction. Tone and pitch-accent are characterized by changes in voice pitch (fundamental frequency) and in some cases changes in voice quality such as creakiness or breathiness (as in Vietnamese, see Ref. 69). These types of changes in the source function are carried most saliently during the vowel portions of the signal. Although stress affects both consonant and vowels, it is marked most clearly by changes in vowel length, in vowel formants, and in fundamental frequency excursions. Prosodic information is carried by both consonants and vowels; however, much of it takes the form of pitch, vowel length, and quality changes. Thus, despite the relatively large information burden that consonant phonemes bear, the portions in the physical signal that are identified with vowels carry much more of the acoustic information in a more robust fashion than the portions of the signal associated with the consonants.

Invariance and Variability

In addition to violating the linearity condition, the speech signal is characterized by a high degree of variability, violating the invariance condition. There are many sources of variability that may be interrelated or independent. Variability can be broken into two broad categories: (1) production related and (2) production independent. It is worth noting that even though production-related variability is complex, it is lawful and is a potentially rich source of information both about the intended meaning of an utterance and about the talker. Production-independent variability derives from such factors as environmental noise or reverberation and may provide the listener with information about the environmental conditions surrounding the conversation; it can be seen as random in its relation to the linguistic meaning of the utterance and to the talker. Understanding how the perceptual process deals with these different types of variability is one of the most important issues in speech perception research. In traditional symbol-processing approaches that treat variation as noise, listeners are thought to compensate for differences through a processes of perceptual *normalization* in which linguistic units are perceived relative to the context (e.g., the prevailing rate of speech) (70,71) or the dimensions of the talkers vocal tract (14). Alternative nonanalytic approaches to speech perception that are based on episodic memory (72) propose that speech is encoded in a way that preserves the fine details of speech-production-related variability. Although these approaches may use some types of variability in the speech per-

ception process, little has been said about the production-independent variability. The following section, while not exhaustive, is a sampling of some well-known sources of variability and their impact on speech perception (for a more detailed review, see Ref. 2). Production-related variability in speech applies both across talkers as a result of physiological, dialectal, and socioeconomic factors, as well as within a talker from one utterance to the next as a result of factors such as coarticulation, rate, prosody, emotional state, level of background noise, distance between talker and hearer, and semantic properties of the utterance. A review of some of the most important sources of variability in speech and their effects on perceptual processes follows.

Coarticulation. The most-studied source of within-talker variability, *coarticulation,* is one source of nonlinearity in speech. In the production of speech, the gestures in the vocal tract are partially or wholly overlapping in time, resulting in an acoustic signal in which there is considerable contextual variation (10,45). The degree to which any one speech gesture is affected or affects other gestures depends on the movements of the articulators and the degree of its constriction as well as factors such as rate of speech and prosodic position. Although coarticulation is often described as a universal physiological aspect of speech, there is evidence for talker-specific variation in the production and timing of speech gestures and in the resulting characteristics of coarticulation (73,74).

The perceptual problems introduced by coarticulatory variation became apparent early in the search for invariant speech cues. Because of coarticulation, there is a complex relationship between acoustic information and phonetic distinctions. In one context, an acoustic pattern may give rise to one percept, whereas in another context the same acoustic pattern may give rise to a different percept (101). At the same time, many different acoustic patterns may cue a single percept (75).

Speaking Rate. Changes in speech rate are reflected in changes in the number and duration of pauses, in durational changes of vowels and some consonants, and in deletions and reductions of some of the acoustic properties that are associated with particular linguistic. For example, changes in VOT and the relative duration of transitions and vowel steady states occur with changes in speaking rates (71).

There is now a large body of research on the consequences of rate-based variability on the perception of phonemes. These findings demonstrate that listeners are sensitive to rate-based changes that are internal or external to the target word. The importance of token-internal rate sensitivity was demonstrated by Miller and Liberman (76). Listeners were presented with a synthetic /ba/–/wa/ continuum that varied the duration of the formant transitions and the duration of the vowel. The results showed that the crossover point between /b/ and /w/ was dependent on the ratio of the formant transition duration to the vowel duration: the longer the vowel, the longer the formant transitions had to be to produce the /wa/ percept. The importance of token-external rate sensitivity was demonstrated in an experiment on the identification of voiced and voiceless stops. Summerfield (70) presented listeners with a precursor phrase that varied in speaking rate followed by a stimulus token. As the rate of the precursor phrase increased, the voiced–voiceless boundary shifted to shorter VOT values. Sommers, Nygaard, and Pisoni (77) found that the intelligibility of isolated words presented in noise was affected by the number of speaking rates that were used to generate the test stimulus ensemble: stimuli drawn from three rates (fast, medium, and slow) were identified more poorly than stimuli from only a single speaking rate.

Prosody. Rate-based durational variation is compounded by many other factors, including the location of syntactic boundaries, prosody, and the characteristics of adjacent segments (29,41,78). It is well known that lexical stress has a dramatic effect on the articulations that produce the acoustic signal. However, lexical stress is only one level of prosodic hierarchy spanning the utterance. Prosody is defined by Beckman and Edwards (78, p. 8) as the organizational framework that measures off chunks of speech into countable constituents of various sizes. Different positions within a prosodic structure lead to differences in articulations, which in turn lead to differences in the acoustic signal. For example, vowels that are in the nuclear-accented syllable of a sentence (primary sentential stress) have a longer duration, a higher amplitude, and a more extreme articulator displacement than vowels in syllables that do not bear nuclear accent (78,79).

Articulations that are at the edges of prosodic domains also undergo systematic variation that result in changes in the acoustic signal such as lengthened stop closures, greater release burst amplitude, lengthened VOT, and less vowel reduction. These effects have been measured for word initial versus noninitial positions and at phrase and sentence (80). Finally, the magnitude of a local effect of a prosodic boundary on an articulation interacts in complex ways with global trends that apply across the utterance. One such trend, commonly referred to as *declination,* is for articulations to become less extreme and for fundamental frequency to fall as the utterance progresses (81). Another global trend is for domain edge effects to apply with progressively more force as the edges of progressively larger domains are reached. These factors interact with local domain edge effects in a way that indicates a nested hierarchical prosodic structure. However, the number of levels and the relative strength of the effect may be a talker-dependent factor (80).

Semantics and Syntax. In addition to prosodic structure, the syntactic and semantic structure have substantial effects on the fundamental frequency, patterns of duration, and relative intensities of vowels (7,82). For example, when a word is uttered in a highly predictable semantic and syntactic position, it will show a greater degree of vowel reduction (*centralization*), with lower amplitude and a shorter duration than the identical word in a position with low contextual predictability (7,8). These production differences are correlated with speech intelligibility; if the two words are isolated from their relative contexts, the word from the low-predictability context is more intelligible than the word from the high-predictability context. This type of effect is hypothesized to be the result of the talker adapting to the listener's perceptual needs (9): the more information the listener can derive from the conversational context, the less effort a talker needs to spend maintaining the intelligibility of the utterance. The reduced speech is referred to as hypoarticulated and the nonreduced speech

is referred to as hyperarticulated. Similar patterns of variability can be seen in many other production-related phenomena such as the Lombard reflex (described later). This variability interacts with other factors like speaking rate and prosody in a complex fashion, making subsequent normalization extremely difficult. Yet, listeners are able to extract and use syntactic, semantic, and prosodic information from the lawful variability in the signal.

Environmental Conditions. Many factors can cause changes in a talker's source characteristics and in the patterns of duration and intensity in the speech signal that are not directly related to the talker or to the linguistic content of the message. These include the relative distance between the talker and the hearer, the type and level of background noise, and transmission line characteristics. For example, in a communicative situation in which there is noise in the environment or transmission line, there is a marked rise in amplitude of the produced signal that is accompanied by changes in the source characteristics and changes the dynamics of articulatory movements, which together are known as the Lombard reflex (83,84). The Lombard reflex is thought to result from the talker's need to maintain a sufficiently high signal-to-noise ratio to maintain intelligibility.

Because the environmental conditions and the distance between talker and hearer are never identical across instances of any linguistic unit, it is guaranteed that no two utterances of the same word in the same syntactic, semantic, and prosodic context will be identical. Furthermore, in natural settings, the level of environmental noise tends to vary continuously so that even within a sentence or word, the signal may exhibit changes. Similarly, if the talker and listener have the ability to communicate using both the visual and auditory channels, the resulting speech signal exhibits selective reductions such as those seen for high semantic context or good signal-to-noise ratios, but when there is no visual channel available, the resultant speech is marked by hyper-articulation that is similar to that seen for the Lombard reflex or for low-semantic predictability contexts (85). Like other types of hypo- and hyperarticulation, the variation based on access to visual information is highly correlated with speech intelligibility.

Physiological Factors. Among the most commonly cited sources of between-talker variation are differences in the acoustic signal based on a talker's anatomy and physiology. The overall length of the vocal tract and the relative size of the mouth cavity versus the pharyngeal cavity determines the relative spacing of the formants in vowels (35). These differences underlie some of the male–female and adult–child differences in vowels and resonant consonants (75,86). Moreover, vocal tract length may also contribute to observed differences in obstruent voicing (87) and fricative spectra (88). Physiological differences between male, female, and child laryngeal structure also contribute to observed differences in the source characteristics such as fundamental frequency, spectral tilt, and noisiness (89). Other types of physiologically based differences among talkers that are expected to have an effect on the acoustic signal include dentition, size and doming of the hard palate, and neurological factors such as paralysis or motor impairments.

The importance of talker-specific variability in the perception of linguistic contrasts was first reported by Ladefoged and Broadbent (14). Listeners were presented with a precursor phrase in which the relative spacing of the formants was manipulated to simulate established differences in vocal tract length. The stimulus was one of a group of target words in which the formant spacing remained fixed. The listener's percepts were shifted by the precursor sentence. In a follow-up experiment, a different group of listeners was presented with the same set of stimuli under a variety of conditions and instructions (90). Even when listeners were told to ignore the precursor sentence or when the sentence and the target word were presented from different loudspeakers, the vowel in the target word was reliably shifted by the formant manipulation of the precursor sentence. This effect was successfully countered only by placing the target word before the sentence or by having the listeners count aloud for 10 seconds between the precursor and hearing the target word.

DIALECTAL AND IDEOLECTAL DIFFERENCES

In addition to physiologically based differences, there are a number of socioeconomic and regional variables that affect the production of speech. Perception of speech from a different dialect can be a challenging task. Peterson and Barney (91) found differences in dialect to be one of the most important sources of confusions in the perception of vowel contrasts. Research on improving communications reliability found that training talkers to avoid dialectal pronunciations in favor of Standard English was much easier than training listeners to adapt to a variety of dialects (92). Differences between individual speaker's styles, or idiolect, also require a certain amount of adaptation on the part of the listener. Dialectal and ideolectal variability have received relatively little attention in the speech perception literature and are generally treated as additional sources of noise, which are discarded in the process of normalization.

Robustness of Speech Perception

In everyday conversational settings, there are many different sources of masking noise and distortions of the speech signal, yet only under the most extreme conditions is perceptual accuracy affected. Much of the robustness of speech comes from the redundant information that is available to the listener. Because the main goal of speech is the communication of ideas from the talker to the hearer (normally under less than optimal conditions), it is not surprising that spoken language is a highly redundant system of information transmission. Even though redundancy of information in a transmission system implies inefficient encoding, it facilitates error correction and recovery of the intended signal in a noisy environment and ensures that the listener recovers the talker's intended message.

The redundancy of speech resides, in part, in the highly structured and constrained nature of human language. Syntactic and semantic context play a large role in modulating the intelligibility of speech. Words in a sentence are more predictable than words spoken in isolation (93–95). The sentence structure (syntax) of a particular language restricts the set of possible words that can appear at any particular point in the sentence to members of appropriate grammatical categories.

The semantic relationships between words also aids in perception by further narrowing the set of words that are likely to appear in a sentence. It has been shown experimentally that limiting the set of possible words aids in identification. For example, Miller et al. (95) found that limiting the vocabulary to digits alone results in an increase in speech intelligibility. More generally, lexical factors like a word's frequency of usage and the number of acoustically similar words have been shown to have a dramatic impact on a words intelligibility (96).

Phonological structure also constrains the speech signal and facilitates the listener's perception of the intended message. Prosodic structure and intonation patterns provide auditory cues to syntactic structure, which reduces the number of possible parses of an utterance. The syllable structure and stress patterns of a language limit the number of possible speech sounds at any particular point in an utterance, which aids in identifying words (30).

Much of the top-down information in language is contextual in nature and resides in the structural constraints on a given language and not in the speech signal itself (97). However, because prosodic, syntactic, and semantic factors create systematic variability in production (discussed in the variability section), the signal contains a significant amount of information about the linguistic structures larger than the segment, syllable, and word. Although the role of suprasegmental information (above the level of the phoneme) has traditionally received less attention in the perception literature, there have been a few studies that reveal the richness of suprasegmental information in the speech signal.

In spectrogram reading experiments, Cole et al. (98) demonstrated that the acoustic signal is rich in information about the segmental, lexical, and prosodic content of an utterance. An expert spectrogram reader who was given the task of transcribing an utterance of unknown content using speech spectrograms alone achieved an 80 to 90% accuracy rate. This finding demonstrates that not only are features that cue segmental contrasts present in the signal, but prosodic and word boundary information is also available. However, it is not clear from these spectrogram reading experiments whether the features that the transcriber used are those that listeners use.

There are numerous other studies that demonstrate the importance of prosodic melody in sentence and word parsing. For example, Lindblom and Svensson (99), using stimuli in which the segmental information in the signal was removed, found that listeners could reliably parse sentences based on the prosodic melody alone. Prosody has been found to play a role in perceptual coherence (1) and to play a central role in predicting words of primary semantic importance (100).

A second source of the redundancy in speech comes from the fact that the physical signal is generated by the vocal tract. As we have already noted, speech sounds are overlapped, or coarticulated, when they are produced, providing redundant encoding of the signal. The ability to coarticulate, and thereby provide redundant information about the stream of speech sounds, serves to both increase transmission rate (101) and provide robustness to the signal (47). Redundancy in the acoustic signal has been tested experimentally by distorting, masking, or removing aspects of the signal and exploring the effect these manipulations have on intelligibility. For example, connected speech remains highly intelligible when the speech power is attenuated below 1800 Hz or when it is attenuated above 1800 Hz (94). This finding indicates that speech information is distributed redundantly across lower and higher frequencies. However, not all speech sounds are affected equally by frequency attenuation. Higher-frequency attenuation causes greater degradation for stop and fricative consonants, whereas lower-frequency attenuation results in greater degradation of vowels, liquids (/r/ and /l/ in English), and nasal consonants (93). For example, the place of articulation distinctions among fricatives are carried in large part by the fricative noise, which tends to be concentrated in higher frequencies. Attenuating these particular frequencies results in an increase in fricative confusions and a decrease in intelligibility.

Speech can be distorted in a natural environment by reverberation. Experiments on the perception of nonsense syllables found that intelligibility was relatively unaffected by reverberation with a delay of less than 1.5 s. Reverberation with a greater delay caused a marked drop off in intelligibility (102). Under extremely reverberatory conditions, individual speech sounds blend together as echoes overlap in a way that causes frequency and phase distortions. Again, not all speech sounds are equally affected by reverberation. Long vowels and fricatives, which have an approximately steady state component, are much less susceptible to degradation than short vowels and nasal and stop consonants, which are distinguished from each other by relatively short and dynamic portions of the signal.

Overall, the intelligibility of individual speech sounds in running speech is in part a function of their intensity. In general, vowels are more intelligible than consonants. More specifically, those consonants with the lowest intensity have the poorest intelligibility. Of these, the least reliably identified in English are the nonsibilant fricatives, such as those found in f̲at, v̲at, th̲in, and th̲is. These fricatives achieve 80% correct identification only when words are presented at relatively high signal-to-noise ratios (93). These fricatives noises are also spectrally similar, adding to their confusability with each other. English is one of the few languages that contrasts nonsibilant fricatives (40), presumably because of their confusability and low intelligibility. By contrast, the sibilant fricatives (for example, those found in the words s̲ap, z̲ap, Confu̲cian, and confu̲sion) have a much greater intensity and are more reliably identified in utterances presented at low signal-to-noise ratios. The next most intelligible sounds are the stop consonants, including /p/, /t/, and /k/ in English, followed by the vocalic consonants such as the nasals and liquids. Vowels are the most identifiable. The low vowels, such as those found in the words cot and caught, are more easily identified than the high vowels, such as those found in peat and pit.

MODELS AND THEORIES

Theories of human speech perception can be divided into two broad categories, those that attempt to model segmentation of the spoken signal into linguistic units (which we refer to as models of speech perception) and those that take as input a phonetic transcription and model the access of the mental lexicon (which we refer to as models of spoken word recognition). Almost all models of speech perception try to identify pho-

nemes in the signal. A few models go straight to the word level and thus encompass the process of word recognition as well. These models are discussed in the section on word recognition.

Models of Human Speech Perception

Most of the theories of speech perception and spoken word recognition have either been formalized to the extent that the predictions of the theory can be tested, or they have been explicitly implemented in computational models. Many of these theories share general theoretical claims. For example, in speech perception, information is extracted from the acoustic signal and used to identify more abstract phonological or lexical units. In spoken word recognition, this information is used to pick out the target word from all other words, particularly among those which are highly similar to and most confusable with the target. Often what differentiates these theories is the way they have been formalized and implemented. It is these differences that are the focus of our discussion.

Invariance Approaches. The most extensively pursued approach to solving the variability problem is the search for invariant cues in the speech signal. This line of research, which dates back to the beginning of modern speech research in the late 1940s, has revealed a great deal of coarticulatory variability. It has resulted in a series of careful and systematic searches for invariance in the acoustic signal that has revealed a wealth of empirical data. Although researchers investigating acoustic-phonetic invariance differ in their approaches, they have in common the fundamental assumption that the variability problem can be resolved by studying more sophisticated cues than were originally considered (6). Early experiments on speech cues in speech perception used copy-synthesized stimuli in which much of the redundant information in the signal had been stripped away. In addition, acoustic analysis of speech using spectrograms focused only on gross characteristics of the signal.

One approach, termed *static* (4), is based on the acoustic analysis of simple CV syllables. This approach focused on complex integrated acoustic attributes of consonants that are hypothesized to be invariant in different vowel. Based on Fant's (35) acoustic theory of speech production, Blumstein and Stevens (12) hypothesized invariant relationships between the articulatory gestures and acoustic features associated with a particular segment. They proposed that the gross spectral shape at the onset of the consonant release burst is an invariant cue for place of articulation. In labial stops (/p/ and /b/), the spectral energy is weak and diffuse with a concentration of energy in the lower frequencies. For the alveolar stops (/t/ and /d/), the spectral energy is strong but diffuse with a concentration of energy in the higher frequencies (around 1800 Hz). Velar stops (/k/ and /g/) are characterized by strong spectral energy that is compact and concentrated in the midfrequencies (around the 1000 Hz).

A different approach, termed *dynamic* (4), has been proposed by Kewley-Port (103). She employed auditory transformations of the signal, looking for invariant dynamic patterns in running spectra of those transformations. The dynamic approach is promising because it can capture an essential element of the speech signal: its continuous nature. More recent static approaches adopted an element of dynamic invariance into their approaches (104).

As is noted by Nygaard and Pisoni (4), any assumption of invariance necessarily constrains the types of processes that underlie speech perception. Speech perception will proceed in a bottom-up fashion with the extraction of invariant features or cues being the first step in the process. Invariance explicitly assumes abstract canonical units and the elimination of all forms of variability and noise from the stored representation. This includes many sources of variation that are potentially useful to the listener in understanding an utterance. For example, indexical and prosodic information is discarded in the reduction of the information in the signal to a sequence of idealized symbolic linguistic invariants.

Even though these approaches to speech sound perception have provided some promising candidates for extraction of invariant features from the signal and have produced invaluable empirical data on the acoustic structure of the speech signal and its auditory transforms, they have done so for only a very limited set of consonants in a very limited set of contexts. For example, the three places of articulation treated by Blumstein and Stevens represent slightly less than one quarter of the known consonant places of articulation in the worlds' languages (27). Even for the same places of articulation, the features found in English may not invariantly classify segments of other language. Furthermore, much of the contextual variability that is noncontrastive in English, and therefore removed in the invariance approach, forms the basis for a linguistic contrast in at least one other language. Therefore, the type of processing that produces invariant percepts must be language-specific.

Motor Theory. One of the ways in which the perception of speech differs from many other types of perception is that the perceiver has intimate experience in the production of the speech signal. Every listener is also a talker. The motor theory and the revised motor theory (13) take advantage of this link by proposing that perception and production are related by a common set of neural representations. Rather than looking for invariance in the acoustic signal, the perceiver is hypothesized to recover the underlying intended phonetic gestures from an impoverished and highly encoded speech signal. The intended gestures of the talker are therefore assumed to be perceived directly via an innate phonetic module conforming to the specifications of modularity proposed by Fodor (105). The phonetic module is proposed to have evolved for the special purpose of extracting intended gestures preemptively (101). That is, the phonetic module gets first pass at the incoming acoustic signal and extracts the relevant phonetic gestures passing the residue on for general auditory processing (106).

A variety of experiments showing that speech is processed differently from nonspeech provide evidence for a neural specialization for speech perception. Some of these findings have subsequently been shown to apply equally as well to nonspeech stimuli (5). Even though some evidence for the specialness of speech still stands, it is uncertain whether appropriate nonspeech controls to compare to speech have been considered. A number of ways of creating complex signals that are more or less acoustically equivalent to speech have been considered; however, these experiments do not explore

whether there are controls that are communicatively or informationally equivalent to speech.

A good example of the importance of testing the evidence with informationally equivalent stimuli can be found in a phenomenon known as duplex perception (107) which has been cited frequently as strong evidence for a speech-specific module (1,13). To elicit duplex perception, two stimuli are presented dichotically to a listener wearing headphones. An isolated third formant transition, which sounds like a chirp, is presented in one ear while the base syllable, which is ambiguous because it has had the third formant transition removed, is presented in the other ear. The isolated formant transition fuses with the base syllable, which is then heard as an unambiguous syllable in the base ear. Additionally, the chirp is perceived separately in the other ear. Duplex perception was found to occur with speech stimuli but not with acoustically equivalent stimuli. However, the informational equivalence of the stimuli was brought into question by Fowler and Rosenblum (108) who found that a natural sound, the sound of a door slamming, patterned more like speech in a duplex perception task, and differently from laboratory generated nonspeech controls (which are complex artificial sound patterns). A door slam is ecologically relevant because it gives the hearer information about an action that has occurred in the world (106). Speech has tremendous social significance and is probably the most highly practiced complex perceptual task performed by humans. These factors have not been adequately considered when explaining differences between speech and nonspeech perception.

A claim of the original formulation of the motor theory was that the percepts of speech are not the acoustic signals that impinge directly upon the ear but rather the articulations made by the speaker. One of the striking findings from early experiments was that there are discontinuities in the acoustic-to-phonemic mapping for stop onset consonants (109). These discontinuities were taken as crucial evidence against an acoustic basis for phoneme categories. However, researchers have found that for some phonemic categories the acoustic mapping is simple whereas the articulatory mapping is complex. For example, American English /r/ can be produced with one or more of three distinct gestures, and there is intraspeaker variation in which different gestures are used (75).

The search for first-order acoustic invariance in speech has been largely unsuccessful, and it is now well known that the articulatory gestures and even their motor commands are not invariant either (110). In the revised motor theory, the articulatory percepts are assumed to be the speaker's intended gestures, before contextual adjustments and other sources of speaker-independent variability in production (13). Thus, in terms of the nature of neural representations, the motor theory's proposed linguistic representations are extremely abstract, canonical symbolic entities that can be treated as formally equivalent to abstract phonetic segments. Because neither acoustic nor articulatory categories provide simple dimensions upon which to base perceptual categories in speech, the coherence categories of these abstractions can be based on either articulatory or acoustic properties, or both.

There are several appealing aspects of the motor theory of speech perception. It places the study of speech perception in an ecological context by linking production and perception aspects of spoken language. It also accounts for a wide variety of empirical findings in a principled and consistent manner. For example, the McGurk effect can be nicely accommodated by a model that is based on perception of gestures, although direct perception (15,108) and FLMP (111) also incorporate visual information but in very different ways. Despite the appeal of the motor theory, there remain several serious shortcomings. The proposed perceptual mechanisms remain highly abstract, making effective empirical tests of the model difficult to design. A more explicit model of how listeners extract the intended gestures of other talkers would go far to remedy this problem. In addition, the abstract nature of the intended gestures involves a great deal of reduction of information and therefore suffers from the same shortcomings that traditional phonemic reduction does: it throws away much of the inter- and intratalker variability, which is a rich source of information to the listener.

Direct-Realist Approach. The direct-realist approach to speech perception (15,108) draws on Gibson's (112) ecological approach to visual perception. Its basic assumption is that speech perception, like all other types of perception, acts directly on events in the perceiver's environment rather than on the sensory stimuli and takes place without the mediation of cognitive processes. An event may be described in many ways but those that are ecologically relevant to the perceiver are termed distal events. The sets of possibilities for interaction with them are referred to as affordances. The distal event imparts structure to an informational medium, the acoustic signal and reflected light in the case of visible speech, which in turn provides information about the event to the perceiver by imparting some of its structure to the sense organs through stimulation. The perceiver actively seeks out information about events in the environment, selectively attending to aspects of the environmental structure.

In speech, the phonetically determined coordinated set of movements of the vocal tract that produce the speech signal is made up of those events that the perceiver is attending to. In this way, the direct-realist approach is like the motor theory. However, rather than assuming a speech-specific module retrieving intended gestures from an impoverished acoustic signal, the direct-realist approach assumes an information-rich signal in which the phonetic events are fully and uniquely specified. Because the perception is direct, the direct-realist approach views variability and nonlinearity in a different light than most other approaches to speech perception, which are abstractionist in nature.

The vocal tract cannot produce a string of static and nonoverlapping shapes, so the gestures of speech cannot take place in isolation of each other. Direct perception of gestures gives the listener detailed information about both the gestural and environmental context. This implies that the perceiver is highly experienced with the signal, and so long as that variation is meaningful, it provides information about the event. Rather than remove noise through a process of normalization, variation provides the perceiver with detailed information about the event that includes the talker's size, gender, dialect region, emotional state, as well as prosodic and syntactic information. Therefore, according to this view, stimulus variation ceases to be a problem of perception and becomes a problem of perceptual organization. While direct perception focuses on the perceived events as gestural constellations roughly equivalent to the phoneme, it is also compatible with the theory to assume the perceived events

are words. Thus, we might also consider direct perception as a model of spoken word recognition. Direct perception shares with the exemplar models (discussed in the word recognition section) the assumption that the variability in the signal is rich in information, which is critical to perception.

Direct perception is appealing because of its ability to incorporate and use stimulus variability in the signal, and because it makes the link between production and perception transparent. However, several important theoretical issues remain unresolved. One potential problem for a model that permits no mediation of cognitive processes are top-down influences on speech perception. As was noted previously, these effects are extremely robust and include phoneme restoration (113), correction of errors in shadowing (19), mishearings (114), lexical bias (33), syntactic and semantic bias (115), and lexical frequency and density bias. Fowler (15) acknowledges this problem and suggests that there may be special mechanisms for highly learned or automatic behavior and for perceivers hypothesizing information that is not detected in the signal. She suggests that even though perception itself must be direct, behavior may often not be directed by perceived affordances. In this way, the direct-realist perspective departs dramatically from other versions of event perception (112), which have nothing to say about cognitive mediation.

Finally, Remez (116) notes that it is not clear that the perceptual objects in linguistic communication are the gestures that create the acoustic signal. Even though visual perception of most objects is unambiguous, speech gestures are very different in terms of their perceptual availability. Fowler proposes that the perception of the articulatory gestural complex is the object of perception, but the articulations themselves are a medium that is shaped by the intended linguistic message. As she notes herself, even though visually identified objects are perceived as such, listener's intuitions are that they perceive spoken language not as a series of sound-producing actions (i.e., gestures) but as a sequence of words and ideas. This difference is not necessarily a problem for the model itself, but it is a problem for the way this approach has thus far been employed.

FLMP. A radically different approach to speech perception is represented by informational models that are built around general cognitive and perceptual processes. Most of these models have been developed to explain phonemic perception, and they typically involve multiple processing stages. One example of this approach is the fuzzy logic model of perception, (FLMP) (111). FLMP was developed to address the problem of integrating information from multiple sources, such as visual and auditory input, in making segmental decisions. The criterion for perception of a particular set of features as a particular perceptual unit such as the phoneme is goodness of the percept's match to a subjectively derived prototype description in memory, arrived at through experience with the language of the listener. In acoustic processing, the speech signal undergoes an acoustic analysis by the peripheral auditory system. Evidence for phonemic features in the signal are evaluated by feature detectors using continuous truth values between 1 and 0. Then feature values are integrated and matched against the possible candidate prototypes. Because fuzzy algorithms are used, an absolute match is not needed for the process to achieve a phonemic percept.

Several aspects of FLMP make it an appealing model. First, it provides an explicit mechanism for incorporating multiple sources of information from different modalities. This is particularly important considering the role that visual input can play in the speech perception process. Second, it provides a good fit to data from a wide variety of perceptual experiments (111). Third, it is one of the only models of speech perception that is mathematically explicit, because it is based on a precise mathematical framework. However, there are several serious shortcomings to the model. The most severe deficiency, noted by Klatt (2) and others, is that it is unclear that the fuzzy values are flexible enough to account for the variation that is observed in the speech signal. Because the model works with features to be matched to stored prototypes in memory, there is still a reliance on exclusivity of invariant features and the dependence of features on a degree of normalization across the many sources of variability observed in conversational speech. Moreover, the model has no connection to the perception–production link. Finally, FLMP employs a large number of free parameters that are deduced from the data of specific experimental paradigms but that do not transfer well across paradigms.

Models of Spoken Word Recognition

Models of spoken word recognition can be broken down into two types: those that act on a phonemic or broad phonetic representations and those that work directly on the acoustic input. Models based on a phonemic level are inspired, or transparently derived, from models of alphabetic reading. Because these models use a unitized input, they explicitly or implicitly assume access to a phonemic or featural representation. These models require either an additional preprocessor which recognizes phonemes, segments, or features, or they assume direct perception of these units from information in the acoustic signal. Models that work on segmental or featural input are by far the most numerous and best known, and only a few that are representative of the diversity of proposals will be discussed here. These are TRACE, NAM, and SHORTLIST. Models that act on the speech signal, or an auditory transformation thereof, necessarily incorporate the speech perception process into the word recognition process. Of the few models of this type, two examples will be discussed: LAFS and exemplar-based models.

The Cohort Model

One of the most influential models of spoken word recognition is the Cohort model (19). In this model, the process of word recognition begins with the activation of a "word-initial cohort" based on the first segment or segments of the target word. As more segments are identified, the cohort dwindles as fewer and fewer candidates match the available information. A crucial difference between words in the cohort model is their so-called recognition point, the point in the input where the target word is the only remaining candidate. The recognition point has been shown to affect performance in a variety of tasks (116a).

The Cohort model deserves a great deal of credit for emphasizing the temporal unfolding of speech and its influence on word recognition. However, it has been shown that word recognition can be achieved when word initial information is unavailable (115) and that the dependence of recognition on

the uniqueness point may be task dependent (116b). Thus, the hypotheses concerning the strict temporal processing of lexical information in the Cohort model are too strong.

TRACE. The TRACE model (117,118) is an example of an interactive activation/competition connectionist model. The most widely discussed version of TRACE takes allophonic-level features as their input. An early form of the model took the speech signal as its input and relied on feature detectors to extract relevant information; however, this version was quite limited, being built around only nine CV syllables produced by a single talker.

TRACE is constructed of three levels representing features, phonemes, and words. The featural level passes activation to the phonemic level, which in turn passes activation to the word level. Within each level, the functional units are highly interconnected nodes each with a current activation level, a resting level, and an activation threshold. There are bidirectional connections between units of different levels and between nodes within a level. Connections are excitatory between units at different levels that share common properties (e.g., among voice, place, and manner features and a particular consonantal phoneme). Connections among units within a level may be inhibitory; for example, as one place feature at one time slice is activated, it will inhibit the activation of other place features. Connections among units within a level may also be excitatory; for example, a stop consonant at one time slice will facilitate segments that can precede or follow it, such as /s/ or a vowel (depending on the phonotactic constraints of the language).

The excitatory and inhibitory links in TRACE have important implications for the types of processing operations within the model. Because of the inhibitory links within a level, TRACE acts in a winner-takes-all fashion. Moreover, the excitatory links provides a mechanism for the contribution of top-down information to the perception of speech sounds. TRACE contrasts with traditional symbolic invariance approaches because it treats coarticulatory variation as a source of information rather than a source of noise; the inhibitory and facilatory links between one time slice and the next allow for adjacent segments to adjust the weighting to a particular feature or phoneme in a given context.

Despite these advantages, there are two major problems with TRACE. The first is that, although it can use the coarticulatory variation in segmental contexts as information, it is unclear how the model would incorporate other sources of lawful variation such as prosody, rate, or differences among talkers. The second is that TRACE's multiple instantiations of the network across time are considered to be neurally and cognitively implausible (119). More recent connectionist models have proposed recurrent neural networks as a way of representing the temporal nature of speech (118).

Connectionist models such as TRACE are similar to FLMP because they rely on continuous rather than discrete representations. Continuous activation levels allow for varying degrees of support for competing perceptual hypotheses. Connectionist models also allow for the evaluation and integration of multiple sources of input and rely on general-purpose pattern-matching schemes with a best-fit algorithm. But the connectionist models and the FLMP differ in the degree to which top-down influences can affect low-level processes. Massaro (111) claims that connectionist models that have top-down and bottom-up connections are too powerful, predicting both attested and unattested results. Massaro argues that FLMP allows top-down *bias* in the perception process, whereas TRACE's two-way connections result in top-down-induced changes in perceptual sensitivity. This is an open issue in need of further research.

The Neighborhood Activation Model. The neighborhood activation model, or NAM (96) shares with TRACE the notion that words are recognized in the context of other words. A pool of word candidates is activated by acoustic/phonetic input. However, the pool of activated candidates is drawn from the similarity neighborhood of the word. A similarity neighborhood is the set of words that is phonetically similar to the target word. Relevant characteristics of the similarity neighborhood are its density and neighborhood frequency. The density of a word is the number of words in a neighborhood. The neighborhood frequency of a word is the average frequency of words in the neighborhood. There is a strong frequency bias in the model which allows it to deal with apparent top-down word frequency effects without positing explicit bidirectional links. Rather than unfolding over time, similarity in NAM is a static property of the entire word. NAM is least developed as a general model of word recognition because it assumes not only a phonemic level but word segmentation as well (see Ref. 120 for a revisions that resolve some of these problems). Moreover, NAM has been implemented only for monosyllabic words. NAM can account for a specific set of lexical similarity effects not treated in other models and is attractive because it is grounded in a more general categorization model based on the Probability Choice Rule (121).

SHORTLIST. SHORTLIST (122) parses a phonemic string into a set of lexical candidates, which compete for recognition. SHORTLIST can be seen as an evolutionary combination of both the TRACE and Marslen-Wilson's Cohort model (19). A small set (the shortlist) of lexical candidates compete in a TRACE-style activation/competition network. The phonemic string is presented gradually to the model, but candidates with early matches to the string have an advantage because of their early activation, much like the original cohort model. However, as Cutler (119) notes, SHORTLIST avoids the cognitive implausibility of TRACE's temporal architecture, which effectively duplicates the network at each time slice. SHORTLIST also avoids the cohort models overdependence on word initial information. The model takes phonemic information as its input and strictly bottom-up information determines the initial candidate set. The candidate set is determined by comparing whole words but with each strong (i.e., stressed) syllable acting as a potential word onset. This use of prosodic information sets this model apart from others and gives SHORTLIST the ability to parse words from a phrase or utterance represented as a string of phonemes and allophones.

LAFS. Lexical Access From Spectra (2), or LAFS, is a purely bottom-up model of word recognition, which compares the frequency spectrum of the incoming signal to stored templates of frequency spectra of words. The stored templates are context-sensitive spectral prototypes derived from subjective experience with the language and consist of all possible diphone (CV and VC) sequences and all cross-word boundaries in the language, resulting in a very large decoding network.

Thus, LAFS addresses the problems of contextual variability by precompiling the coarticulatory and word boundary variations into stored representations in an integrated memory system. The model attempts to address interspeaker and rate-based variability by using a best-fit algorithm to match incoming spectra with stored spectral templates. LAFS fully bypasses the intermediary featural and segmental stages of processing; the perceptual process consists of finding the best match between the incoming spectra and paths through the network.

The advantages of such a strategy are numerous and have been discussed in detail by Klatt (2). Rather than discarding allophonic and speaker-specific detail through reduction to an abstract symbolic representation such as features or segments, the input spectra are retained in full detail. This frees the model from dealing with problems of acoustic invariance across contexts. LAFS does not make segmental phonemic-level decisions because it performs recognition at the word level; consequently, there is less data reduction than in traditional phonemic-based models. More information can be brought to bear on the lexical decision, thereby reducing the probability of error and increasing the ability of the system to recover from error (122a). Because the stored prototypes are based on subjective learning, there can be local tuning, and there is less chance of overgeneralization. The perceptual process is explicit, being based on a distance/probability metric (123) and the scoring strategy is uniform throughout the network.

Despite the power of the approach, there are several problems with the LAFS strategy that Klatt (2) acknowledges and some that have been raised since then. The most serious is that, even though LAFS is constructed to accommodate coarticulatory and word-edge variability, it is unlikely that the distance metric is powerful enough to accommodate the full range of variability seen in spoken language. Furthermore, it is nearly impossible to preprocess and store in memory all the sources of variability cited in the preceding variability section. Finally, much of the stimulus variability in speech comes not in spectra alone but in timing differences as well (Klatt cites the example of variable onset of prenasalization), and LAFS is not built to accommodate much of the temporal nature of speech variation. LAFS is obviously constructed to model a fully developed adult's perception process and contains some developmentally implausible assumptions (but see Ref. 124 for a developmentally oriented adaptation of LAFS called WRAPSA). Its structure involves *a priori* knowledge about all possible diphones in the language and all cross-word boundary combinations; different languages have varying inventories of speech sounds and different constraints on how these sounds can combine within and across words, yet the model depends on these being precompiled for speech perception and word identification to proceed. Cutler (119) notes that the redundancy inherent in precompiling all word boundaries for every possible word pair separately is psychologically implausible. In addition, recent phonetic research has found different boundary effects at multiple levels of the prosodic hierarchy. Requiring precompiled boundaries at the foot, intonation phrase, and possibly other levels adds to the psychological implausibility of the model. Finally, because the model is explicitly bottom-up, it cannot properly model the top-down factors like lexical, prosodic, and semantic bias in the lexical decisions.

Exemplar-Based Models of Word Recognition. Like LAFS, exemplar-based models bypass the reduction of the speech signal to featural and segmental units in the word identification process. However, unlike LAFS, exemplar models are instance-based; they do not rely on precompiled prototypes stored in memory. In exemplar models, there are no abstract categories (whether learned prototypes or innate features and phonemes). Instead, the set of all experienced instances of a category form the basis for the category. The process of categorization, therefore, involves computing the similarity of the stimulus to every stored instance of every category (125). Although this type of model behaves as if it works on idealized prototype categories, categorization is a result of computations and the decision process rather than stored prototypes of the stimulus. Exemplar models of perception and memory are fairly widespread in cognitive psychology, but they have only rarely been applied to speech perception and spoken word recognition (for further background on exemplar models in speech perception, see Ref. 72).

As discussed at the beginning of this chapter, one of the motivations for proposing that the speech signal is reduced to abstract categories such as phonemes and features has been the widespread belief that memory and processing limitations necessitate data reduction. However, more recent empirical data suggest that earlier theorists largely overestimated memory and processing limitations. There is now ample evidence in speech perception and word recognition literature of long-term preservation of instance-specific details about the acoustic signal. Goldinger (72) discusses in detail the motivation for exemplar models and cites evidence of episodic memory for such language-relevant cases as faces, physical dynamics, modality of presentation, exact wording of sentences, and talker-specific information in spoken words. Taken together, this recent evidence has inspired some psycholinguists and speech researchers to reconsider exemplar based approaches to speech perception and spoken word recognition. Although limited, one of the more explicitly implemented models has been proposed by Johnson (16), which is based on Kruschke's (126) connectionist ALCOVE model.

In a description of his model, Johnson (16) discusses several potential problems that an exemplar approach to speech perception must address to be a realistic model of human spoken word recognition. Because most exemplar models have been developed for the perception of static images, these models must be revised and elaborated upon to take into account the time-varying nature of spoken language. This problem is addressed by considering the role of short-term auditory memory in the processing of speech. The incoming signal is sliced into auditory vectors in both the frequency and time domains. As the signal is processed and encoded, the spectral vectors in the short-term auditory buffer are matched to all stored vectors, and matches are activated adding to previous activation levels thereby representing a veridical short-term memory of the signal. Introducing time in this way, permits the modeling of temporal selective attention and segmentation strategies. For example, language specific segmentation strategies such as those that inspired the SHORTLIST model might be modeled by probing the matrix cyclically for boundary-associated acoustic events.

A second problem that must be addressed if exemplar models are to be considered cognitively plausible is that of memory limitations. Although there is now a great deal of evidence

that much fine detail about specific instances of spoken language is encoded and stored in memory, it is implausible that each experienced auditory pattern is stored at a separate location in the brain or that these instances can be retrieved. Both Goldinger (72) and Johnson (16) use vector abstractions of spoken words to reduce the amount of stored data. Vector representations also represent the encoding of words in a multidimensional similarity space, a feature common to many other models of word recognition. The explicit use of a similarity space also provides another level of data reduction. Different words which have the same value on some dimensions are combined in that location. Johnson's model uses a few dimensions that are tied to the acoustic–phonetic characteristics of the words. Thus, this model is an exemplar-based model which focuses on speech perception. Goldinger's model uses more dimensions, which include segmental, semantic, voice, and context information at a more abstract level. Goldinger's model thus focuses more on spoken word recognition.

Top-down influences on speech perception pose problems for fully bottom-up processing models. It might seem that an exemplar model would leave no room for lexical or semantic bias in the decision process. However, usage frequency, recency, and contextual factors can be modeled with base-activation levels and attention weights. For example, a high-frequency lexical item would have a high base-activation level that is directly tied to the frequency of occurrence with a time decay factor. Because syntactic and semantic conditions increase the predictability of a set of words, the base activation rises. Attention weights can be adjusted to model selective attention exhibited by the shrinking and expanding of the perceptual space (125). One example of such perceptual distortion is the *perceptual magnet effect* where the perceptual space appears as if it is warped by prototypes (127), resulting in decreased sensitivity to changes along a dimension within the range variation of a particular category but increased sensitivity across a category boundary.

Finally, Johnson suggests that exemplar models are also capable of incorporating the production–perception link. As a talker produces an utterance, he/she is also hearing the utterance. Therefore, the set of auditory memory traces that are specific to words produced by the talker can be linked to a set of equivalent sensory-motoric exemplars or articulatory plans.

Like the direct perception approach, exemplar-based models of perception bring a radical change from past assumptions. Instead of treating stimulus variation in speech as noise to be removed to aid in the perception of abstract units of speech, variability is treated as inherent to the way experiences with speech are stored in memory. Therefore, variation is a source of information that may be used by the perceiver depending on the demands of the listening situation. The appeal of this approach is its ability to account for a very wide variety of speech perception phenomena in a consistent and principled manner. The extensive work on exemplar modeling in other areas of perception means that, unlike much of the traditional work in speech perception, the fundamentals of the approach have been worked out explicitly. However, because the approach is in its infancy in the field of speech perception, it remains to be seen how it performs when tested in a more rigorous fashion across many different environments.

CONCLUSION

Research on human speech perception has shown that the perceptual process is highly complex in ways beyond our current understanding or theoretical tools. Speech perception relies on both visual and auditory information which are integrated as part of the perceptual process. Near-perfect performance is achieved despite an enormous amount of variability both within and across talkers and across a wide variety of different environmental conditions. In the early days of speech research, it was believed that the perceptual process relied on a few invariant characteristics of the segments that differentiated larger linguistic units like words and utterances. Even though we may yet find higher-order relational invariants that are important features for defining the linguistic categories in language, it has already been demonstrated that listeners use the lawful variability in the acoustic signal when perceiving speech. Variability cannot be removed, discarded, or normalized away in any psychologically plausible model of speech perception. Some of the approaches discussed here, which rely on more elaborate notions of perceptual categories and long-term encoding, incorporate the variability and nonlinearity inherent in the speech signal directly into the perceptual process. These new approaches to speech perception treat the speech signal as information-rich and use lawful variability and redundant information rather than treat these properties of speech as extraneous noise to be discarded. We believe that these new approaches to the traditional problems of invariance and nonlinearity provide a solution to the previously intractable problem of perceptual constancy despite variability in the signal.

BIBLIOGRAPHY

1. M. Studdert-Kennedy, Speech perception, in N. J. Lass (ed.), *Contemporary Issues in Experimental Linguistics,* New York: Academic Press, 1976, pp. 213–293.
2. D. H. Klatt, Review of selected models of speech perception, in W D. Marslen-Wilson (ed.), *Lexical Representation and Process,* Cambridge, MA: MIT Press, 1989, pp. 169–226.
3. J. L. Miller, Speech perception, in D. N. Osherson and H. Lasnik (eds.), *An Invitation to Cognitive Science,* Cambridge, MA: MIT Press, 1990, pp. 69–93.
4. L. C. Nygaard and D. B. Pisoni, Speech perception: New directions in research and theory, in J. L. Miller and P. D. Eimas (eds.), *Speech, Language, and Communication,* San Diego: Academic Press, 1995, pp. 63–96.
5. S. D. Goldinger, D. B. Pisoni, and P. A. Luce, Speech perception and spoken word recognition: Research and theory, in N. J. Lass (ed.), *Principles in Experimental Phonetics.* St. Louis, MO: Mosby, 1996, pp. 277–327.
6. T. M. Neary, Speech perception as pattern recognition, *J. Acoust. Soc. Amer.,* **101**: 3241–3254, 1997.
7. P. Lieberman, Some effects of semantic and grammatical context on the production and perception of speech, *Lang. Speech,* **6**: 172–187, 1963.
8. C. A. Fowler and J. Housum, Talkers signalling of new and old words in speech and listeners perception and use of the distinction, *Memory Lang.,* **26**: 489–504, 1987.
9. B. Lindblom, Explaning phonetic variation: A sketch of the H and H theory, in W. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling,* Dordrecht: Kluwer, 1990, pp. 403–439.

10. A. M. Liberman, Some results of research on speech perception, *J. Acoust. Soc. Amer.,* **29**: 117–123, 1957.
11. N. Chomsky and G. A. Miller, Introduction to the formal analysis of natural language, in R. D. Luce, R. Bush, and E. Galanter (eds.), *Handbook of Mathematical Psychology,* New York: Wiley, 1963, pp. 269–321.
12. S. E. Blumstein and K. N. Stevens, Perceptual invariance and onset spectra for stop consonants in different vowel environments, *J. Acoust. Soc. Amer.,* **66**: 1001–1017, 1980.
13. A. M. Liberman and I. G. Mattingly, The motor theory of speech perception revised, *Cognition,* **21**: 1–36, 1985.
14. P. Ladefoged and D. E. Broadbent, Information conveyed by vowels, *J. Acoust. Soc. Amer.,* **29**: 948, 1957.
15. C. A. Fowler, An event approach to the study of speech perception from a direct-realist perspective, *J. Phonetics,* **14**: 3–28, 1986.
16. K. Johnson, Speech perception without speaker normalization, in K. Johnson and J. W. Mullennix (eds.), *Talker Variability in Speech Processing,* San Diego: Academic Press, 1997, pp. 145–165.
17. K. I. Forster, Accessing the mental lexicon, in R. J. Wales and E. Walker (eds.), *New Approaches to Language Mechanisms,* Amsterdam: North-Holland, 1976.
18. A. G. Samuel, Phonetic prototypes, *Percept. Psychophys.,* **31**: 307–314, 1982.
19. W. D. Marslen-Wilson and A. Welsh, Processing interactions during word-recognition in continuous speech, *Cogn. Psychol.,* **10**: 29–63, 1978.
20. A. Cutler and D. Norris, The role of strong syllables in segmentation for lexical access, *J. Exp. Psychol.: Hum. Percept. Perform.,* **14**: 381–410, 1988.
21. R. Jacobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis,* Cambridge, MA: MIT Acoustics Laboratory, 1952.
22. M. Halle, Speculations about the representation of words in memory, in V. Fromkin (ed.), *Phonetic Linguistics,* New York: Academic Press, 1985, pp. 101–114.
23. C. P. Browman and L. Goldstein, Gestural specification using dynamically-defined articulatory structures, *J. Phonetics,* **18**: 299–320, 1990.
24. J. A. Goldsmith, *Autosegmental & Metrical Phonology.* Oxford, UK: Blackwell, 1990.
25. D. E. Broadbent, Information processing in the nervous system, *Science,* **150**: 475–462, 1965.
26. C. D. Creelman, The case of the unknown talker, *J. Acoust. Soc. Amer.,* **29**: 655, 1957.
27. P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages.* Oxford, UK: Blackwell, 1996.
28. C. Hockett, *Manual of Phonology.* Bloomington: Indiana Univ. Press, 1955.
29. J. Pierrehumbert and M. Beckman, *Japanese Tone Structure,* Cambridge, MA: MIT Press, 1988.
30. A. Cutler, The comparative perspective on spoken-language processing, *Speech Commun.,* **21** (1–2): 3–15, 1997.
31. A. Cutler et al., The syllables differing role in the segmentation of French and English, *J. Memory Lang.,* **25**: 385–400, 1986.
32. T. G. Bever, J. Lackner, and R. Kirk, The underlying structures of sentences are the primary units of immediate speech processing, *Percept. Psychophys.,* **5**: 191–211, 1969.
33. W. F. Ganong, Phonetic categorization in auditory word perception, *J. Exp. Psychol.,* **6**: 110–125, 1980.
34. R. E. Remez, Units of organization and analysis in the perception of speech, in M. E. H. Schouten (ed.), *The Psychophysics of Speech Perception,* Dordrecht: Martinus Nijhoff, 1987, pp. 419–432.
35. G. Fant, *Acoustic Theory of Speech Production,* The Hague: Mouton, 1960.
36. A. K. Syrdal and H. S. Gopal, A perceptual model of vowel recognition based on the auditory representation of American English vowels, *J. Acoust. Soc. Amer.,* **79**: 1086–1100, 1986.
37. L. Gerstman, Classification of self-normalized vowels, *IEEE Trans. Audio Electroacoust.,* **AU-16**: 78–80, 1968.
38. K. N. Stevens and A. S. House, Perturbation of vowel articulations by consonantal context: An acoustic study, *J. Speech Hear. Res.,* **6**: 111–128, 1963.
39. B. Lindblom and M. Studdert-Kennedy, On the role of formant transitions in vowel recognition, *J. Acoust. Soc. Amer.,* **42**: 830–843, 1967.
40. I. Maddieson, *Patterns of Sound,* Cambridge, UK: Cambridge Univ. Press, 1984.
41. I. Lehiste, *Suprasegmentals,* Cambridge, MA: MIT Press, 1970.
42. T. J. Hudak, Thai, in B. Comrie (ed.), *The World's Major Languages,* Oxford, UK: Oxford Univ. Press, 1987, pp. 757–776.
43. J. D. McCawley, *The Phonological Component of the Japanese Grammar,* The Hague: Mouton, 1968.
44. P. Ladefoged, I. Maddieson, and M. T. T. Jackson, Investigating phonation types in different languages, in O. Fujimura (ed.), *Vocal Physiology: Voice Production, Mechanisms and Functions,* New York: Raven Press, 1988, pp. 297–317.
45. P. C. Delattre, A. M. Liberman, and F. S. Cooper, Acoustic loci and transitional cues for consonants, *J. Acoust. Soc. Amer.,* **27**: 769–773, 1955.
46. J. D. O'Connor et al., Acoustic cues for the perception of initial / w, j, r, l/ in English, *Word,* **13**: 22–43, 1957.
47. R. Wright, *Consonant clusters and cue preservation in Tsou,* Ph.D. dissertation, Univ. of California at Los Angeles, 1996.
48. O. Fujimura, M. J. Macchi, and L. A. Streeter, Perception of stop consonants with conflicting transitional cues: A cross-linguistic study, *Lang. Speech,* **21**: 337–345, 1978.
49. A. van Weiringen, Perceiving dynamic speechlike sounds, Ph.D. dissertation, Univ. of Amsterdam, 1995.
50. C. Schadle, *The Acoustics of Fricative Consonants,* Cambridge, MA: MIT Press, 1985.
51. J. M. Heinz and K. N. Stevens, On the properties of voiceless fricative consonants, *J. Acoust. Soc. Amer.,* **33**: 589–596, 1961.
52. K. S. Harris, Cues for the discrimination of American English fricatives in spoken syllables, *Lang. Speech,* **1**: 1–7, 1958.
53. G. A. Miller and P. E. Nicely, An analysis of perceptual confusions among some English consonants, *J. Acoust. Soc. Amer.,* **27**: 329–335, 1955.
54. P. J. Bailey and Q. Summerfield, Information in speech: Observations on the perception of [s]-stop clusters, *J. Exp. Psychol.,* **6**: 536–563, 1980.
55. D. Kewley-Port, D. B. Pisoni, and M. Studdert-Kennedy, Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants, *J. Acoust. Soc. Amer.,* **73**: 1779–1793, 1983.
56. A. Walley and T. Carrell, Onset spectra and formant transitions in the adult's and child's perception of place of articulation in initial stop consonants, *J. Acoust. Soc. Amer.,* **73**: 1011–1022, 1983.
57. K. Kurowski and S. E. Blumstein, Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants, *J. Acoust. Soc. Amer.,* **76**: 383–390, 1984.
58. A. S. House, Analog studies of nasal consonants, *J. Speech Hear. Res.,* **22**: 190–204, 1957.
59. A. Malécot, Acoustic cues for nasal consonants, *Language,* **32**: 274–278, 1956.

60. S. Hawkins and K. N. Stevens, Acoustic and perceptual correlates of the nonnasal-nasal distinction for vowels, *J. Acoust. Soc. Amer.,* **77**: 1560–1575, 1985.
61. A. M. Liberman et al., Tempo of frequency change as a cue for distinguishing classes of speech sounds, *Psychol. Monogr. (Gen. Appl.),* **68**: 1–13, 1956.
62. P. Shinn and S. E. Blumstein, On the role of the amplitude envelope for the perception of [b] and [w], *J. Acoust. Soc. Amer.,* **75**: 1243–1252, 1984.
63. L. Lisker and A. D. Abramson, A cross-language study of voicing in initial stops: Acoustic measurements, *Word,* **20**: 384–422, 1964.
64. B. H. Repp, Relative amplitue of aspiration noise as a voicing cue for syllable-initial stop consonants, *Lang. Speech,* **22**: 173–189, 1979.
65. S. D. Soli, Structure and duration of vowels together specify fricative voicing, *J. Acoust. Soc. Amer.,* **72**: 366–378, 1982.

65a. D. W. Massaro and M. M. Cohen, Evaluation and integration of visual and auditory information in speech perception, *J. Exper. Psychol.: Human Perception Performance,* **9**: 751–753, 1983.

66. W. H. Sumby and I. Pollack, Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Amer.,* **26**: 212–215, 1954.
67. H. McGurk and J. MacDonald, Hearing lips and seeing voices, *Nature,* **264**: 746–748, 1976.

67a. G. Fant, Descriptive analysis of the acoustic aspects of speech, *Logos,* **5**: 3–17, 1962.

68. G. N. Clements, Principles of tone assignment in Kikuyu, in G. N. Clements and J. Goldsmith (eds.), *Autosegmental Studies in Bantu Tone.* Dordrecht: Foris, 1984.
69. D. Nguyen, Vietnamese, in B. Comrie (ed.), *The World's Major Languages,* Oxford, UK: Oxford Univ. Press, 1987, pp. 777–796.
70. Q. Summerfield, On articulatory rate and perceptual constancy in phonetic perception, *J. Exp. Psychol.: Hum. Percept. Perform.,* **7**: 1074–1095, 1981.
71. J. L. Miller, Rate-dependent processing in speech perception, in A. Ellis (ed.), *Progress in the Psychology of Language.* Hillsdale, NJ: Erlbaum, 1987.
72. S. D. Goldinger, Words and voices: Perception and production in an episodic lexicon, in K. Johnson and J. W. Mullennix (eds.), *Talker Variability in Speech Processing,* San Diego: Academic Press, 1997, pp. 33–66.
73. K. N. Stevens, Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds, in A. Rigault and R. Charbonneau (eds.), *Proc. 7th Int. Congr. Phonetic Sci.,* The Hague: Mouton, 1972, pp. 206–232.
74. K. Johnson, P. Ladefoged, and M. Lindau, Individual differences in vowel production, *J. Acoust. Soc. Amer.,* **94**, 701–714, 1993.
75. R. Hagiwara, Acoustic realizations of American /r/ as produced by women and men. Ph.D. dissertation, Univ. of California at Los Angeles, 1995.
76. J. L. Miller and A. M. Liberman, Some effects of later occurring information on the perception of stop consonant and semivowel, *Percept. Psychophys.,* **25**: 457–465, 1979.
77. M. S. Sommers, L. C. Nygaard, and D. B. Pisoni, The effects of speaking rate and amplitude variability on perceptual identification, *J. Acoust. Soc. Amer.,* **91**: 2340, 1992.
78. M. Beckman and J. Edwards, Articulatory evidence for differentiating stress categories, in P. A. Keating (ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III,* Cambridge, UK: Cambridge Univ. Press, 1994, pp. 7–33.
79. K. de Jong, The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation, *J. Acoust. Soc. Amer.,* **97**: 491–504, 1995.
80. C. Fougeron and P. A. Keating, Articulatory strengthening at edges of prosodic domains, *J. Acoust. Soc. Amer.,* **101**: 3728–3740, 1997.
81. M. Varya and C. A. Fowler, Declination of supralaryngeal gestures in spoken Italian, *Phonetica,* **49**: 48–60, 1992.
82. I. Lehiste, Role of duration in disambiguating syntactically ambiguous senences, *J. Acoust. Soc. Amer.,* **60**: 1199–1202, 1976.
83. E. Lombard, Le signe de l'élévation de la voix, *Ann. Mal. Oreille, Larynx, Nez Pharynx,* **37**: 101–119, 1911.
84. H. Lane and B. Tranel, The Lombard sign and the role of hearing in speech, *J. Speech Hear. Res.,* **14**: 677–709, 1971.
85. A. H. Anderson et al., Limited visual control of the intelligibility of speech in face-to-face dialogue, *Percept. Psychophys.,* **59** (4): 580–592, 1997.
86. R. A. W. Bladon, C. G. Henton, and J. B. Pickering, Towards an auditory theory of speaker normalization, *Lang. Commun.,* **4**: 59–69, 1984.
87. J. E. Flege and K. P. Massey, English prevoicing: Random or controlled, paper presented at the Linguistic Society of America, Albuquerque, NM, 1980.
88. M. F. Schwartz, Identification of speaker sex from isolated, voiceless fricatives, *J. Acoust. Soc. Amer.,* **43**: 1178–1179, 1968.
89. D. H. Klatt and L. C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers, *J. Acoust. Soc. Amer.,* **87**: 820–857, 1990.
90. P. Ladefoged, *Three Areas of Experimental Phonetics,* Oxford, UK: Oxford Univ. Press, 1968.
91. G. E. Peterson and H. L. Barney, 1952.
92. J. W. Black and H. M. Mason, Training for voice communication, *J. Acoust. Soc. Amer.,* **18**: 441–445, 1946.
93. H. Fletcher, *Speech and Hearing.* Princeton, NJ: Van Nostrand-Reinhold, 1929.
94. N. R. French and J. C. Steinberg, Factors governing the intelligibility of speech sounds, *J. Acoust. Soc. Amer.,* **19**: 90–119, 1947.
95. G. A. Miller, G. A. Heise, and W. Lichten, The intelligibility of speech as a function of the context of the test materials, *J. Exp. Psychol.,* **16**: 329–335, 1951.
96. P. A. Luce and D. B. Pisoni, Recognizing spoken words: The neighborhood activation model, *Ear Hear.,* **19** (1): 1–36, 1998.
97. F. Jelinek, *Statistical Methods for Speech Recognition.* Cambridge, MA: MIT Press, 1998.
98. R. Cole et al., Speech patterns on paper, in R. Cole (ed.), *Perception and Production of Fluent Speech,* Hillside, NJ: Erlbaum, 1978.
99. B. Lindblom and S. G. Svensson, Interaction between segmental and nonsegmental factors in speech recognition, *IEEE Trans. Audio Electroacoust.,* **AU-21**: 536–545, 1973.
100. A. Cutler, Phoneme-monitoring reaction time as a function of preceding intonation contour. *Percept. Psychophys.,* **20**: 55–60, 1976.
101. A. M. Liberman, *Speech: A Special Code,* Cambridge, MA: MIT Press, 1996.
102. J. C. Steinberg, Effects of distortion on telephone quality, *J. Acoust. Soc. Amer.,* **1**: 121–137, 1929.
103. D. Kewley-Port, Time-varying features as correlates of place of articulation in stop consonants, *J. Acoust. Soc. Amer.,* **72**: 379–389, 1983.
104. M. Mack and S. E. Blumstein, Further evidence of acoustic invariance in speech production: The stop-glide contrast, *J. Acoust. Soc. Amer.,* **73**: 1739–1750, 1983.
105. J. A. Fodor, *The Modularity of Mind,* Cambridge, MA: MIT Press, 1983.

106. W. W. Gaver, What in the world do we hear?: An ecological approach to auditory event perception, *Ecol. Psychol.,* **5** (1): 1–29, 1993.

107. T. C. Rand, Dichotic release from masking for speech, *J. Acoust. Soc. Amer.,* **55**: 678–680, 1974.

108. C. A. Fowler and L. D. Rosenblum, Duplex perception: A comparison of monosyllables and slamming doors, *J. Exp. Psychol.: Hum. Percept. Perform.,* **16** (4): 742–754, 1990.

109. A. M. Liberman, P. C. Delattre, and F. S. Cooper, The role of selected stimulus-variables in the perception of unvoiced stops, *Amer. J. Psychol.,* **65**: 497–516, 1952.

110. P. F. MacNeilage, Motor control of serial ordering of speech, *Psychol. Rev.,* **77**: 182–196, 1970.

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

112. J. J. Gibson, *The Senses Considered as Perceptual Systems.* Boston: Houghton-Mifflin, 1966.

113. R. M. Warren, Perceptual restoration of missing speech sounds, *Science,* **176**: 392–393, 1970.

114. S. Garnes and Z. Bond, A slip of the ear: A snip of the ear? A slip of the year? in V. Fromkin (ed.), *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, Hand,* New York: Academic Press, 1980.

115. A. Salasoo and D. B. Pisoni, Interaction of knowledge sources in spoken word identification, *J. Memory Lang.,* **24**: 210–231, 1985.

116. R. E. Remez, Realism, language, and another barrier, *J. Phonetics,* **14**: 89–97, 1986.

116a. W. D. Marslen-Wilson, Functional parallelism in spoken word-recognition, *Cognition,* **25**: 71–102, 1987.

116b. M. Taft and G. Hambly, Exploring the cohort model of word recognition, *Cognition,* **22**: 259–282, 1986.

117. J. L. McClelland and J. L. Elman, The TRACE model of speech perception, *Cogn. Psychol.,* **18**: 1–86, 1986.

118. J. L. Elman, Connectionist approaches to acoustic/phonetic proessing, in W. D. Marslen-Wilson (ed.), *Lexical Representation and Process,* Cambridge, MA: MIT Press, 1989, pp. 227–260.

119. A. Cutler, Spoken word recognition and production, in J. L. Miller and P. D. Eimas (eds.), *Speech, Language, and Communication,* San Diego: Academic Press, 1995, pp. 97–137.

120. E. T. Auer, Dynamic processing in spoken word recognition: The influence of paradigmatic and syntagmatic states, Ph.D. dissertation, Univ. at Buffalo, Buffalo, NY, 1993.

121. R. D. Luce, *Individual Choice Behavior,* New York: Wiley, 1959.

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

122a. J. L. Miller, Decision units in the perception of speech, *IRE Trans. Inf. Theory,* 81–83, 1962.

123. F. Jelinek, The development of an experimental discrete dictation recognizer, *Proc. IEEE,* **73**: 1616–1624, 1982.

124. P. Jusczyk, *Discovering Spoken Language,* Cambridge, MA: MIT Press, 1997.

125. R. M. Nosofsky, Exemplar-based accounts of relations between classification, recognition, and typicality, *J. Exp. Psychol.: Learn., Memory, Cognition,* **14**: 700–708, 1988.

126. J. K. Kruschke, ALCOVE: An exemplar based connectionist model of category learning, *Psychol. Rev.,* **99**: 22–44, 1992.

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

Reading List

J. Cutting and L. Kozlowski, Recognizing friends by their walk: Gait perception without familiarity cues, *Bull. Psychon. Soc.,* **9**: 353–356, 1977.

D. H. Klatt, Speech perception: A model of acoustic-phonetic analysis and lexical access, *J. Phonetics,* **7**, 279–312, 1979.

T. M. Neary, Context effects in a double-weak theory of speech perception, *Lang. Speech,* **35**: 153–172, 1992.

RICHARD WRIGHT
University of Washington, Seattle

STEFAN FRISCH
University of Michigan, Ann Arbor

DAVID B. PISONI
Indiana University, Bloomington