Chapter 8

Speech Perception and Spoken Word Recognition: Research and Theory

Stephen D. Goldinger, David B. Pisoni, Paul A. Luce

Chapter Outline

Key Terms

INTRODUCTION

The study of speech perception is concerned with the listener's ability to perceive the acoustic waveform produced by a speaker as a string of meaningful words and ideas. By this definition, speech perception has been researched since at least the turn of the century, when one of the earliest empirical studies was published by Bagley (1900-1901; see Cole and Rudnicky, 1983). Bagley's experiments addressed a surprisingly wide variety of topics that have since been rediscovered, including phonemic restoration, semantic priming, importance of word-initial information, and sentence context effects on word recognition. A common theme of Bagley's experiments was their focus on the influence of semantic and lexical knowledge on the perception of distorted words. As Cole and Rudnicky (1983) observed, Bagley anticipated many empirical phenomena and theoretical accounts of speech perception and spoken word recognition that remain central to discussions today.

If Bagley's results and arguments were presented today, they would most likely be considered relevant to language perception rather than to speech perception per se. *Speech perception* has, for a variety of reasons, come to refer more specifically to phoneme perception than to the perception of words or phrases. Unlike a process such as visual perception, in which the recognition of objects or motion is available to the observer, speech perception as the researcher defines it is a process of which we are generally unaware. As Darwin (1976, p. 175) comments, "Our conscious perceptual world is composed of greetings, warnings, questions, and statements; while their vehicle, the segments of speech, goes largely unnoticed and words are subordinated to the framework of the phrase or sentence." Despite the truth of Darwin's observation, the majority of research on speech perception in the past three decades has focused only on the unnoticed vehicle, phonetic perception. This rather myopic approach has resulted in a large theoretical body of literature that is somewhat divorced from more general theories of perception and mainstream cognitive psychology. For example, only recently have serious efforts been applied to model the process of speech perception not as an end in itself but as subservient to word recognition (Pisoni and Luce, 1987). This chapter considers the effects of this segregated research and theorizing on our understanding of speech as the front end of language.

Numerous papers and chapters review the theories and data in speech perception (Studdert-Kennedy, 1974, 1976; Darwin, 1976; Cutting and Pisoni, 1978; Pisoni, 1978; Pisoni and Luce, 1986; Luce and Pisoni, 1987; Miller, 1990). In large part, the fundamental issues in speech perception and the data relevant to those issues have remained unchanged over the past several years. Accordingly, although this chapter will address several fundamental issues in speech perception, it is not a comprehensive review of the empirical literature in the field. Nevertheless, this chapter is fairly eclectic, and we hope to address a sufficiently wide range of topics. We do not marshal evidence for one particular theory or class of theories at the expense of all others; instead, we examine and evaluate a wide range of theories. Finally, throughout the chapter we examine how research and theory in speech perception have or have not developed over the years with respect to the fundamental "problems" of speech. The next section begins with a review of several of the long-standing basic issues in speech perception.

BASIC ISSUES IN SPEECH PERCEPTION

Linearity, Lack of Acoustic-Phonetic Invariance, and the Segmentation Problem

Since the mid-1950s no finding has influenced speech research and theory more profoundly than the failures of the speech signal to satisfy the linearity and invariance conditions. The *linearity* condition assumes that for each perceived phoneme, there must be a particular corresponding stretch of sound in the utterance (Chomsky and Miller, 1963). For example, if the listener perceives that phoneme X occurs before phoneme Y, the stretch of sound associated with phoneme X must precede the stretch of sound associated with phoneme Y in the physical signal. The *invariance* condition assumes that for each phoneme X, a specific set of acoustic correlates must occur in all phonetic contexts. Under these conditions, recognition of phoneme X implies that the features for X occurred in the speech signal in a discrete time window and that

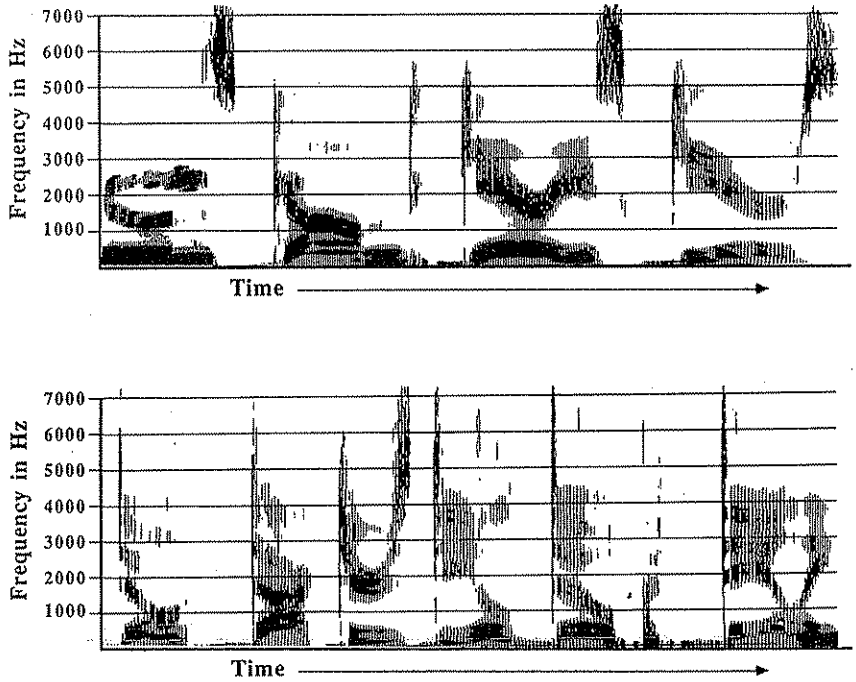


FIGURE 8-1 Spectrograms showing acoustic-phonetic invariance for the word-initial consonants /g/ and /d/. The upper spectrogram shows the sentence, "Goons gummed Gary's gears." The lower spectrogram shows the sentence, "Don't doctors deal dope daily?"

no other features or temporal distributions of features could have occurred.

Neither the linearity nor the invariance condition is met in natural speech, primarily because of the way speech is produced: the speech articulators move continuously in production so that the shape of the vocal tract for each intended phoneme is influenced by the shapes for the preceding and following phonemes. Coarticulation results in overlapping features, *smearing*, among neighboring phonemes. Hockett (1955) likens the relation between intended phonemes and the physical speech signal to a series of Easter eggs that are pushed through a wringer. The effect of the speaker's coarticulatory wringer is to create a speech signal in which there is rarely a stretch of sound that corresponds uniquely to a given phoneme. Instead, the cues overlap in time, resulting in what Liberman et al. (1967) termed the "encoded nature of speech." Coarticulation produces complex mappings between acoustic cues and perceived phonemes. Acoustic features for phonemes vary widely as a function of varying phonetic contexts, speaking rates, and other factors. Figure 8-1 shows the variations in the acoustic forms of second-formant transitions for the phonemes /d/ and /g/ as a function of varying vowel contexts (Liberman et al., 1954; Delattre, Liberman, and Cooper, 1955). Comparison of the various physical manifestations of the /d/s and /g/s demonstrates acoustic-phonetic variability. Although the second-formant transitions provide the cues for place of articulation necessary to recognize these phonemes, the acoustic realizations of transitions are clearly not invariant.

The failure of the speech signal to satisfy the linearity and invariance conditions is perhaps the most important puzzle in speech perception. It constitutes what Studdert-Kennedy (1983) calls the "animorphism paradox"—the invariant units of perception do not correspond to invariant acoustic segments in the signal. Indeed, the problems of **acoustic-phonetic invariance** have guided speech research since its beginning, and many researchers are working on these problems today. While some researchers have continued the quest for invariant aspects of the acoustic signal (Stevens and Blumstein, 1978, 1981; Kewley-Port, 1982, 1983; Kewley-Port and Luce, 1984; Sussman, 1989, 1991; Sussman, McCaffrey, and Matthews, 1991; Sussman, Hoemeke, and Ahmed, 1993; Fowler, 1994), and others have addressed the problems of invariance via theoretical innovation (Liberman and Mattingly,

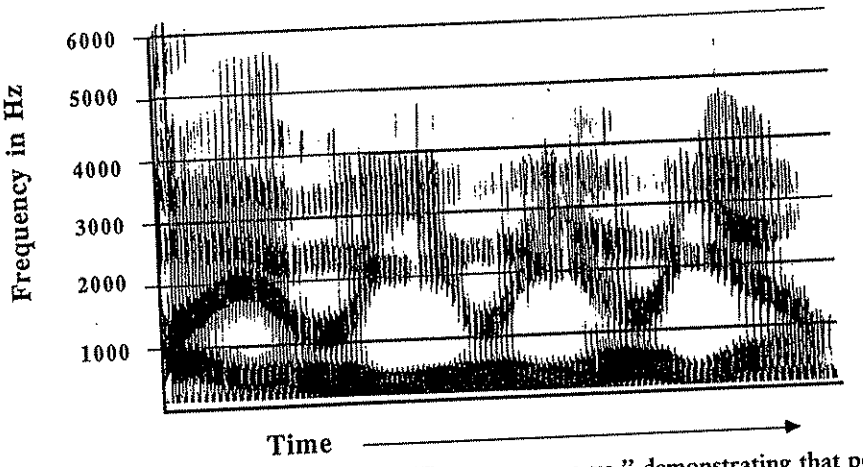


FIGURE 8-2 Spectrogram of the utterance, "I owe you a yo-yo," demonstrating that perceptual segmentation is not clearly reflected in acoustic segmentation.

1985; McClelland and Elman, 1986), the problem of contextual variability in speech remains central in research.

Coarticulation poses another problem for research, namely the lack of clear segmentation in the speech signal. Although listeners perceive speech as a series of discrete phonemes and words, physical temporal boundaries between phonemes are not reliably found in the spoken utterance. Although the sentence in Figure 8-2 displays almost no physical landmarks on which to base segmentation, it does not pose any special difficulty for listeners. The rate of information transmission in speech is enormous, clearly requiring that listeners somehow convert the continuous waveform into discrete, abstract units for cognitive processing (Liberman et al., 1967; Neisser, 1967). However, the speech signal does not lend itself to simple segmental analysis. Although it is possible to segment speech according to purely acoustic criteria (Fant, 1962), the segmentation provided by such algorithms typically does not correspond to the segmented representation that a listener would perceive. The importance and the difficulty of segmentation become immediately apparent when considering recent attempts to develop speech recognition devices; lack of segmentation, linearity, and invariance have been intractable problems.

Units of Analysis in Speech Perception

Nonlinearity, the lack of acoustic-phonetic invariance, and the nonsegmental nature of speech create another problem, the selection of a minimal unit of perceptual analysis. Given the information-rich speech waveform and limited channel capacities of the auditory system and auditory memory, it is clear that raw sensory information must be encoded into some scheme that can be efficiently processed (Broadbent, 1965; Liberman et al., 1967). Consider the estimate of Liberman, Mattingly, and Turvey (1972) that the conversion of speech sounds into phonemes reduces the information transfer rate of speech from approximately 40,000 bits per second to 40 bits per second. The conversion of phonemes into higher units of linguistic analysis further reduces the bit rate.

Figure 8-3 shows several possible units of analysis. The question for theories of speech perception has typically concerned selection of the "best" or most natural coding unit; claims of primacy have been made for phonetic features, phonemes, syllables, and words. Researchers in generative linguistic theory have even proposed units as large as the clause or sentence (Miller, 1962; Bever, Lackner, and Kirk, 1969).

Debates concerning the primacy of various units were prevalent in the literature for several years. A long-standing debate in the 1970s that has returned to prominence recently (see Pitt and Samuel, 1993) centered on claims that the syllable is a more basic perceptual unit than the phoneme (Savin and Bever, 1970; Massaro, 1972). Massaro (1972) argued that syllables are more discretely represented in the speech signal than phonemes, so selection of the syllable as the primary unit of perception resolves the

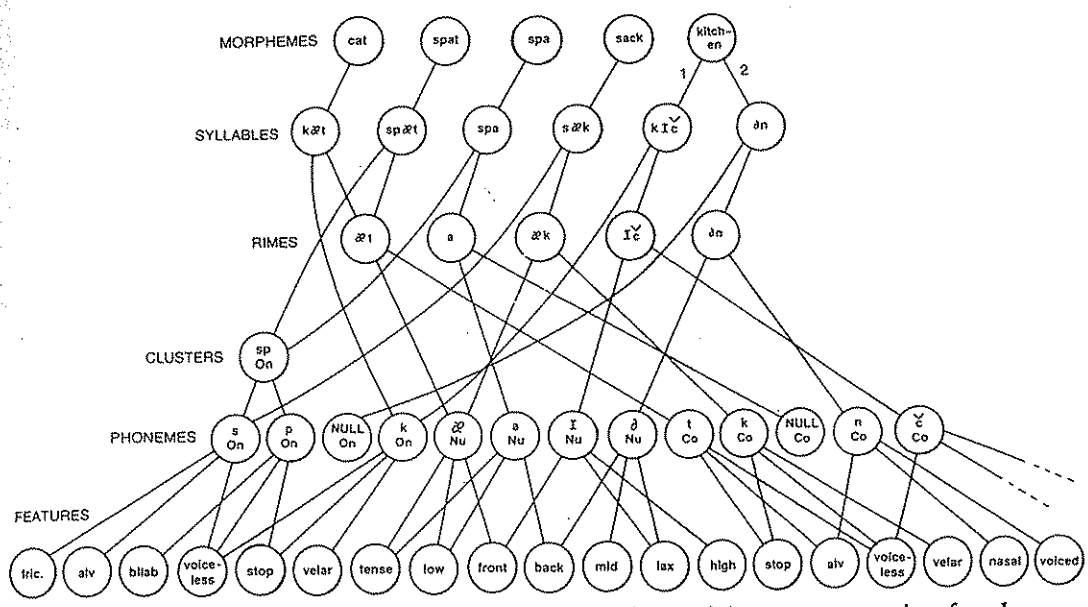


FIGURE 8-3 A section of a speech processing network containing numerous units of analysis, including morphemes, syllables, syllable rimes, consonant clusters, phonemes, and phonetic features. (From Dell, G. [1986]. A spreading activation theory of retrieval in sentence production. *Psychological Review, 93,* 283-321.)

problems of segmentation and invariance quite easily. Unfortunately, invariance is no more tractable with syllable-sized units than with phoneme-sized units. Furthermore, the information conveyed by syllables may depend on retrieval of their segmental constituents, so the issue of primary units is not resolved (Hawles and Jenkins, 1971; Pisoni, 1978).

Recent theories in speech perception imply that during comprehension of fluent speech, the primacy of any particular unit of speech may be less important than the obligatory units of speech and their interactions (McClelland and Elman, 1986). Although problems of coarticulation have discouraged researchers from positing the phoneme as an obligatory unit, numerous alternatives have been proposed. Examples include syllables (Cole and Scott, 1974a, 1974b; Studdert-Kennedy, 1974, 1980; Massaro and Oden, 1980; Segui, 1984), context-sensitive allophones (Wickelgren, 1969, 1976), and context-sensitive spectra (Klatt, 1979). All of these approaches have attempted to alleviate the problem of acoustic-phonetic invariance via the proposal of units that are relatively invariant in continuous speech. Although there is still ample reason to consider the importance of segmental representations in speech perception and word recognition (Pisoni and Luce, 1987), other context-sensitive perceptual units incorporate contextual variability directly in their representations and may therefore prove more robust to the problems of coarticulated speech.

These four fundamental problems have shaped speech perception research and theory for nearly four decades and will no doubt figure prominently in future work as well. Although other issues have characterized speech research in recent years, these issues capture the essence of the problem of speech perception: how does the listener convert the continuously varying speech waveform into a series of discrete representations for linguistic analysis? Any reasonable theory of speech perception must address this fundamental question. The next section addresses other issues that have been less focal but no less interesting. These include the specialization of speech, the problem of perceptual constancy, and the importance of suprasegmental and source information in speech.

FURTHER ISSUES IN SPEECH PERCEPTION

Specialization of Speech Perception

For many years, Liberman and his colleagues at Haskins Laboratories have proposed a view of speech perception as a specialized process requiring specialized neural mechanisms unique to humans (Studdert-Kennedy, 1980; Mattingly

and Liberman, 1988; Liberman and Mattingly, 1989). Early support for the claim that speech is special came from a well-known study by Liberman et al. (1957), who generated a synthetic continuum of consonant-vowel (CV) syllables ranging from /b/ to /d/ to /g/ by changing the second-formant transitions in graded steps. Although the physical changes between adjacent stimuli were small, subjects' identification responses were sharply discontinuous. Despite the graded steps in the continuum, subjects' perception of the syllables shifted abruptly, falling into natural categories for the phonemes /b/, /d/, and /g/. Moreover, when subjects were asked to discriminate among tokens from the stimulus continuum, their discrimination of tokens from different phonemic categories was nearly perfect, but their discrimination of tokens within the same phonemic category was nearly at chance. The phenomenon of discontinuous, categorical perception for speech sounds was markedly different from typical results of psychophysical experiments employing nonspeech stimuli such as pure tones. Nonspeech continua are perceived continuously, resulting in discrimination functions that are monotonic with respect to the physical scale. These differences between speech and nonspeech perception led researchers to propose that speech perception is subserved by specialized mechanisms distinct from mechanisms for general audition (see Repp [1983a] for a comprehensive review of the categorical perception literature).

A number of other phenomena have been purported to demonstrate the specialized nature of speech. These include findings of phonetic discrimination in infants, the rigidity of adult phonetic categories, cross-modal cue integration, cue trading relations, and duplex perception. These phenomena are considered in this chapter. However, for the development of speech perception, see Aslin and Pisoni, 1980; Eimas, et al., 1971; Walley, Pisoni, and Aslin, 1981; and Chapter 9 of this volume.

Perception of Speech and Nonspeech Signals

Some of the earliest empirical support for the claims of specialization for speech came from the categorical perception of speech stimuli compared with the continuous perception of nonspeech stimuli. The explanation for these differences offered by Liberman (1970a, 1970b), Liberman et al. (1967), and Studdert-Kennedy and Shankweiler (1970) was based on the **motor theory** of speech perception, in which speech perception is assumed to be mediated by knowledge of articulation. Considering the stimulus continuum examined by Liberman et al. (1957), although the physical scale was composed of many graded steps of second formant transitions, production of /b/, /d/, and /g/ corresponds to three discrete, discontinuous places of articulation. Listeners' perception of these sounds does not follow the continuous physical attributes of the signal but seems to follow the abstract, discontinuous places of articulation. The fact that nonspeech signals are continuously perceived was taken as further support that at least for stop consonants, speech perception entails a specialized **speech mode of perception.**

The motor theory account of categorical perception, and the generality of the data themselves, were challenged by researchers who believed that the same phenomena could be explained via general principles of auditory perception (Massaro, 1972, 1987; Cutting, 1978; Schouten, 1980; Pastore, 1981). A problem with using the basic psychophysical studies as the contrast to the speech perception studies was that neither the nonspeech stimuli nor nonspeech categorization tasks were adequately matched to their speech counterparts (Pisoni, 1991). More recently, a number of experiments using analogs of speech have demonstrated that subjects can perceive continuously varying stimuli categorically even though they reportedly hear the stimuli as nonspeech events such as tones or beeps. Such demonstrations of categorical perception for nonspeech signals imply that generic psychophysical principles may account for categorical perception; perception may be discontinuous, ostensibly without reference to articulatory knowledge. Accordingly, these researchers have attempted to account for categorical perception of speech stimuli via general auditory processing of acoustic stimuli, whether speech or nonspeech.

Lisker and Abramson (1964, 1967) demonstrated that categorical perception between voiced and voiceless stops (/b/ versus /p/, /d/ versus /t/, /g/ versus /k/) is determined by voice onset time. VOT is the silent interval between the burst release at the articulators and the onset of voicing. In voiceless stops, there is typically a long lag between the burst release and voicing; in voiced stops, the lag is shorter and may even

be negative (voicing begins before the stop is released). The finding that the temporal coordination of articulatory gestures determines categorical perception is consistent with an articulation-based mode of speech perception. However, similar findings have been obtained in experiments using nonspeech materials. Miller et al. (1976) created nonspeech VOT analogs by generating stimuli that contained aperiodic noise bursts followed by periodic buzzing. The interval between the noise burst and the buzz was varied in small steps, following Lisker and Abramson's earlier VOT experiments. Subjects asked to classify the stimuli according to "noise" vs. "no noise" showed categorical perception and discrimination very similar to those found with speech stimuli. Similarly, Pisoni (1977) employed stimuli that were even less speechlike than those used by Miller et al. (1976) and still observed categorical perception. Pisoni presented stimuli composed of only two tones, one at 500 Hz and one at 1500 Hz. The low tone either preceded or followed the high tone by as much as 50 ms, with graded steps in between. Categorical identification and discrimination closely resembled those obtained by Lisker and Abramson (1967). In addition, Jusczyk et al. (1980) found that infants perceive the two-tone stimuli categorically, just as they perceive speech stimuli (Eimas et al., 1971).

Comparison of the perception of speech and nonspeech signals reveals that other phenomena believed to demonstrate specialized speech processing can be explained by general auditory mechanisms. In a study of the effect of perceived speaking rate on phonetic classification, Miller and Liberman (1979) generated a series of synthetic speech stimuli ranging from /ba/ to /wa/ by gradually changing the duration of the formant transitions leading into the steady-state vowel formants. The most important manipulation was varying the duration of the syllables, hence the perceived speaking rate of the syllables. Miller and Liberman found that, as perceived speaking rate increased, subjects' category boundaries shifted toward /wa/, implying that at faster speaking rates, listeners accept shorter transitions as /w/. Miller and Liberman accounted for these data by proposing that specialized perceptual mechanisms compensate for changes in speaking rate in the perception of stops versus glides (see also Miller and Wayland, 1993). Eimas and Miller (1980) demonstrated the same compensatory phenomenon with infant subjects, implying that the specialized mechanism is innate.

However, it was later found that the compensation for speaking rate could be obtained using nonspeech analogs of speech. Pisoni, Carrell and Gans (1983) generated nonspeech analogs (three component tones) of the Miller and Liberman (1979) stimuli. Subjects categorized these stimuli as either "gradual onset" or "abrupt onset" and displayed a category boundary shift dependent upon duration that bore a striking resemblance to the speech data (Fig. 8-4). From these data Pisoni, Carrell, and Gans (1983, p. 320) suggested that postulation of specialized, rate-sensitive mechanisms for speech may be unwarranted. Instead, they argued that "context effects in discrimination may simply reflect the operation of fairly general auditory processing capacities." Indeed, Oller, Eilers, and Ozdamar (1990) proposed a simple psychophysical model based on linear regression to account for the rate compensation effect. Finally, Jusczyk et al. (1983) replicated the Pisoni, Carrell, and Gans (1983) experiments, showing that 2-month-old infants exhibit the boundary shift for nonspeech as well as speech.

These speech-nonspeech comparison studies suggest that the proposal of specialized mechanisms for speech perception may be unwarranted. However, despite the studies demonstrating the similarities of speech and nonspeech perception, important differences have also been observed (Pisoni, 1991). A number of studies have shown that when listeners are induced to process nonspeech auditory signals in a speech mode (for instance, when told they will hear poor-quality synthetic speech and should label the stimuli using phonetic categories), their perception changes markedly. Such demonstrations are typically provided by between-subjects experiments using a common pool of perceptually ambiguous stimuli that can be heard as either speech or nonspeech, according to the listener's expectations. In one condition subjects are told that they will hear synthetic speech, and in the other they are told that they will hear beeps or tones. After some performance measure is collected from subjects, they are typically queried to ensure that they actually thought the stimuli sounded like either speech or tones, depending upon their assigned group.

When subjects were presented with the **sine wave speech** sentence "Where were you a year ago?" Remez et al. (1981) found that simply

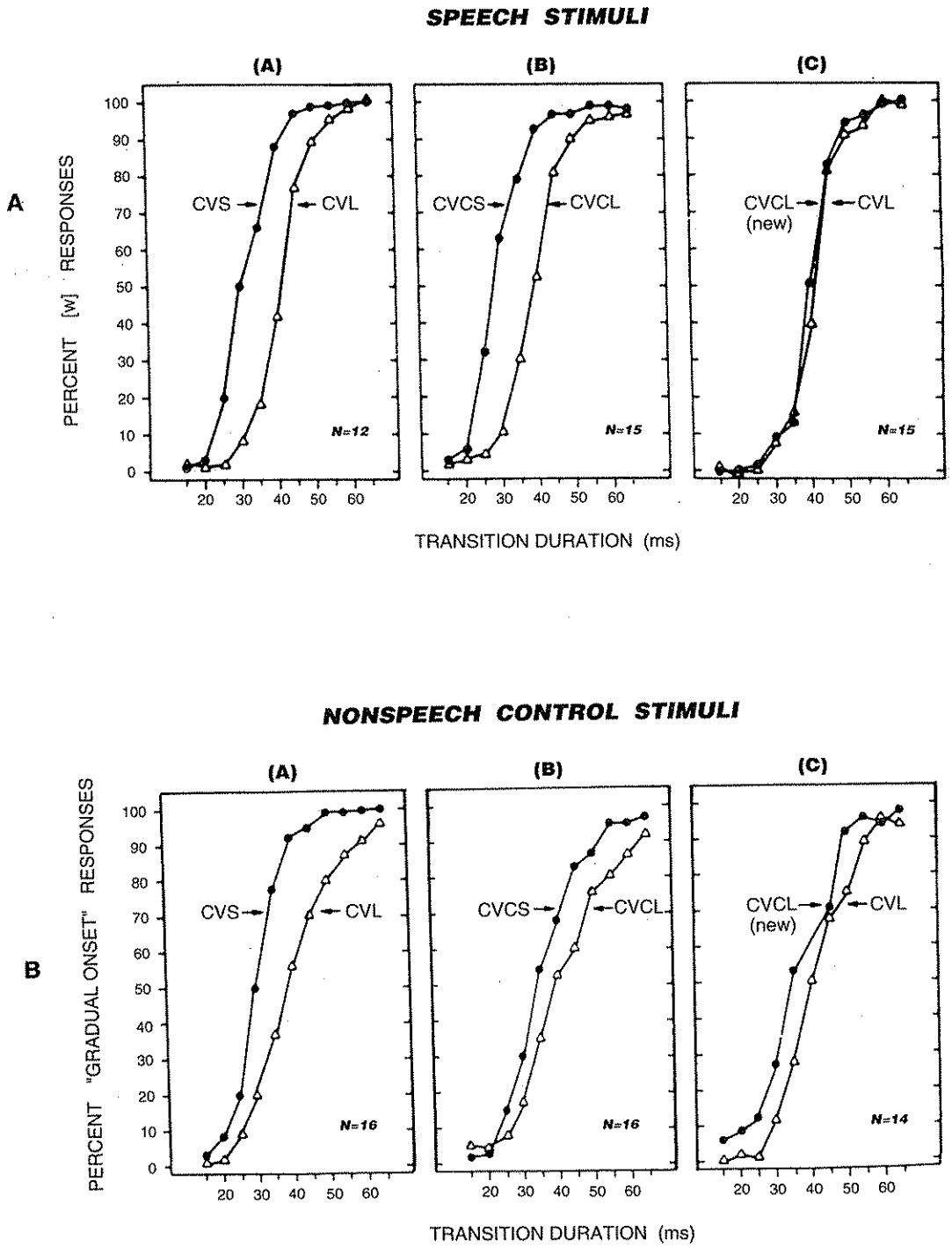


FIGURE 8-4 *A* shows categorization functions for synthetic speech stimuli (Miller and Liberman, 1979). *B* shows similar categorization functions for nonspeech stimuli that subjects heard as tones. (Adapted from Pisoni, D. B., Carrell, T. D., & Gans, S. J. [1983]. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics, 34,* 314-322.)

informing subjects that the signal was speech changed subjects' perception from a series of whistles and beeps to a correctly transcribed sentence (see also Bailey, Summerfield, and Dorman, 1977; Grunke and Pisoni, 1982; Tomiak, Mullennix, and Sawusch, 1987). Of course, the Remez et al. (1981) results may have been due to either a qualitative change in perception or any number of changes in response biases. In an experiment that left less room for a bias interpretation, Schwab (1981) found substantial backward masking and upward spread of masking for sine wave stimuli heard as tones. However, all masking was eliminated when subjects heard the stimuli as speech.

The major difference between perception in the speech and nonspeech modes for ambiguous stimuli appears to be the difference between holistic and componential analysis. It seems that listeners in the speech mode spontaneously forego detailed spectral analysis of the stimuli and make their speech categorizations based on entire, complex configurations of cues.* Conversely, subjects in a nonspeech mode behave more analytically, actually hearing the component parts of the stimuli individually. Further evidence of this was provided by Tomiak, Mullennix, and Sawusch (1987) in an experiment using the Garner (1974) speeded classification task. When told that they would classify nonspeech patterns, subjects separately processed the component dimensions of noise-tone analogs of fricative-vowel syllables. Irrelevant variation in the noise spectra did not affect reaction times for classifying tones. However, when subjects were told that the stimuli were synthetic fricative-vowel syllables, the components were processed integrally so that irrelevant variation in either dimension increased reaction times to classify stimuli along the other dimension. Finally, in an experiment reported by Grunke and Pisoni (1982), subjects were asked to identify ambiguous stimuli with either phonetic or acoustic labels, depending on their assignment to conditions. The stimuli were composed of either one, two, or three component tones. In the one- and two-tone conditions, subjects who used the acoustic categories "rising" and "falling" performed better than subjects who used phonetic categories. However, when a third tone was added, acoustic classifications were greatly impaired and phonetic classifications were substantially improved. Apparently the third tone made the signal more speechlike to listeners in the speech mode and noisier to listeners in the nonspeech mode.

Given these and similar findings, it is apparent that speech and nonspeech modes of perception differ in fundamental, qualitative respects (see Fowler [1990] for a different view). However, the basis of these differences remains to be explained. Are the differences due to the selective operation of different perceptual modules, response strategies, or attentional capacities? This question carries great theoretical importance and merits deeper investigation.

In sum, studies comparing speech and nonspeech perception have repeatedly called into question the strong claims regarding the specialized nature of speech perception. However, the evidence and arguments on both sides are equivocal, and the implications of these studies are subject to interpretation. To appreciate the degree to which the meaning of these studies is in the eye of the beholder, compare the conclusions from two reviews of the speech-nonspeech literature on categorical perception:

> The nonspeech studies to this point do more than just refute the view that categorical perception is specific to speech. They demonstrate that there are certain important similarities in the ways certain classes of speech and nonspeech sounds are perceived. (Jusczyk, 1986, p. 43).

> In summary, despite a few suggestive results, there is no conclusive evidence so far for any significant parallelism in the perception of speech and nonspeech. (Repp, 1983a, p. 50).

Duplex Perception

Duplex perception is a phenomenon discovered by Rand (1974) and recently cited as strong evidence for a dissociation of phonetic perception from general auditory perception (Liberman, 1982; Repp, 1982; Studdert-Kennedy, 1982; Liberman and Mattingly, 1985, 1989). The general procedure for eliciting the duplex percept is simple. A listener is presented with two simultaneous, dichotic stimuli. One ear hears an isolated third-formant transition that sounds like a nonspeech chirp. At the same time the other ear receives a base syllable. This base syllable consists of the first two formants, complete with transitions, and the third formant without a transition. Typically, the transition presented in isolation completes the syllable to

*Another interpretation may be that subjects in a speech mode process components of the signal separately but in accordance with well-learned combinations.

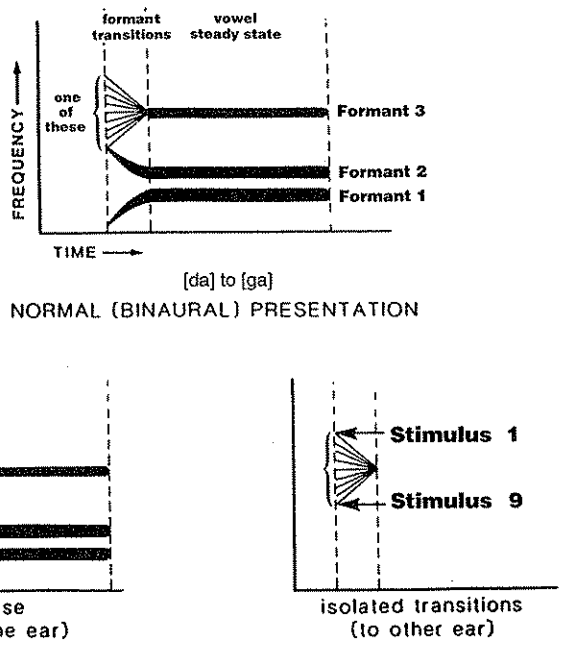


FIGURE 8-5 **Stimuli used in a duplex perception experiment. The upper panel shows a synthetic speech syllable with a range of third formant transitions. When presented binaurally, these syllables range from /da/ to /ga/. The lower left panel shows the constant syllable base that is always presented in the dichotic listening task. The lower right panel shows a series of third-formant transitions, ranging from /ga/ to da/, which are combined with the syllable base.** (From Mann, V. A., & Liberman, A. M. [1983]. Some differences between phonetic and auditory modes of perception. *Cognition, 14,* 211-235.)

create a /da/ or /ga/, sometimes in graded steps along a continuum (Fig. 8-5). When the base and the transition are presented dichotically, the listener's percept is duplex; that is, the completed syllable is perceived and the nonspeech chirp is heard at the same time. Liberman and Mattingly (1989) argue that the phonetic module and a separate general auditory module each respond to different aspects of the stimuli, thus creating the duplex percept.

Several further findings support the claim that segregated modes of processing are responsible for the separate percepts. For example, in one study, Mann et al. (1981; reported in Liberman, 1981) presented a series of different third-formant transitions such that upon fusion the entire syllables consisted of a continuum from /da/ to /ga/. When subjects attended to the nonspeech side of the percept, continuous discrimination functions typical of nonspeech were obtained. When subjects attended to the speech percept, categorical discrimination functions were obtained, implying that the separate percepts are subserved by separate processing systems. Other experiments have demonstrated that various stimulus or procedural manipulations can affect the speech or nonspeech percept independently, again implicating separate processors (Isenberg and Liberman, 1978; Bentin and Mann, 1990; Nygaard and Eimas, 1990; Nygaard, 1993).

Taken together, these findings on duplex perception support the claim of an independent and specialized phonetic recognition system. As Repp (1982, p. 102) concludes:

> Duplex perception phenomena provide evidence for the distinction between auditory and phonetic modes of perception. They show that, in the duplex situation, the auditory mode can gain access to the input from the individual ears, whereas the phonetic mode operates on the combined input from both ears. The "phonological fusion" discovered by Day (1968)—two dichotic utterances such as "banket" and "lanket" yield the percept "blanket"—is yet another example of the abstract, nonauditory level of integration that characterizes the phonetic mode.

Similarly, Whalen and Liberman (1987, p. 171) describe the phenomenon of duplex perception as evidence for the preemptiveness

of speech, arguing that the speech module gets the first crack at interpreting an auditory signal: "The phonetic mode takes precedence in processing the transitions, using them for its special linguistic purposes until, having appropriated its share, it passes the remainder to be perceived by the nonspeech system as auditory whistles."

These interpretations of duplex perception are not universal; several lines of counterevidence have been offered. Pastore et al. (1983) observed duplex perception for musical chords (two notes in one ear, a third note in the other), casting doubt on the claim that duplex perception is a solely speech-based phenomenon. In addition, Nusbaum, Schwab, and Sawusch (1983) demonstrated that listeners use the information in the third-formant transition independently of the base, casting doubt on the claim that the transition in isolation is a true nonspeech signal (see, however, Nusbaum, 1984, and Repp, 1984). This finding by Nusbaum, Schwab, and Sawusch (1983) implies that subjects in the studies reported earlier could have generated their phonetic decisions without any process of auditory fusion between the ears.

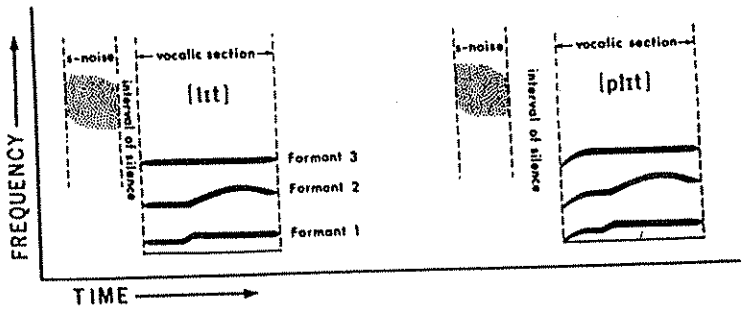
An especially strong challenge to the speech module interpretation of duplex perception comes from Fowler and Rosenblum (1990, 1991). Borrowing language and ideas from Gibson's (1966) event perception, Fowler and Rosenblum argue that duplex perception may demonstrate not the preemptiveness of speech per se but simply the preemptiveness of any meaningful event. The argument is that human sensory systems have evolved to recognize important objects and events around us ("affordances," in Gibson's [1979] terminology). Therefore, our perceptual and cognitive systems are naturally attuned to perceive meaning from any stimulation. Accordingly, Fowler and Rosenblum predicted that duplex perception would occur whenever two acoustic fragments, when integrated, specify a natural event and when one of the fragments has any unnatural quality. Fowler and Rosenblum (1991, pp. 51, 52) write, "Under these conditions, the integrated event should be preemptive and the intense fragment should be duplexed regardless of the type of natural sound-producing event that is involved, whether it is speech or nonspeech, and whether it is profoundly biologically significant or biologically trivial."

To demonstrate duplex perception for a biologically trivial event, Fowler and Rosenblum dichotically presented a low-pass filtered recording of a slamming metal door to one ear and the remaining high-frequency noise to the other ear. Alone, the base sounded like a wooden door slamming, and the chirp sounded to the authors like a can of rice being shaken (recall that we are biased to perceive sounds as events). When the stimuli were played together with the chirp at a higher amplitude than the base, most subjects reported the duplex perception of metal door + chirp. This demonstration of duplex perception for such a completely nonspeech signal calls into question both the relevance of duplex perception to speech research and the specialized nature of speech perception (see Hall and Pastore [1992] for a similar demonstration of duplex perception using musical chords). Findings such as Fowler and Rosenblum's clearly underscore the need for deeper investigation into duplex perception before it is too richly interpreted.

Trading Relations and Integration of Cues

A third class of findings cited as evidence for the specialization of speech perception comes from studies of cue trading and cue integration (see Repp, 1982). The speech signal is replete with cues to phonetic contrasts, and several different cues may indicate a single contrast (Delattre et al., 1952; Denes, 1955; Harris, et al., 1958; Hoffman, 1958; Repp, 1982). This makes it possible that when the utility of one cue is reduced, another cue becomes primary. It is assumed that trading relations occur because the cues are phonetically equivalent with respect to the contrast in question. The cues may trade in importance when necessary or may integrate to provide robust contrasts when all cues are provided equally. Examples of cue trading have been provided by Denes (1955) and Fitch et al. (1980), who demonstrated the perceptual equivalence of closure durations and first-formant transitions in signaling the contrast between minimal pairs such as *slit–split* (Repp, 1982). In Figure 8-6 the phonetic trading relation of closure duration and formant transition is clearly evident.

Phonetic trading relations have been cited as evidence of a speech mode of perception for two main reasons: First, trading relations can occur between both spectral and temporal cues distributed over relatively long intervals. Repp (1982) argues that it is difficult to imagine that such cues



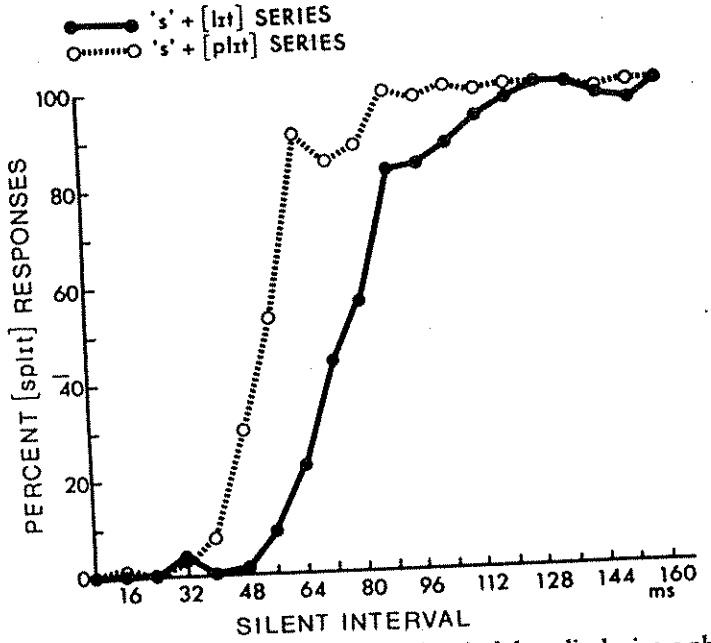


FIGURE 8-6 Examples of stimulus materials and typical data displaying a phonetic trading relation. The upper panel shows synthetic stimuli for the syllables "slit" and "split"; the duration of the silent closure intervals varies along a continuum during a perceptual experiment. The lower panel shows identification data for both syllables as a function of closure duration. When proper transitions are present, a short closure duration is sufficient to perceive "split"; when transitions are absent, a longer duration is sufficient to perceive "split". (From Fitch, H. L., Hawles, T., Erickson, D. M., & Liberman, A. M. [1980]. Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception and Psychophysics, 27,* 343-350.

would be integrated into a single percept unless some speech-specific system were mediating perception. Repp argues further that listeners must possess abstract articulatory knowledge to integrate such disparate cues (see also Liberman and Mattingly, 1985). Repp (1982, p. 95) suggests that:

> Trading relations may occur because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic formation by perceiving the most plausible articulatory act. This explanation requires that the listener have at least a general model of human vocal tracts and of their ways of action.

The second reason that trading relations have been cited as evidence for the speech mode of perception comes again from comparisons of speech and nonspeech perception. Best, Mor-

rongiello, and Robson (1981) reported two experiments using sine wave speech, which may be heard as either speech or nonspeech, depending primarily upon the listener's expectation. They found that listeners in a speech mode exhibit cue trading and integration, but listeners in a nonspeech mode do not. They considered these findings proof that the integration and perceptual equivalence of multiple cues are specific to speech.

Their conclusion has been challenged. For example, Massaro and Oden (1980) have presented a model of speech perception that accounts for trading relations while making no assumptions of specialized processing (see also Massaro, 1972, 1987, 1989; Massaro and Cohen, 1976, 1977; Oden and Massaro, 1978; and Derr and Massaro, 1980). Massaro and Oden argue that multiple features corresponding to a single phonetic contrast are extracted independently from the speech waveform and are integrated multiplicatively into a unitary percept. The weight given to each feature in this integration is determined by the strength, or certainty, of the feature's presence. By this account, speech perception reduces to a "prototypical instance of pattern recognition" (Massaro and Oden, 1980, p. 131).

Repp (1983b, p. 132) arrived at a conclusion similar to that of Massaro and Oden, stating that trading relations ". . . are not special because, once the prototypical patterns are known in any perceptual domain, trading relations follow as the inevitable product of a general pattern matching operation. Thus, speech perception is the application of general perceptual principles to very special patterns."

In short, as in the earlier debates regarding speech versus nonspeech perception and duplex perception, the evidence provided by trading relations is ambiguous with respect to claims of a specialized speech mode of processing.

Cross-Modal Cue Integration (The McGurk Effect)

Another recent finding in speech perception attributed to specialized mechanisms is *cross-modal cue integration,* or the McGurk effect (MacDonald, 1976; MacDonald and McGurk, 1978; Summerfield, 1979, 1983; Roberts and Summerfield, 1981). The phenomenon is a perceptual illusion, demonstrated as follows: A subject is presented with a video display of a talker (or synthesized face; see Massaro and Cohen, 1990) articulating simple CV syllables and hears spoken syllables synchronized with the visual display. The McGurk effect occurs when the visual and auditory syllables are incongruous. The listener typically reports hearing neither the spoken syllable nor the lip-read syllable, but something in between. For example, when presented with a face that articulates /ma/ and an auditory syllable /ta/, most subjects report hearing /na/ (Fig. 8-7).

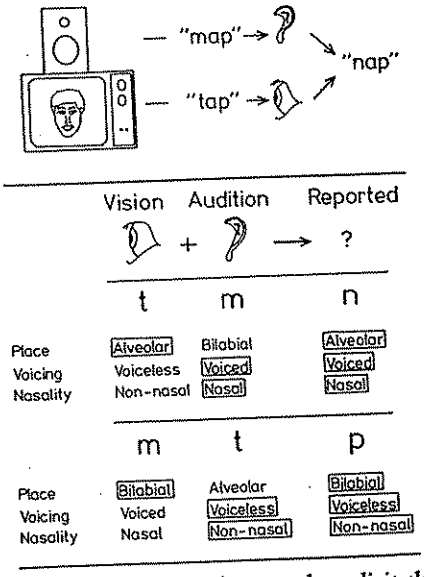


FIGURE 8-7 The procedure used to elicit the McGurk illusion. The video displays a face articulating the syllable /tap/ and the audio channel outputs the syllable /map/. Audiovisual integration leads the subject to perceive the syllable /nap/. (From Summerfield, Q. [1983]. Audio-visual speech perception, lipreading and artificial stimulation. In Lutman, M. E., & Haggam, M. P. (Eds.), *Hearing Science and Hearing Disorders.* London: Academic Press.)

According to subjective reports of the McGurk illusion, the effect is quite striking. Liberman (1982) points out that the procedure affects listeners' experience of hearing the syllable as an integrated event, to an extent that listeners cannot determine the degree to which their perception of syllable identity is due to either source of information. For example, Repp (1982, p. 102) reports,

> I have experienced this effect myself (together with a number of my colleagues at Haskins) and can confirm that it is a true perceptual phenomenon and not some kind of inference or bias in the face of conflicting information. The observer really believes that he or she hears what, in fact, he or she only sees on the screen; there is little awareness of anything odd happening.

The McGurk illusion has been interpreted as particularly strong evidence for a specialized speech perceptual system that makes reference to articulatory gestures. Fowler and Rosenblum (1991, p. 37) speculate, "Why does integration occur? One answer is that both sources of information, the optical and the acoustic, provide information about the same event of talking, and they do so by providing information about the talker's phonetic gestures." However, there are detractors to this position. Massaro and Cohen (1983, 1990, 1993) have shown that their fuzzy logical model of perception provides precise accounts of the McGurk and MacDonald data without postulation of any speech-specific mechanisms. In addition, the generality of the phenomenon is limited. Easton and Basala (1982) found that the illusion is not invoked if whole words are used instead of syllables (although Dekle, Fowler, and Funnell, 1992, reported otherwise). This finding and Massaro's model suggest that the illusion is the product of general perceptual biases that are revealed only by highly ambiguous stimuli. Finally, one of the principal assumptions of cognitive psychology is that humans routinely perform intricate information processing that may involve any number of stages, computations, heuristics, or biases without any awareness of the operations they perform. Accordingly, despite the impressions of listeners regarding the illusion, the fact that a phenomenon seems truly perceptual does not allow us to conclude by fiat that the results cannot be due to biases (Neisser, 1967; Cutting, 1987).

Two further findings related to cross-modal integration do seem to tip the scales back in favor of a specialized-processing account. Miller (1990) cites 4- and 5-month-old infants' sensitivity to auditory-articulatory correspondence as strong evidence for innately specified perceptual mechanisms (Kuhl and Meltzoff, 1982; MacKain et al., 1983). Kuhl and Meltzoff (1982) found that infants prefer to watch a display of an articulating face if the accompanying spoken syllables match the articulation rather than incongruent audiovisual displays. Furthermore, Roberts and Summerfield (1981) used the McGurk phenomenon in a clever test of selective adaptation (see Eimas and Corbit, 1973). They presented subjects with an auditory syllable /bɛ/ and visual syllable /gɛ/, producing the percept of /dɛ/. However, on a test of adaptation, the perceived audiovisual syllable had the same effects as a purely auditory /bɛ/ on a /bɛ/–/dɛ/ series; subjects' phonetic perception of the stimulus as /dɛ/ was not reflected in their adaptation data. Studdert-Kennedy (1982, p. 7) considers this finding a powerful indication of the dissociation of general auditory and phonetic perception:

> I take [the procedure of] audio-visual adaptation to demonstrate unequivocally the on-line dissociation of auditory and phonetic perception. Moreover, following Summerfield (1979), I take the results of the audio-visual adaptation study to demonstrate that the support for phonetic perception is information about the common source of acoustic and optical information, namely, articulatory dynamics.

Studdert-Kennedy may be correct. Alternatively, we may assume, as in an information-processing model of speech perception (Cutting and Pisoni, 1978), that the pathway from audition to phonetic perception is composed of processing stages (see also Studdert-Kennedy, 1974, 1976). The locus of the phonetic perception of the McGurk paradigm and the locus of the adaptation effect could be separated so that the adaptation manipulation affects some stage of processing that precedes the audiovisual integration. This seems likely. Presumably the integration of information from vision and audition occurs somewhat late in the speech perception process. If so, the Roberts and Summerfield (1981) data may not imply a strict auditory versus phonetic dissociation; the adaptation stimulus may simply affect precategorical phonetic perception, an explanation that would be compatible with either a motor theory or an auditory theory. Perhaps the only firm conclusion is that the McGurk effect, like duplex perception, may eventually constitute compelling evidence for specialized speech perception based on articulatory gestures. For the present, however, more complete investigation of these phenomena is clearly necessary.

Role of Linguistic Experience in Speech Perception

An important but neglected issue relevant to the question of specialization concerns the role of linguistic experience on adult speech perception (see Studdert-Kennedy et al., 1970). It has long been known that infants can categorically discriminate among the phonemes of their native language and among many other nonnative phonemes. With continued linguistic experi-

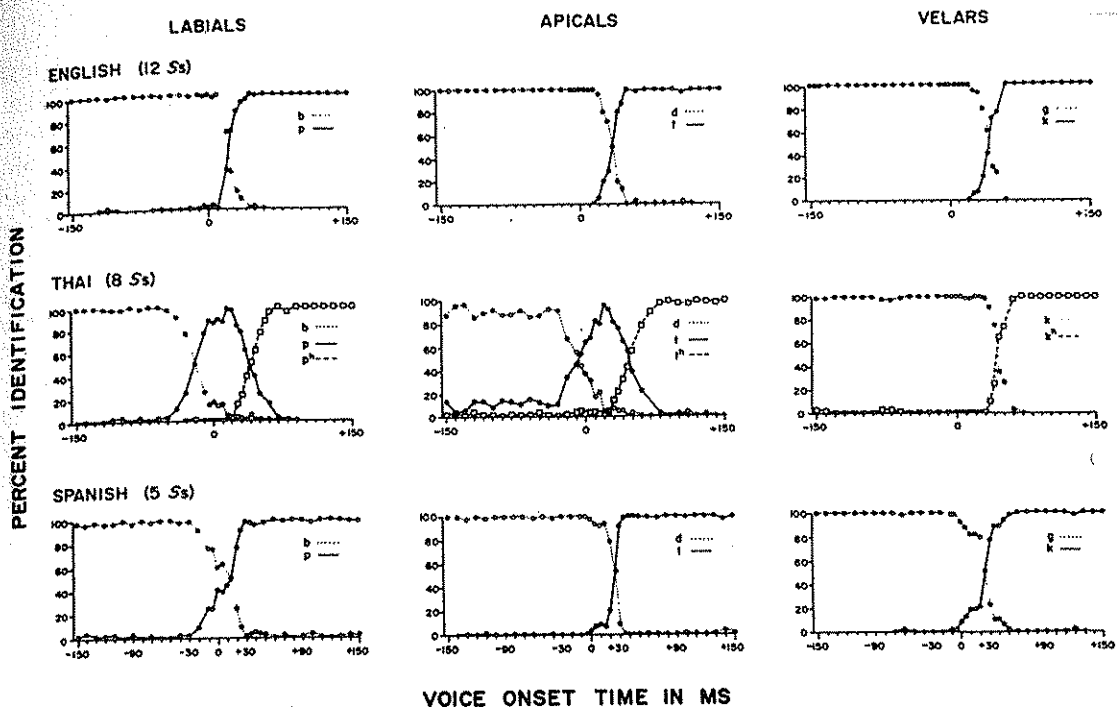


FIGURE 8-8 Cross-language identification data for labial, apical, and velar stops ranging in voice onset time from −150 to +150 msec. Categorization is clearly affected by the linguistic background of the listener. (Adapted from Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. [1982]. Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 297-314.)

ence, the listener's ability to discriminate between speech sounds not phonemically contrastive in the native tongue seems to be virtually lost (Strange and Jenkins, 1978; Aslin and Pisoni, 1980; Aslin, 1985; Logan, Lively, and Pisoni, 1991; Pisoni, Logan, and Lively, 1994). Maturation appears to pare down the set of all possible contrasts (or at least most; see Best, MacRoberts, and Sithole, 1988) that listeners can originally discriminate to only the set required for the native language (see Polka, 1992; Best, 1994; Polka and Werker, 1994). Evidence for the language-specific discrimination abilities of adults was first provided by Lisker and Abramson (1964, 1967; see also Abramson and Lisker, 1967). They investigated the abilities of speakers of varying languages to perceive three sets of synthetic speech stimuli that formed continua along the dimension of VOT that corresponded to labial, velar, and palatal sounds. The results of their experiments demonstrated that, in general, subjects from different linguistic backgrounds identified and discriminated the stimuli according to the contrastive phonological categories of their languages. The cross-language identification functions obtained by Lisker and Abramson (1967), shown in Figure 8-8, demonstrate the influence of the native language on perceptual classification.

Beyond the influence of the native phonemic repertoire on the typical identification of speech sounds, many studies have demonstrated the inflexibility of the adult listener's phonemic categories. Training an adult speaker of one language to discriminate reliably between phonemes of another language is very difficult and requires extensive training to obtain even small improvements (Strange, 1972; Vinegrad, 1972; Strange and Jenkins, 1978; Strange and Dittmann, 1984). Therefore, it was argued that the development of phonetic categories may require a plastic neural substrate that becomes less flexible after a critical period (Eimas, 1975). This view of the nature and development of phonetic categories is clearly compatible with the assumption, recently defended by Liberman and Mattingly (1989), that speech perception is modular. Fodor (1983) describes perceptual modules as innately specified, neurally hardwired and nonmodifiable. Fodor's modularity is therefore

compatible with the view that speech perception is subserved by perceptual and memory systems that are flexible only in infancy, becoming autonomous and impenetrable as early as possible.

More recent research, however, has demonstrated that significant improvements in discrimination of nonnative phonetic contrasts can be obtained using laboratory training procedures. In one experiment, Pisoni et al. (1982) trained English-speaking subjects to perceive three categories along a VOT continuum where only two categories naturally exist. A more recent example comes from training procedures employed by Logan, Lively, and Pisoni (1991; see also Pisoni, Logan, and Lively, 1994) to teach Japanese listeners to discriminate /r/ from /l/. Previous research showed that training Japanese listeners to distinguish these phonemes is extremely difficult and usually produces only marginal results (Goto, 1971; MacKain, Best, and Strange, 1981; Mochizuki, 1981; Strange and Dittmann, 1984). However, Logan, Lively, and Pisoni argued that neither the stimulus materials nor the training procedures employed in most of these studies were ideal for teaching listeners the intended contrast. For example, Strange and Dittmann (1984) used laboratory training procedures. Their methods failed and they concluded that training procedures are ineffective in modifying the phonetic categories of adult listeners. However, several aspects of their methodology call this global conclusion into question. First of all, the stimuli were synthetic tokens of "rock" and "lock", and no other stimuli were used. In addition, the procedure was a standard same-different discrimination task with limited feedback.

Logan, Lively, and Pisoni used natural tokens of minimal pairs contrasting /r/ and /l/. Furthermore, to provide listeners with more robust categories, they presented tokens produced by a variety of talkers. In addition, the target phonemes occurred in a variety of phonetic environments. Clearly, these natural and variable tokens contain far more cues and more ecological validity than the synthetic tokens employed in earlier studies. Using these varied materials and extensive feedback, Logan, Lively, and Pisoni observed substantial improvements in discrimination for all of their subjects. Similar preliminary results have been reported by Pruitt et al. (1990) in training English listeners to distinguish Hindi retroflex-dental consonants.

Findings such as these, as well as a large body of developmental data (Aslin and Pisoni, 1980; and Chapter 9 of this volume) have prompted several researchers (e.g., Jusczyk, 1985, 1986) to propose that phonological categories develop and are maintained by general attention and categorization mechanisms. These theories assume that the phonological inventory for any given language can be derived by selectively attending to relevant contrastive dimensions while selectively ignoring variation along irrelevant dimensions. Nosofsky (1986, 1987) has shown that this kind of selective attention strategy applied in simple category learning tasks can account for a wide variety of findings in the literature on categorization, perceptual identification, and the nature of psychological similarity. Logan, Lively, and Pisoni (1991) also refer to these attentional mechanisms to explain their learning data and to account for the learning failures of earlier studies. These proposals imply that the processes of speech perception rely on general cognitive principles of pattern recognition, attention, and categorization rather than highly specialized mechanisms unique to speech perception. However, we cannot determine whether these training procedures affect early phonetic perceptual processes or some later decisional processes. Clearly, we are still a long way from complete understanding of these issues, especially the developmental aspects of phonetic perception. For the present, however, we can maintain that adult phonetic categories are not rigid, as has been suggested, and that their flexibility is consistent with a view of speech perception based on general cognitive mechanisms.

Studies of Speech Perception in Nonhumans

One final area of research that merits consideration in this discussion is speech perception by nonhuman animals. The logic that motivates such research is simple: When strong claims were made that categorical perception was a speech-specific phenomenon, researchers attempted to demonstrate categorical perception of nonspeech signals. Similarly, when claims were made that categorical perception was uniquely human and speech-specific, researchers attempted to demonstrate that nonhuman animals with auditory systems roughly analogous to the human auditory system could also perceive speech sounds categorically. Clearly, animals do

not derive phonetic content from human speech, so any discrimination or categorization data provided by the animals must reflect general auditory and classification processes.

In studies of speech discrimination by monkeys, Morse and Snowdon (1975) and Waters and Wilson (1976) found preliminary evidence that monkeys perceive place of articulation categorically. More convincing evidence was provided in experiments conducted by Kuhl and Miller (1975, 1978) on perception of speech by chinchillas. In experiments on categorization (as indicated in an avoidance-conditioning task), it was demonstrated that, for stimulus continua varying VOT, chinchillas' categorization boundaries were remarkably similar to those for English-speaking listeners (Fig. 8-9). Moreover, Kluender, Diehl, and Killeen (1987) demonstrated that Japanese quail can learn apparently robust phonetic categories for stop consonants /b/, /d/, and /g/. The quail learned the stops in CV syllables followed by four different vowels and later could discriminate the three stops in the context of eight novel vowels; this generalization implies some form of abstraction of the category.

What are we to make of this? The results are certainly suggestive: If nothing else, they imply that given an auditory system similar to the human and a rudimentary ability to distinguish between stimuli, animals tend to respond differentially to speech signals that correspond to human phonetic categories. This implies that we need not hypothesize specialized, articulatory-based perceptual mechanisms to account for human speech perception. Unfortunately, the results of the animal studies can be taken only as suggestive. There is no reason to assume, for instance, that human languages would have evolved phonetic contrasts that were especially difficult for our auditory systems to discriminate (Stevens, 1972). The animal data may simply illustrate that phonetic categories are evolutionarily well conceived. Furthermore, since we have no access to the animals' experience, we have no basis for assuming that anything speechlike is perceived at all. In short, examining their behavior is rather like examining the behavior of a Turing machine—it may resemble human performance, but that does not mean it derives from the same underlying mechanisms (see also Repp, 1983a).

Finally, what are we to conclude about the entire debate concerning the specialization of speech perception? There are viable arguments on both sides of the issue. This debate has been fruitful for the sake of continuing research—more data have been generated and energy devoted to its resolution than to any other issue in speech perception. At the same time, the specialization hypothesis may be empirically unassailable. This may be true especially now that the specialization mechanism has been described in terms of the modularity hypothesis. Fodor (1985) describes three necessary characteristics of any experiment considered a bona fide counterexample to the modularity of a perceptual system: (1) The experiment must demonstrate the influence of background information (higher cognitive processes) on perceptual output. (2) This influence must clearly involve the perceptual system; it cannot reflect postperceptual processing or a decisional criterion shift. (3) The cognitively penetrated system must be the usual system for natural perception in the given domain, not involving some backup systems that are required only in special circumstances, such as in perceiving degraded stimuli. Consider, for example, the finding that mere instructions change the percept of sine wave speech from a sequence of tones into a sentence (Remez et al., 1981). At first glance, this appears to violate the impenetrable nature of the phonetic module, whose operations are supposed to be impervious to the listener's beliefs and expectations. (Fodor's preferred examples are optical illusions, such as the Mueller-Lyer illusion, which persists even when the observer knows that the lines are of equal length.) Clearly the sine wave speech demonstration satisfies the first condition, but the second and third are questionable. Furthermore, almost any experiment aimed at demonstrating the nonspecialized nature of phonetic perception may fail to satisfy at least one of these criteria. The challenge for future research is to address the relevant issues while circumventing these pitfalls.

Normalization Problems in Speech Perception

The problems posed for theories of speech perception by the inherent nonlinearity, variability, and nonsegmental nature of the speech signal arise from the basic assumption that listeners must somehow map distorted information in the speech signal onto canonical linguistic representations in memory. Typically, researchers in

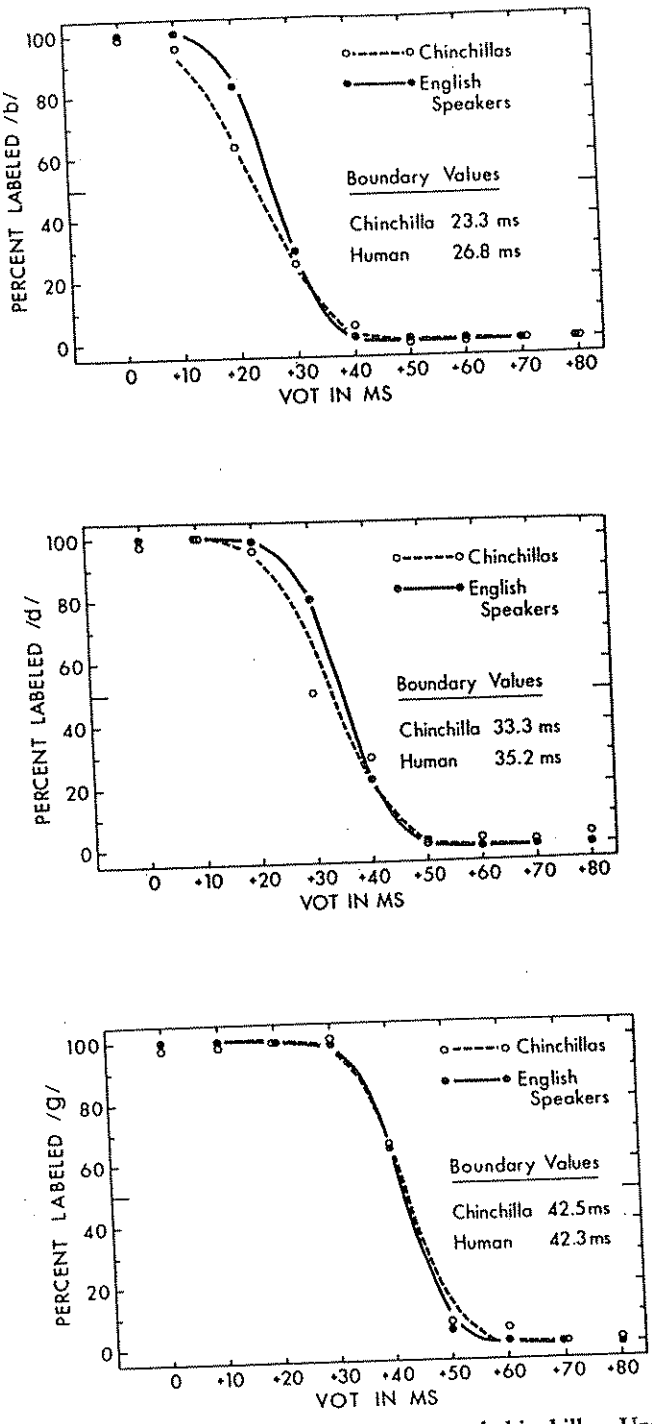


FIGURE 8-9 **Phonetic categorization data for humans and chinchillas. Upper panel, /b/-/p/ continuum; middle panel, /d/-/t/ continuum; lower panel, /g/-/k/ continuum. In all three cases, the phonetic boundaries of humans and chinchillas are quite similar.** (Adapted from Kuhl, P. K., & Miller, J. D. [1978]. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63,* 905-917.)

speech perception have limited their study of variability to the effects of different phonetic contexts. However, many factors beyond phonetic context influence the acoustic realizations of phonetic contrasts. Collectively, perceptual accommodations to variations in speech patterns to recover canonical linguistic units fall into the category of *perceptual normalization.* Recent research on normalization focuses on sources of variation such as talkers' vocal tract differences and speaking rate differences (although the problem of perceptual constancy is also introduced by a speaker with a mouthful of food, a singing voice, etc.).

Individuals differ in the sizes and shapes of their vocal tracts (Joos, 1948; Peterson and Barney, 1952; Fant, 1973), glottal characteristics (Carr and Trill, 1964; Monsen and Engebretson, 1977; Carrell, 1984), their idiosyncratic articulatory strategies for producing speech (Ladefoged, 1980; Johnson, Ladefoged, and Landau, 1993), and the dialects of their native regions (Abercrombie, 1967). This produces wide variability in production of the same words and phrases across individuals. Nevertheless, human listeners accurately perceive speech across virtually all (reasonably intelligible) speakers without any apparent difficulty. At present, little is known about the perceptual processes responsible for the implied perceptual compensations, nor is it known whether perceptual compensation occurs at all.

A related matter is time and rate normalization. Speech is a temporally distributed signal. Accordingly, the cues to individual phonetic contrasts in speech are distributed in time and are substantially influenced by changes in speaking rate. Moreover, the acoustic durations of phonetic segments are influenced by the locations of syntactic boundaries in fluent speech, by syllabic stress, and by the component features of adjacent segments (Gaitenby, 1965; Lehiste, 1970; Klatt, 1975, 1976, 1979). Segmental durations are modified further by contextual factors in speech. For example, vowels of words spoken in sentences are approximately half the duration of vowels of the same words spoken in isolation (Luce and Pisoni, 1987). In sum, phonetic contrasts in conversational fluent speech are characterized by widespread durational variation. Furthermore, it is well known that some durational variation in speech carries important information about numerous phonetic contrasts, word boundaries, and so on. In English, numerous phonetic contrasts are distinguished by durational cues. Thus, the listener must attend to and use durational cues to stress, phonemic contrasts, and pragmatics while ignoring irrelevant durational variations due to particular talkers or circumstances (Port, 1977; Miller, 1980).

Indexical Information in Speech

The human voice conveys information about a speaker's age and gender, as well as more cultural information such as regional origin, temperament, and social group membership. Such aspects of speech, known as **indexical information** (Abercrombie, 1967), do not, in general, relate directly to processes of phonetic perception (other than adding still more variability) but are heavily used in linguistic communication nonetheless. For example, most of us are reasonably expert at discriminating a New England accent from a Japanese accent, just as we are reasonably expert at discriminating the speech patterns of children from those of adults. Indexical information also alerts the listener to the speaker's identity and to important changes in the physical or emotional state of the speaker. Ladefoged and Broadbent (1957) call these aspects of the voice "personal information."

The use of indexical information in everyday communication is pervasive, as when we infer quite extensive information about a speaker's origin and background; in many societies speech patterns are commonly associated with social status (Abercrombie, 1967). Aside from cultural speech patterns, speech patterns of an individual speaker are richly informative. We are remarkably sensitive to the emotional or physical state of a speaker within our own culture, and we can readily identify people we know from their voice alone (Van Lancker, Kreiman, and Emmorey, 1985; Van Lancker, Kreiman, and Wickens, 1985). We also recognize signature voices; for instance, most of us can identify even a poor impersonation of W. C. Fields or Porky Pig. Finally, the entire realm of changes we call "tone of voice" is pervasive in communication and is readily perceived as anger, depression, or joy. Occasionally, tone of voice modifies the semantic content of an utterance, as in a sarcastic comment. Finally, research has demonstrated that listeners incidentally store detailed information about speakers' voices and implied connotative states when listening to speech

(Geiselman and Bellezza, 1977; Geiselman and Crawley, 1983).

These facts raise two apparently contradictory questions in consideration of variability among talkers. The first concerns the listener's ability to recognize the segments of the language despite the idiosyncratic variability introduced by each new voice. The second concerns the listener's ability to exploit such variability to perceive the characteristics of the talker and the communicative situation.

Talker Variability in Speech Perception and Word Recognition

Although Joos (1948) described the problem of talker variability, one of the first empirical demonstrations of its effects was provided by Ladefoged and Broadbent (1957; see also Peters, 1955a, 1955b). Ladefoged and Broadbent presented listeners with the synthesized sentence "Please say what this word is:" followed by "bit", "bet", "bat", or "but". The carrier phrase was altered in different conditions by raising (by 30%) or lowering (by 25%) either the first or second formant or both. This manipulation changed the perceived dimensions of the talker's vocal tract. Ladefoged and Broadbent observed reliable changes in subjects' identification of the target syllables according to the perceived talker. The authors concluded that the carrier phrase allowed the listener to calibrate the vowel space for the talker and to adjust interpretations of the target vowels accordingly (see also Gerstman, 1968). Following this early demonstration, a number of studies sought to investigate and explain the relative constancy of natural vowel perception across talkers (see Shankweiler, Strange, and Verbrugge, 1977; Johnson, 1990). The guiding notion for all such studies was the idea that listeners must somehow extrapolate the entire vowel space of any given talker from a small speech sample (Joos, 1948; Lieberman, Crelin, and Klatt, 1972).

In further research, however, Verbrugge et al. (1976; see also Shankweiler, Strange, and Verbrugge 1977) questioned the premise of this approach. They noted that despite talker variability, listeners' error rates in vowel identification tasks are low (only 4% in Peterson and Barney's 1952 experiments). Verbrugge et al. reexamined vowel identification across talkers and found that accuracy is generally quite high despite talker variability. They also found that providing examples of a speaker's point vowels did not improve listeners' performance, contrary to the notion of calibration. Finally, they found that listeners adjust their criteria according to perceived rate of articulation much more than to perceived length of vocal tract. Verbrugge et al. (1976) concluded that talker normalization either requires very little prior information or does not occur in speech perception at all (see also Strange et al., 1976). Instead they suggested that adjustment to talkers may have more to do with tracking articulatory dynamics than with frequency-based calibration (see also Green, Stevens, and Kuhl, 1994). Verbrugge and Rakerd (1986) presented listeners with /bVd/ syllables spoken by males and females. The syllables had the middle 60% removed, leaving only the beginning and ending transitions with silence in between. Their results showed that considerable vowel identity information is contained in the transitions and that this information is independent of the talker. Verbrugge and Rakerd (1986, p. 56) concluded, "This strongly suggests that a dialect's vowels can be characterized by higher-order variables (patterns of articulatory and spectral *change*) that are independent of a specific talker's vocal tract dimensions."

The fundamental claim of these reports—that talker normalization involves recovery of underlying articulatory dynamics—is familiar. These findings imply that variability introduced from individual talker characteristics may be resolved in the same manner as all other acoustic-phonetic variability. This treatment of perceptual normalization finds support from studies of development as well. Experiments conducted by Kuhl (1979) and by Kuhl and Miller (1982) demonstrated that 6-month-old infants could accurately discriminate vowels produced by three different talkers; the infants did not attend to the changing voice characteristics of the stimuli but focused on vowel identity instead. Holmberg, Morgan, and Kuhl (1977) obtained similar results, finding that infants' discrimination of fricatives was not affected by talker variability (however, see Carrell, Smith, and Pisoni, 1981). Finally, Jusczyk, Pisoni, and Mullennix (1992) examined the effects of talker variability on infants' discrimination of CVC syllables. They found that infants could discriminate syllables, such as /bug/ and /dug/, equally well in single-talker and multiple-talker conditions. All of these

findings are consistent with a notion that some specialized system, perhaps a phonetic module sensitive to articulatory gestures, is involved in talker normalization, just as hypothesized with regard to more general problems of perceptual constancy in speech (e.g., Liberman and Mattingly, 1985, 1989).

Nonetheless, the data pertaining to effects of talker variability are equivocal. Essentially, the claims of noneffects of talker variability come from tasks of perceptual identification or discrimination with few attentional or time constraints. Despite these demonstrations of the listener's remarkable accuracy in perceiving speech from varying talkers, several experiments have shown reliable effects of talker variability on speech perception and word recognition. Creelman (1957) investigated the effects of talker variability on the recognition of phonetically balanced words that were presented in lists of tokens spoken by 1, 2, 4, 8, or 16 talkers. Perceptual identification of these words in noise showed that words in lists produced by 2 or more talkers were recognized slightly less accurately (differences on the order of 7% to 10%) than words in the single-talker list.

Later experiments with larger sets of stimulus materials have shown larger effects of talker variability. Summerfield and Haggard (1973; see also Summerfield, 1975) observed slower reaction times to recognize spoken words in multiple-talker blocks than in single-talker blocks. Mullennix, Pisoni, and Martin (1989) investigated the effects of talker variability on word recognition using a large sample of CVC monosyllables. Words were presented in lists spoken by either one talker or by 15 talkers, and subjects performed either perceptual identification of words in noise or auditory naming of nondegraded words. Mullennix, Pisoni, and Martin observed large and reliable effects of talker variability. Word recognition was slower and less accurate with multiple talkers than within single talkers. Moreover, as shown in Figure 8-10, talker variability was a more robust effect and was less sensitive to changes in task demands than other variables known to affect word recognition, such as **word frequency** and neighborhood density (see Luce, 1986; Luce, Pisoni, and Goldinger, 1990). Finally, Mullennix, Pisoni, and Martin also found that talker variability interacted with signal degradation, implying that noise and talker variability affect a common stage of processing. From these data they suggested that talker variability affects early stages of speech perception responsible for immediate phonetic perception.

Talker Variability in Memory and Attention

Recent experiments in memory and attention provide further insights into the nature of talker variability effects. Martin et al. (1989) investigated serial recall of 10-item word lists spoken by either a single talker or by 10 different talkers and found that recall of multiple-talker lists was less accurate than recall of single-talker lists, but only for items in early list positions. Moreover, they found that recall of digits visually presented before the spoken lists was less accurate if the subsequent lists were multiple-talker lists than if they were single-talker lists. Finally, they found that the differences in recall between single- and multiple-talker lists were unaffected by a post-perceptual distractor task (following Peterson and Peterson, 1959). From these converging lines of evidence, Martin et al. suggested that word lists produced by multiple talkers require greater attention for rehearsal in working memory than the same lists produced by a single talker.

Further evidence of the attention-demanding nature of talker variability was provided by Goldinger, Pisoni, and Logan (1991), who presented for serial recall single-talker and multiple-talker lists at varying speeds. They found that talker variability interacted strongly with presentation rate, whereas other stimulus variables, such as word frequency, did not (Fig. 8-11). At relatively fast presentation rates, recall of single-talker lists was superior to recall of multiple-talker lists, as in the Martin et al. experiments. At very slow rates, however, recall of multiple-talker lists was more accurate than recall of single-talker lists, suggesting that voice information is retained in long-term memory (see Schacter and Church, 1992; Palmeri, Goldinger, and Pisoni, 1993; Church and Schacter, 1994). The rate manipulation has long been assumed to affect the rehearsal processes of the recall task (Murdock, 1962; Rundus, 1971), so this result suggests that talker variability taxes these attention-demanding stages of processing. Another interesting finding reported by Lightfoot (1989) was that subjects' familiarity with the talkers' voices also modifies the differences in recall of single- and multiple-talker lists. When subjects were trained to recognize the voices of

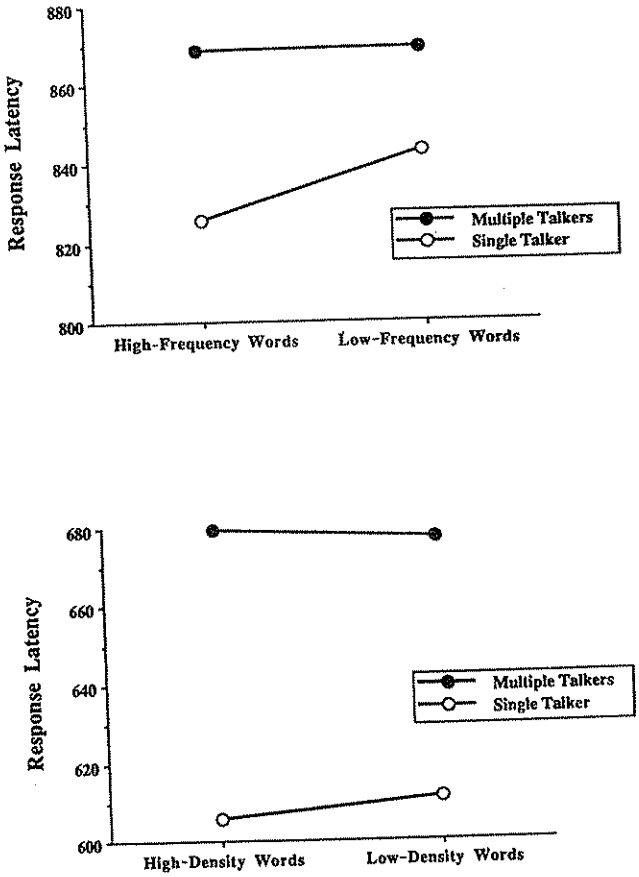


FIGURE 8-10 Auditory word naming data. The upper panel displays mean naming latencies for words from single- and multiple-talker stimulus sets as a function of word frequency. The lower panel displays mean naming latencies for words from single- and multiple-talker stimulus sets as a function of similarity neighborhood density. The effects of talker variability are more robust than the effects of either of the other stimulus variables. (From Mullennix, J. W.,Pisoni, D. B., & Martin, C. S. [1989]. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365-378.)

the various talkers and associate them with fictional names (Brad, Mary, Jane, Sam, etc.), multiple-talker lists were recalled better than single-talker lists, even at relatively fast presentation rates. Recently, Nygaard, Sommers, and Pisoni (1994) used the same procedure to show that familiarity with a speaker's voice improves recognition of novel words produced by the speaker.

Jusczyk, Pisoni, and Mullennix (1992) further elucidated the effects of talker variability on memory. They observed, as Kuhl and her colleagues reported earlier (Holmberg, Morgan, and Kuhl, 1977; Kuhl, 1979; Kuhl and Miller, 1982), that infants recognize phonemic constancy very well despite variation of the stimulus voices. However, Jusczyk, Pisoni, and Mullennix also employed a variation of the high-amplitude sucking (HAS) procedure (Eimas et al., 1971) that included a 2-minute delay between the habituation to one syllable and the presentation of a new syllable. This manipulation let them assess the effects of talker variability on infants' ability to encode and remember phonetic structure. They found that infants who heard speech from a single talker were able to detect a phonetic change across the 2-minute delay but the infants who heard speech from multiple talkers were not. These results, taken together with the adult data, suggest that maintaining perceptual constancy across talkers requires extra attention.

Mullennix and Pisoni (1990) demonstrated the influence of talker variability on selective attention. They employed the Garner (1974) speeded classification procedure to investigate

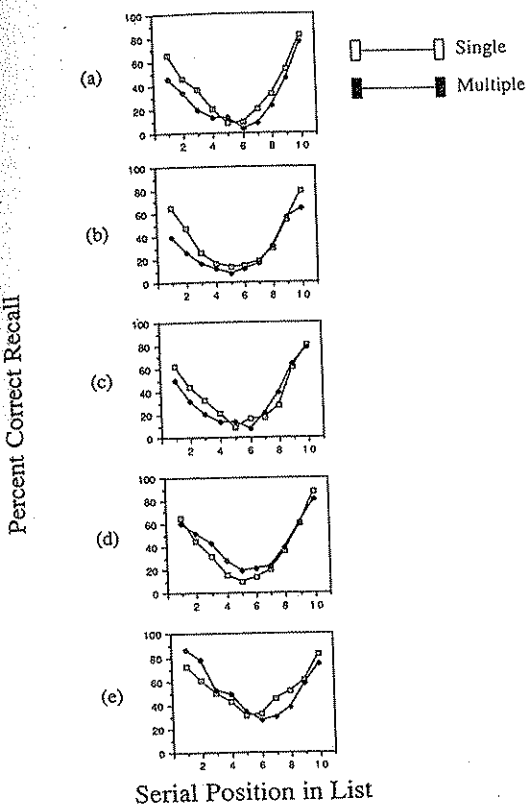


FIGURE 8-11 **Serial recall data. The five panels display recall of single- and multiple-talker word lists presented at five different rates. One word was spoken every (a) 250 ms, (b) 500 ms, (c) 1000 ms, (d) 2000 ms, or (e) 4000 ms. The single-talker advantage in primacy recall is reversed at the slower ISIs.** (From Goldinger, S. D., Pisoni, D. B., & Logan, J. S. [1991]. The nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152-162.)

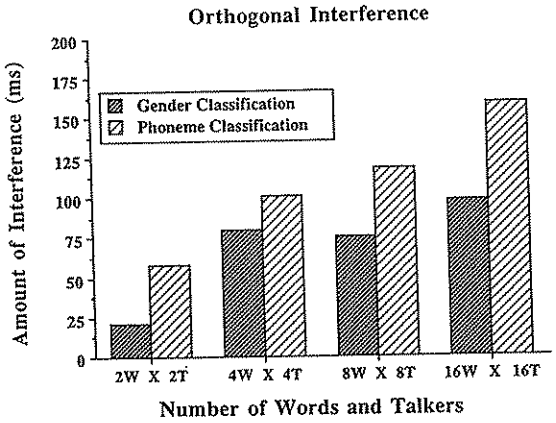


FIGURE 8-12 **Garner speeded classification data. The dark bars show the amount of interference that phonetic variability caused in gender classification; light bars show the amount of interference that talker variability caused in phonetic classification. Across all stimulus sets, the dimensions of voice and phoneme were perceived integrally.** (Adapted from Mullennix, J. W., & Pisoni, D. B. [1990]. Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379-390.)

processing dependencies between phonetic variability and talker variability. Subjects classified monosyllabic words according to either the voicing of the initial phoneme (/b/ versus /p/) or the gender of the talker. They found that irrelevant variations in phonetic constitution or voice could not be ignored; variation along either dimension slowed classification along the other dimension. However, a large asymmetry was observed, showing that variability along the voice dimension impaired classification along the phonetic dimension more than vice versa. These data, shown in Figure 8-12, suggest that the processing of voice information and phonetic information are qualitatively different but also depend on one another, sharing a limited-capacity cognitive system (see Cutting and Pisoni, 1978). Mullennix and Pisoni suggest that both indexical information and phonetic information are processed in a mandatory fashion, following Fodor (1983; for a similar suggestion see Miller, 1987). However, the implied modules may function as a cascade system (McClelland, 1979), such that the output of the phonetic module is more strongly affected by the output of the voice module than vice versa (see also Nusbaum and Morin, 1992).

In summary, the available data on the effects of talker variability in speech perception, word recognition, attention, and memory all indicate that indexical information deserves more thorough consideration in theoretical discussions of speech perception than it has traditionally received. Talker-related information affects perception of speech and memory of spoken material, attracts selective attention, and is routinely encoded in parallel with linguistic information (Geiselman and Bellezza, 1976, 1977; Palmeri et al., 1993). The traditional approach to the study of speech perception has considered only abstract linguistic units without regard to the media that carry them. Further investigation into the generality and nature of normalization effects in speech should provide valuable insights into speech perception and perhaps the architecture of general perceptual systems as well.

Prosody and Timing in Speech Perception

Another neglected topic is the role of prosodic information in language perception. *Prosody* is the melody, timing, rhythm, and amplitude of fluent speech, and it is typically thought of as changes in the acoustic correlates of stress, such as fundamental frequency and vowel duration (Lehiste, 1970; Huggins, 1972). Most of the emphasis in speech perception research and theory has been on the segmental analysis of phonemes, whereas **suprasegmental information** has received only cursory consideration. Although the role of prosody has been researched more vigorously in recent years, a wide gap remains between research on the perception of isolated segments and features and on sentences with full prosody and natural rhythm (see Cohen and Nooteboom, 1975). However, it is becoming apparent that prosodic factors may link phonetic segments, features, and words to grammatical processes at higher levels of analysis. Moreover, prosody seems to provide useful information about the lexical, syntactic, and semantic content of the spoken utterance. We briefly review several findings that illustrate the importance of prosodic information in the perception of connected speech (Huggins, 1972; Darwin, 1975; Nooteboom, Brokx, and de Rooij, 1978; Studdert-Kennedy, 1980).

Differences in fundamental frequency can provide important cues to the proper parsing of speech into constituents for syntactic analysis. In acoustic analyses of connected speech, Lea (1973) found that a drop in fundamental frequency usually occurred at the end of each major syntactic constituent of a sentence, and a rise in fundamental frequency occurred in the beginning of the following constituent. In more detailed analyses, Cooper and Sorenson (1977) found reliable rise-fall patterns at the boundaries between the main clauses of a sentence, between main and embedded clauses, and between major phrases. Lindblom and Svensson (1973; see also Svensson, 1974) have shown that listeners can parse speech that is devoid of segmental cues but maintains prosodic integrity (see also Nakatani and Schaffer, 1978). These findings and others (Collier and t'Hart, 1975; Klatt and Cooper, 1975; Cooper, 1976; Klatt, 1976) demonstrate the importance of prosody as a cue to phrasal grouping.

Another function of prosody is the maintenance of perceptual coherence (Studdert-Kennedy, 1980). As an example, Darwin (1975) had listeners shadow a sentence played to one ear while another sentence was presented to the other ear. At some point the prosodic contours of the sentences were switched, but their lexical, syntactic, and semantic content remained unchanged. Shadowing often spontaneously followed the prosodic contour across ears rather than the syntax or semantics of the message to which subjects were originally attending. Nooteboom, Brokx, and de Rooij (1978) suggest that prosodic contours maintain the "perceptual integrity" of the signal and provide evidence that the continuity of fundamental frequency and formant frequencies underlies this integrity (see Bregman, 1978 and 1990; Remez et al., 1994).

Cutler and her colleagues (1976, 1977, 1979, 1981) demonstrated yet another important function of prosody in speech perception. Cutler has shown that prosodic contours enable listeners to predict where sentence stress will fall. Because sentence stress is usually placed on words of primary semantic importance, the ability to predict stress placement presumably guides attention to the most important words in the sentence. Thus, prosody appears to guide attention to high-information stretches of fluent speech. To demonstrate that attention follows the predicted sentence stress, Cutler and her colleagues demonstrated faster phoneme-monitoring reaction times for words predicted by prosodic contour to receive stress, regardless of the word's actual acoustic realization or form class. A word in a sentence position of predicted stress is responded to faster than the same recorded token in another sentence position.

These demonstrations of the role of prosody in guiding attention have led Cutler and others to propose accounts of word recognition in which prosody is considered a primary source of information rather than marginally relevant variability (see Cutler, 1976, 1989; Grosjean and Gee, 1987). These approaches all emphasize the prominence of strong syllables in fluent speech and suggest that such syllables may focus attention and initiate segmental analysis and lexical access. This approach, recently dubbed the *metrical segmentation strategy* (Cutler and Butterfield, 1992; Cutler et al., 1992; McQueen, Norris and Cutler, 1994), contrasts with more temporally constrained left-to-right models of speech perception and word recognition such as

cohort theory (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980; Marslen-Wilson, 1987) , that assume word beginnings are necessarily processed first. As Cutler (1989), p. 354) says:

> The major problem for lexical access in natural speech situations is that word starting points are *not* specified. The evidence presented here has shown how prosodic structure, in particular metrical prosodic structure, can offer a way out of this dilemma. Where do we start with lexical access? In the absence of any better information, we can start with any strong syllable.

Finally, all of these useful prosodic cues make acoustic-phonetic invariance far more problematic. The durations of phonetic segments vary widely across stressed and unstressed syllables and in varying syntactic environments (Oller, 1973; Klatt, 1974, 1975; Luce and Charles-Luce, 1985). Spoken stress also entails wide spectral variations in formant frequencies and fundamental frequency (Lehiste, 1970). The durational variations of speech timing provide useful cues to lexical identity and syntactic structure, but at the cost of further removing anything resembling canonic phonemes from the signal. This potentially contradictory nature of prosodic information underscores the necessity of some theoretically sound resolution of the problem of invariance—apparently, as more meaningful variation is added to the signal, perception is improved rather than impaired.

This chapter has identified and discussed many of the long-standing issues in speech perception as well as several issues that researchers have recently explored. We now focus on individual theories and models of speech perception. We briefly introduce and comment on only a few models in the literature (see Klatt [1989] for a more extensive review), some of the most important and influential classes of theories, particularly those that should figure prominently in future research.

THEORETICAL APPROACHES TO SPEECH PERCEPTION

The perception of spoken language, encompassing all processes from peripheral auditory coding of the speech signal to comprehension of the message, is very complex. Many sources of knowledge and multiple levels of representation interact in myriad combinations. To date, the complexity of language has precluded the formulation of theories of language perception that are both global and empirically testable. Therefore, the situation in language perception research is similar to other areas of cognitive science: most investigators have examined only the details of specific phenomena and paradigms rather than more complex or integrative issues (Newell, 1973).

The remaining sections of this chapter illustrate this situation by their unfortunate dichotomy. In this section we review several models of speech perception, and in the following section, several models of spoken word recognition. This segregation is largely due to the orientations of the models themselves. Although there are a few notable exceptions (e.g., Klatt's LAFS model and the TRACE model), most of these models were formulated to explain either the identification of phonemes in the speech signal or the mapping of phonemic strings onto lexical representations in memory. Very few models specify the integrated processes of speech perception and word recognition (Pisoni and Luce, 1987). One trend, especially in the connectionist movement, is toward grouping these processes into unitary models. Another trend is toward justifying the segregation of processes considered in different models by arguing that the processes are segregated in perception. The concept of the phonetic module (Liberman and Mattingly, 1985, 1989) clearly justifies narrow consideration of phonetic perception without regard to the mapping of speech representations onto lexical representations. Although we believe the former trend will prove more fruitful in the long run, we recognize the value of the earlier models and discuss them next, beginning with the most influential of all models of speech perception, the motor theory.

Motor Theory of Speech Perception

The original motor theory described by Liberman et al. (1967, p. 452) was based on the assumption that ". . . speech is perceived by processes that are also involved in its production." This view of speech perception was motivated by the fact that a listener is also a speaker, and a close link exists between the acoustic forms of speech sounds and their underlying articulation. Therefore, an effective and economical means of perceiving speech is to perceive the articulatory gestures that produce sounds. Advocates of the

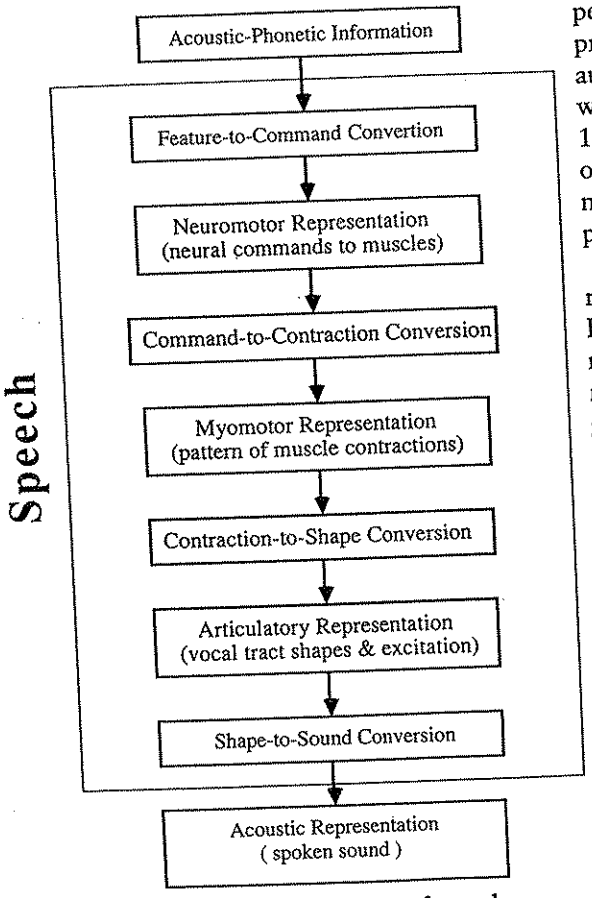


FIGURE 8-13 The motor theory of speech perception. The theory posits conversion of acoustic-phonetic information to a speech representation via articulatory knowledge. (From Cooper, F. S. [1972]. How is language conveyed by speech? In Kavanagh, J. F., & Mattingly, I. G. (Eds.), *Language by ear and by eye.* Cambridge, MA: MIT Press.

motor theory argue that a solution to the invariance problem lies in the more reliable nature of articulatory gestures (compared with acoustic phonemes) as units of perception (Fig. 8-13).

Although, for many years, the original motor theory has held a dominant position in accounts of speech perception, the link between the theory and the data is rarely more than suggestive. As the review of evidence in the section *Specialization of Speech Perception* showed, the evidence in support of motor theory is ambiguous. For example, much of the early support for motor theory came from the finding that synthetic stop consonants were perceived categorically, whereas steady-state vowels were perceived continuously, apparently paralleling their respective articulatory origins. However, subsequent research showed that the differences in perception between consonants and vowels were primarily due to their differing demands on auditory short-term memory (Fujisaki and Kawashima, 1969, 1970, 1971; Pisoni, 1971, 1973, 1975). This general cognitive explanation of the continuous-categorical distinction eliminated the need to appeal to articulation for the perception of stops.

The motor theory has been revised in two key regards (Liberman and Mattingly, 1985, 1989). First, whereas the original model was based on recognition of observable gestures, the revised model is based on perception of intended gestures. Gestures are a set of movements by the articulators that result in a phonetically relevant vocal tract configuration. Each intended gesture of the language has properties that specify it uniquely, and each intended gesture is invariant, such that each segment of the language maps uniquely to a distinctive gesture. The second important modification is that gestures are perceived directly (following Gibson, 1966) by an innate phonetic module.

The revised motor theory makes four basic claims considered in turn by Klatt (1989). The first claim is that speech production and perception are linked psychologically so that they share common representations and processes. Second, the basic unit of speech perception is the underlying intended articulatory gesture associated with a phonetic segment rather than the actual physical motions implied by the acoustics. Third, perception of the intended gesture is direct, performed by a specialized module. Fourth, the model is supported by the claim that no other model can account for the the wide array of phenomena to which the motor theory has been applied over the years.

With respect to the link between production and perception, Klatt (1989) agrees that the processes must be linked in some sense (as inverses, at least), but he also notes that there is no simple way to relate the processes to make articulatory perception any easier than acoustic perception. Considering the direct perception of intended gestures, Klatt notes that while the position is attractive and would solve many problems of variability, no mechanisms described in the theory can perform this feat. Furthermore, Klatt argues that technology demonstrates the extreme difficulty of determining vocal tract shapes from speech acoustics, but the motor theory is based on faith that this transformation is possible. In contrast to the premises of Newtonian mechanics, we cannot be certain

that speech is a reversible event. Finally, regarding the uniqueness of the motor theory in accounting for a wide range of phenomena, Klatt argues that the revised motor theory is so abstract that it is essentially no different from auditory theories such as LAFS, and he therefore suggests that the account is no longer unique. Klatt (1989, p. 180) concludes, "An attractive motor theory *philosophy* has been described by Liberman and Mattingly, but we are far from the specification of a motor-theory *model* of speech perception." His point is well taken; the motor theory and the revised theory are based primarily in logic, parsimony, intuitive appeal, and a measure of faith, rather than empirical support.

Direct-Realist Approach to Speech Perception

Fowler (1986, 1990) and Fowler and Rosenblum (1990, 1991) have outlined the framework for a *direct-realist* approach to speech perception. This approach assumes that, as in Gibson's (1966) view of visual perception, speech perception entails the recognition of natural phonetic events. As in the motor theory, Fowler assumes that the relevant events perceived in speech are the speaker's phonetically structured articulations. In the language of event perception, there is a fundamental distinction between the event and the informational medium. For example, an object such as a chair is an event in the world. When our eyes gaze upon the chair, we perceive it via light that is structured by the edges, contours, and colors of the chair. We do not perceive the light per se. Instead, the light is merely the medium by which the chair is perceived. The suggestion for speech perception is very similar to this example—articulatory events lend unique structure to the acoustic waveform, just as chairs lend structure to light. Accordingly, Fowler suggests that articulatory events are directly perceived via the acoustic medium.

The direct-realist approach to speech perception is similar to the motor theory in many respects. However, there are important differences. Most notably, the two theories approach the signal in different ways. Motor theory maintains that the acoustic signal is subjected to computations to retrieve underlying gestures. In contrast, the direct-realist approach maintains that no cognitive mediation whatsoever is necessary; the acoustic signal is "transparent" with respect to the underlying structure of speech (Liberman and Mattingly, 1985). Following this difference, Fowler and Rosenblum (1991) argue that phonetic perception need not be modular, suggesting instead that general perceptual principles can be invoked to perceive the distal events of speech.

The direct-realist approach and motor theory are attractive for many of the same reasons (Studdert-Kennedy, 1986): Direct-realism has intuitive appeal, and it fares well with many of the data that the motor theory can explain. Moreover, it stems from a respected tradition of event perception theories. However, it must meet many challenges. Most important, of course, is the need for empirical support, in which regard it is similar to the motor theory (although evidence is growing; see Dekle, Fowler, and Funnell, 1992; Fowler, 1994). Forgiving the lack of critical data, many logical and theoretical challenges can be offered as well (see commentaries on Fowler's [1986] target article; Diehl and Kleunder, 1989). For example, as Remez (1986) notes, it is not clear what the proper perceptual objects in linguistic communication are. Fowler has adopted a physical perceptual object that is capable of structuring the acoustic media—the articulatory gesture — and has made its recognition the central task of speech perception. However, articulations are not ends in themselves. Unlike chairs, articulations are another medium, because language is symbolic. Strings of articulations are perceived as words and ideas, so gesture perception does not fully explain speech perception. Moreover, as noted by Diehl (1986), Porter (1986), and Remez (1986), chairs and gestures are also very different in terms of their perceptual availability. We know unambiguously when we are looking at a chair; we do not have such access to phonetic gestures. A direct-realist theory might claim that our unambiguous recognition of words and sentences implicitly demonstrates our recognition of gestures, but other less circular accounts are available (Massaro, 1986). The resolution of these and other theoretical vagaries, as well as the further collection of relevant data, will be important to the direct-realist position.

Information-Processing Theories of Speech Perception

Perhaps the polar opposite to the direct-realist perspective is the information-processing perspective. The theories and models that fall into this category are oriented toward general cog-

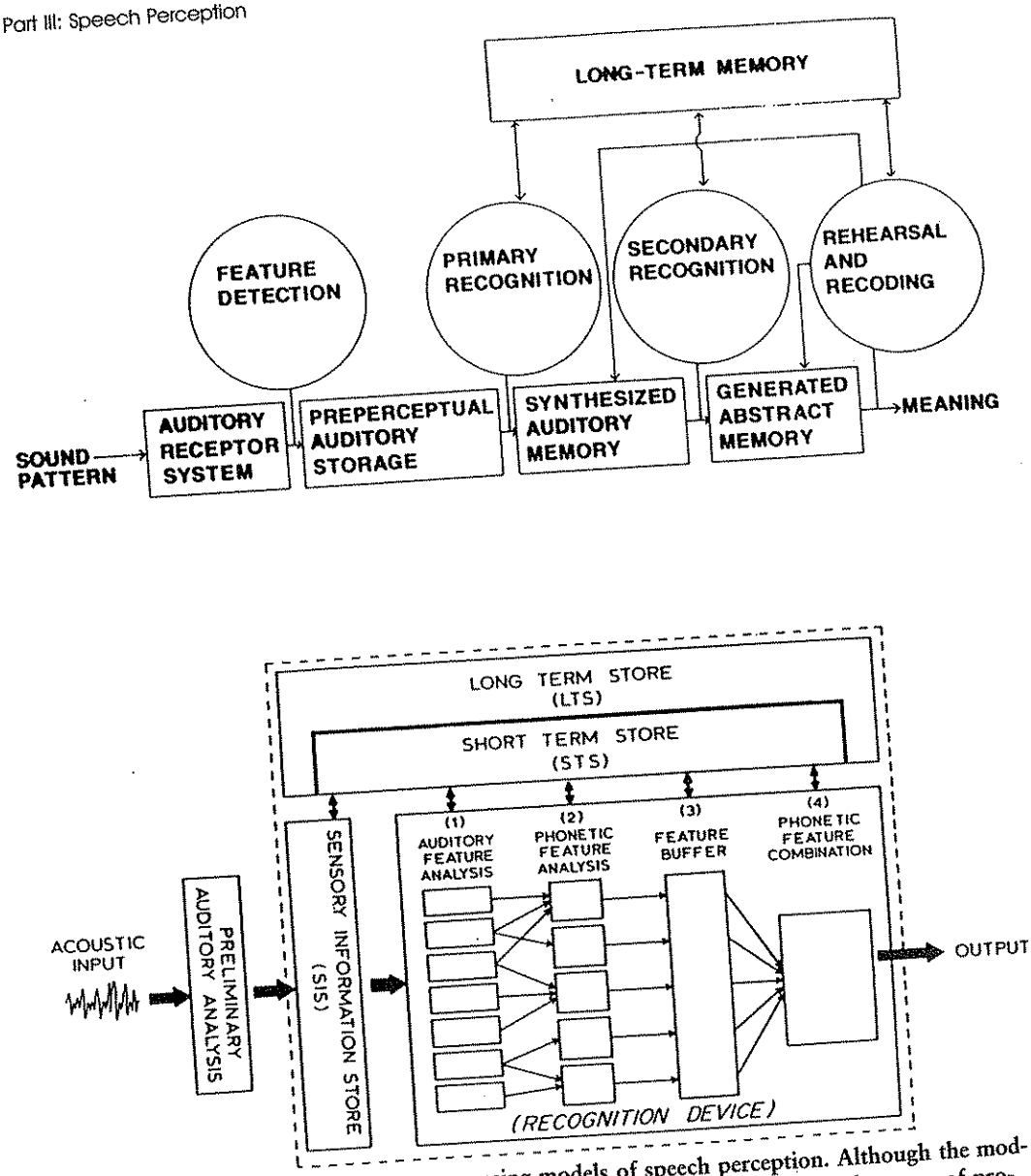


FIGURE 8-14 Two information-processing models of speech perception. Although the models differ in detail, both entail the information-processing elements of multiple stages of processing, hierarchical levels of representation, and strong interactions among perceptual buffers and memory stores. (Top panel from Oden, G. C., & Massaro, D. W. [1978]. Integration of featural information in speech perception. *Psychological Review, 85,* 172-191. Bottom panel from Pisoni, D. B., & Sawusch, J. R. [1975]. Some stages of processing in speech perception. In Cohen, A., & Nooteboom, S. G. (Eds.), *Structure and process in speech perception.* Heidelberg: Springer-Verlag [pp. 16-34].)

nition and perception. All of these models assume distinctive, hierarchically organized levels of processing. Moreover, all or most of these theories assume that limited-capacity perceptual and memory stores are intimately involved in speech analysis (Cutting and Pisoni, 1978). This view contrasts sharply with the revised motor theory and the direct-realist framework. Two typical information-processing stage models are shown in Figure 8-14. Both exhibit multiple levels of representation and processes that interact with and depend upon memory stores and control processes. Beyond the stages of processing shown in the figure, once the information has followed the meaning or output arrow, the recognized linguistic units enter still higher

stages of processing that derive syntactic and semantic content.

Studdert-Kennedy (1974, 1976) was the first to advocate an approach to speech perception based on stages of perceptual processing. He proposed four stages of speech processing: (1) auditory, (2) phonetic, (3) phonological, and (4) lexical, syntactic, and semantic (see review and discussion by Pisoni and Luce, 1986, 1987; Luce and Pisoni, 1987). As the division of Studdert-Kennedy's stages imply, this approach to speech perception synthesizes information-processing psychology and linguistic theory. An advantage of this framework is the clear division of processes of speech perception; working within such a framework provides a well-defined division of topics for investigation.

The basic appeal of stage theories is their reliance on generally accepted mechanisms of cognition and perception. As such, information-processing models introduce certain advantages over modular theories such as motor theory. For example, they can account for the effects of reduced attention or increased memory load on speech perception (Nusbaum and Schwab, 1986). Unfortunately, like many theories of speech perception, information-processing theories have typically been quite vague and not subject to direct empirical tests.

Klatt's LAFS Model

Although we introduce Klatt's lexical access from spectra (LAFS) model in this section, LAFS is a model of spoken word recognition as well. LAFS is one of the few models that successfully addresses several critical issues of speech perception along with access to the mental lexicon and the nature of lexical representations in long-term memory.

LAFS assumes direct, noninteractive access to lexical entries based on context-sensitive spectral sections (Klatt, 1979). It also assumes that adult listeners have a dictionary of all legal diphone sequences stored in memory. Associated with each diphone sequence is its prototypical spectral representation. These spectral representations are proposed to resolve problems associated with contextual variability of individual segments. In a sense, LAFS resolves the problems of variability by precompiling coarticulatory effects directly into the representations of an input word and comparing these derived spectra to prototypes in memory. Word recognition is accomplished when a best match is found between the input spectra and the diphone representations. In this portion of the model, word recognition is directly based on spectral representations of the sensory input, with no intermediate levels of computation corresponding to segments or phonemes.

An important aspect of LAFS is its explicit avoidance of any levels of representation corresponding to phonemes. Instead, the model assumes a precompiled, acoustic-based lexicon of words in a network of diphone power spectra. These spectral templates are assumed to be context-sensitive units, similar to "Wickelphones," because they represent the acoustic correlates of phonemes in different phonetic environments (Wickelgren, 1969). Klatt argues that diphone concatenation is sufficient to capture much of the context-dependent variability observed for phonetic segments in spoken words (see also Marcus, 1984). Word recognition in LAFS proceeds similarly to the workings of the computerized HARPY speech recognition system, in that power spectra are computed every 10 ms and compared with the stored representations (see Klatt [1979] for details on HARPY). When finished, the best path through the diphone network is the optimal phonetic transcription of the signal. Klatt's model is an example of an extreme bottom-up recognition process and may be contrasted to more interactive models of word recognition that we consider below, such as cohort theory and TRACE.

Massaro's Fuzzy Logical Model of Perception

Massaro's fuzzy logical model of perception (FLMP) (Massaro, 1972, 1987, 1989; Massaro and Cohen, 1976, 1977, 1993; Oden and Massaro, 1978; Derr and Massaro, 1980) Massaro and Oden, 1980; was developed to account for feature integration in speech perception, regardless of the nature of the relevant features. For example, FLMP can account for the integration of multiple acoustic cues in the speech waveform as well as audiovisual integration. In this brief introduction, we restrict our attention to the recovery of phonemes from the speech signal, noting only that integration of information from other sources is possible in the model and is accomplished by processes similar to those described here.

FLMP assumes three operations in phoneme

identification. First, *feature evaluation* determines the degree to which any given acoustic-phonetic feature is present in a stretch of sound. Unlike more conventional feature detector theories, FLMP assumes that features are evaluated along a continuous scale rather than an absolute feature present-feature absent dichotomy. Features are assigned continuous, "fuzzy" values ranging from 0 to 1, indicating the degree of certainty that the feature appears in the signal (Zadeh, 1965). The second operation in FLMP is *prototype matching*, in which the feature profiles derived by the earlier operations are compared with prototypes of phonemes stored in memory. Phoneme prototypes are stored as sets of propositions that describe ideal representations of the acoustic correlates of each phoneme. The prototype-matching operation specifies the degree of correspondence between ideal phonemes and the input sets of features. The final operation, *pattern classification*, determines the best match between the candidate phonemes and the input by using goodness of fit algorithms. FLMP provides flexibility in pattern classification by using a variety of logical rules for feature integration so that perfect matches between the input and the prototypes are not required for phoneme identification.

FLMP is appealing for several reasons. First, it is a very general framework that demonstrates how acoustic information (as well as other information) can be mapped onto representations in long-term memory without the postulation of specialized speech procedures or modules. In fact, Massaro (1987, 1989) specifically rejects the notions of specialized or modular processes in speech perception. Second, the model argues that speech perception is not necessarily categorical but can be explained by integration of continuously evaluated features. The framework is therefore consistent with the data reported by Barclay (1972), Pisoni (1973), and others that continuous information remains available in speech perception, despite the categorical identification and discrimination functions obtained in typical studies (e.g., Liberman et al., 1957). Finally, FLMP is one of the only models of speech perception proposed in terms of a precise mathematical framework (Townsend, 1989). However, the quantification has been a source of criticism as well as praise. FLMP employs large numbers of free parameters to account for patterns of data, and the parameter settings do not easily transfer across experimental paradigms (Jenkins, 1989; Warren, 1989). Finally, Massaro's 1987 suggestions that the FLMP framework may be extended to all forms of perception are attractive, but considerable testing and evaluation are clearly required by these claims.

THEORETICAL APPROACHES TO SPOKEN WORD RECOGNITION

The theories and models described in the previous section are models of speech perception, meaning that they primarily address phonetic perception, independent of higher-level lexical or linguistic processes (with the exception of LAFS). In this section we introduce several models of spoken word recognition, models primarily concerned with the rapid location of lexical entries in memory once the speech perception system has specified the necessary sublexical components of the input. This separation of the focus of theories is unfortunate and appears inappropriate (Pisoni and Luce, 1987), especially in light of the data on lexical effects in speech perception (Ganong, 1980; Samuel, 1986; Samuel and Ressler, 1986; Nygaard, 1993). Nevertheless, most of the models considered here assume that some input, perhaps resembling a string of phonemes, is provided by early processes of speech perception and is then compared to the mental lexicon until a best match is found. Very few models of word recognition or lexical access (except TRACE) are concerned with the entire range of processes that subserve word recognition.

The myopic nature of theories of word recognition and lexical access is primarily attributable to their origins. Most theories were designed to account for findings in *visual* word recognition, so assumptions of invariance are easily justified, although most models of visual word recognition allow for some variability. A very general assumption has been that models of visual word recognition can account for spoken word recognition as well, given rudimentary modifications to respect the temporal distribution of the speech signal (Marslen-Wilson and Tyler, 1980; Grosjean and Gee, 1987; Tyler and Frauenfelder, 1987; Cutler, 1989). While the validity of this assumption is subject to debate (Bradley and Forster, 1987), it has isolated the processes of word recognition sufficiently to allow for the development of precise, albeit simplified, theories. While ignoring questions related to the problems of early speech percep-

tion, models of word recognition focus primarily on explaining basic phenomena such as word frequency effects, context effects, types of knowledge sources brought to bear on word recognition, and the nature of representations in the mental lexicon. Indeed, these considerations largely characterize models of word recognition and are the basis of extensive experimentation and debate. Furthermore, one of the fundamental debates about models of word recognition is the distinction between modular and interactive processes (Bradley and Forster, 1987; Tanenhaus and Lucas, 1987). In this discussion we pay special attention to the models' respective approaches to all of these basic phenomena and theoretical distinctions.

In this section we briefly examine five models of word recognition:* the logogen theory, cohort theory, Forster's search theory, the neighborhood activation model, and the TRACE model. It should be noted that these are only some of the models described in the literature, but we hope this review will capture and communicate several of the key issues in spoken word recognition. We begin with one of the earliest models of word recognition, the logogen theory.

Logogen Theory

In Morton's (1969, 1979, 1982) **logogen** theory, these passive sensing devices are associated with each word in the mental lexicon. Each logogen contains all of the information about a given word, such as its meaning, possible syntactic functions, and its phonetic and orthographic structure. A logogen monitors discourse for any information indicating that its particular word is present in the signal, and once such information is encountered, the activation level of the logogen is raised. Given sufficient activation, the logogen crosses a threshold; the information about the referent word is made available to the response system and the word is recognized (Fig. 8-15).

Several important features of the logogen theory have been either strongly rejected or incorporated into later models. First is the emphasis on multiple interactive knowledge sources in word recognition. An important feature of the theory is that logogens monitor all possible sources of information, including higher-level semantic and syntactic information from the discourse and lower-level sensory information. (However, logogens do not "talk to each other," meaning that any given logogen is oblivious to the activity levels of other logogens.) Thus, information from several levels can combine to push the activation level of a logogen toward its threshold. In this sense, logogen theory is highly interactive, and context effects are incorporated into the early stages of word recognition. Words that are readily predicted by the semantic and syntactic context are activated and recognized more quickly than those not well predicted by context. A second important feature of the logogen theory is its portrayal of "word frequency effects." It posits that frequency differences among words produce adjustments in the recognition thresholds of their logogens. Thus, a common word has a lower threshold than a rarely used word and therefore requires less sensory or contextual input for recognition. The characterization of word frequency as a direct coding in recognition thresholds, resting activation levels, or activation functions has been adopted in many later models of word recognition (e.g., Marslen-Wilson, 1987).

Taken together, the two major assumptions of the logogen theory place the word recognition stage as the locus of both context and frequency effects. The approach is highly interactive, and its portrayal of context effects has been challenged by theorists who prefer a more modularist approach to language processing (Forster, 1979, 1990; Bradley and Forster, 1987). Likewise, the theory characterizes word frequency as an integral and automatic aspect of word recognition. However, some theorists argue that word frequency may be better characterized as a form of perceptual or response bias, as demonstrated by the task-dependent magnitude of frequency effects (Balota and Chumbley, 1984; Luce, 1986).

The details of logogen theory have changed somewhat over the years, but the basic mechanisms have remained the same. For example, Morton (1982) divided the logogen system into separate visual and auditory subsystems, but the fundamental notion of the passive threshold device that monitors information from a variety of sources has remained. Unfortunately, logogen

*In the remainder of this chapter, the following distinction is employed. *Word recognition* means only the recognition of an acoustic-phonetic pattern as a token of a given word held in memory. *Lexical access* is the moment when all information about the recognized word becomes available to working memory (see Morton, 1969; Pisoni and Luce, 1987).

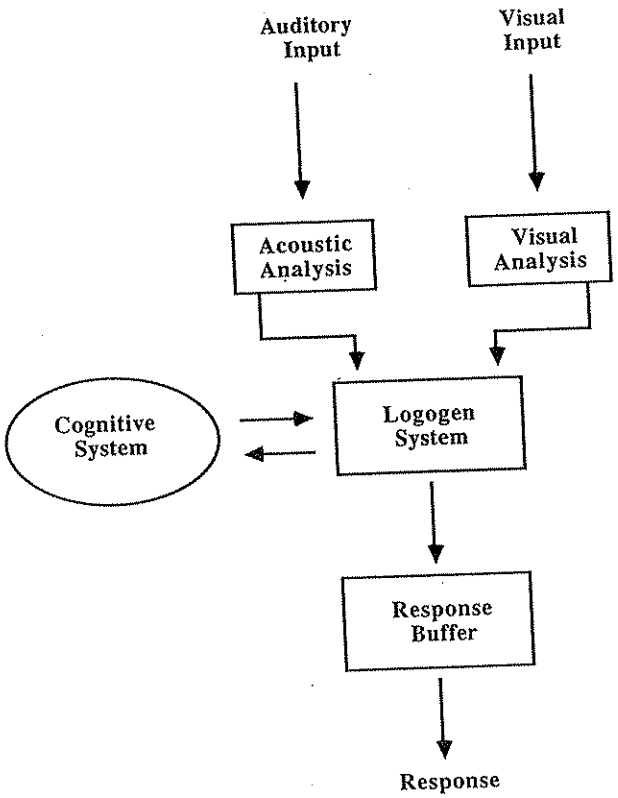


FIGURE 8-15 The logogen theory of word recognition. The theory emphasizes the central role of the logogen system and its interaction with the general cognitive system. The logogen system accounts for frequency effects, and the cognitive system accounts for context effects. (Adapted from Morton, J. [1979]. Word recognition. In Morton, J., & Marshall, J. D. (Eds.), *Psycholinguistics 2: Structures and processes.* Cambridge, MA: MIT Press [pp. 109-156].)

theory is rather vague. It helps us conceptualize how an interactive system works and how word frequency may operate, but it says little about precisely how acoustic-phonetic and higher-level sources of information are integrated, the time course of word recognition, or the structure of the lexicon.

Cohort Theory

Marslen-Wilson's **cohort** theory (Marslen-Wilson, 1975, 1980b, 1987; Marslen-Wilson and Tyler, 1975, 1980; Marslen-Wilson and Welsh, 1978) posits two stages in word recognition, one autonomous and one interactive. In the first autonomous stage, acoustic-phonetic information at the beginning of an input word activates all words in memory that have the same word-initial information. For example, if "slave" is presented to the system, all words beginning with /s/ are activated. The words activated on the basis of word-initial information constitute a cohort. Activation of a cohort is autonomous in the sense that only acoustic-phonetic information can specify it. At this stage, which Marslen-Wilson (1987, 1990, 1993) calls *access,* word recognition is a completely data-driven, or bottom-up, process.

Once a cohort is activated, all possible sources of information come to bear on the selection of the appropriate word. Thus, further acoustic-phonetic information may eliminate "sight" and "save," leaving only words that begin with /sl/, such as "sling" and "slave." Note that access is based on acoustic-phonetic information and is assumed to operate in a strictly left-to-right temporal fashion. At the later integration stage of word recognition, however, higher-level knowledge may also eliminate candidates from the cohort. Thus, if "sling" is inconsistent with the available semantic or syntactic information, it will be eliminated from the cohort. At the integration stage of word recognition, the theory is highly

/s/	/l/	/e/	/v/
sight	slow	sleigh	SLAVE
safe	slip	slave	
sing	slack	•	
so	slide	•	
sound	slave		
slip	•		
store	•		
•			
•			

FIGURE 8-16 Elimination of hypothesized lexical candidates from the word-initial cohort for "slave."

interactive.* Figure 8-16 shows the elimination process: Upon isolation of a single word in the cohort, word recognition is accomplished.

An important feature of cohort theory is its sensitivity to the temporal nature of speech. It gives priority to the beginnings of words and assumes strict left-to-right processing of acoustic-phonetic information. Cohort theory also embraces the notion of optimal efficiency (Marslen-Wilson, 1980a, 1987; Tyler and Marslen-Wilson, 1982), the principle that the word recognition system selects the appropriate word candidate from the cohort at the earliest possible point (the *recognition point*). This means that the word recognition system will commit to a decision as soon as sufficient acoustic-phonetic and higher-level sources of information are consistent with a single candidate.

Although earlier discussions of cohort theory made no mention of word frequency, Marslen-Wilson (1987, 1990) suggested that frequency in cohort theory operates similarly to the logogen theory. Specifically, Marslen-Wilson proposed that word recognition may not require absolute elimination of all members of a cohort but merely a comparison of relative activation levels among candidates (following Luce, 1986), with the activation levels modified by the activation-elimination processes described above. Word frequency is assumed to modify the individual rates of activation of the words constituting the cohort, with common words becoming active faster than rarely used words. Like the logogen theory, then, cohort theory portrays word frequency as an integral aspect of the early phases of word recognition.

Marslen-Wilson's cohort theory has attracted considerable attention for several reasons, including its relatively precise description of the word recognition process, its novel claim that all relevant words in the mental lexicon are activated in the initial stage of access, and the priority it affords to word beginnings, a popular notion in the literature (Cole and Jakimik, 1980). However, the theory is not without its shortcomings, both theoretically and empirically. First, Marslen-Wilson (1987, 1989; Warren and Marslen-Wilson, 1987) has argued that the theory requires no conventional linguistic units, such as phonemes, in order to function. He proposes that to maintain optimal efficiency the word recognition system exploits coarticulatory information that crosses phonemic boundaries (e.g., nasalization of a vowel preceding a nasal consonant) in real time, avoiding unnecessary decisional delays. Unfortunately, the data on this point are ambiguous, and the argument could be made that nasalization of a vowel is primarily a cue to phonemic rather than lexical identity. Affording priority to the lexical cue may be efficient, but it may not be correct. Nor is it clear that candidates can be efficiently eliminated from the cohort without the use of phonemic dichotomies (see Pisoni and Luce [1987] for further discussion).

Another problem with cohort theory is error recovery. For example, if "foundation" is perceived as "thoundation" due to mispronunciation or misperception, the word-initial cohort will not, according to the theory, contain the word candidate "foundation." Although Marslen-Wilson allows for some residual activation of acoustically similar word candidates in the cohort so that a second pass through the cohort structure may occur, it is still unclear how error recovery is accomplished when the intended word is not a member of the original activated cohort.

Finally, several studies have challenged some of cohort theory's stronger assumptions, especially the concept of maximally early decisions in word recognition. For example, although preliminary evidence from the gating task

*At least to a degree. In his revisions, Marslen-Wilson (1987) suggested that the effects of top-down context on the word selection process may be limited, perhaps so that context can have only a facilitatory effect for consistent words but not an inhibitory effect for inconsistent words.

(Grosjean, 1980; Tyler, 1984) supported the notion of early isolation points, later experiments showed that many words are not recognized until well after their acoustic offsets in continuous speech (Grosjean, 1985; Bard, Shillcock, and Altmann, 1988; Connine, Blasko, and Titone, 1993). Moreover, cohort theory predicts that the time it takes to decide an item is a nonword is a function of its **isolation point,** the point in the stimulus at which the item could not constitute an English word (e.g., "lotato" should be rejected faster than "potavo"). However, Goodman and Huttenlocher (1988) and Taft and Hambly (1986) have shown that lexical decisions are not reliably predicted by isolation points. Despite these problems, cohort theory is one of the most important theories in spoken word recognition, primarily because it was developed to explain spoken rather than visual word recognition, and it therefore respects the temporal nature of speech.

Forster's Autonomous Search Theory

In contrast to logogen and cohort theory, Forster's (1976, 1979) theory of word recognition and lexical access is autonomous in the strictest sense. Whereas Morton's and Marslen-Wilson's theories allow for parallel processing of information, linguistic processing in Forster's theory is completely serial. The theory posits three separate linguistic processors: lexical, syntactic, and message. The latest version of Forster's theory incorporates a fourth nonlinguistic processor, the general processing system (GPS). Forster's model may be considered the word-recognition embodiment of several of Fodor's (1983) principles of modularity in perceptual processing (see Tanenhaus and Lucas, 1987; Forster, 1989, 1990), strongly emphasizing algorithmic, noninteractive processing among separate components that are hierarchically organized.

In the first stage of Forster's model, information from peripheral perceptual systems is submitted to the lexical processor. The processor then searches for an entry in three peripheral access files: an orthographic file for visual input, a phonetic file for auditory input, and a syntactic-semantic file for either form of input. Search of the peripheral files is assumed to proceed in *frequency order,* with higher-frequency words searched before lower-frequency words. Word recognition is accomplished at the level of the peripheral access files, where the input pattern is matched to a stored representation. Once an entry is located in these files, lexical access is accomplished by locating the entry in the master lexicon, where all other information about the word is stored (Fig. 8-17).

Upon location of an item in the master lexicon, information pointing to its location in the master list passes to the syntactic processor, which builds a syntactic structure of the discourse. Information passes from the syntactic processor to the message processor, which builds a conceptual structure of the message. Each of the three processors—lexical, syntactic, and message—can pass information to the GPS. However, the GPS cannot influence processing. Rather, it only incorporates general conceptual knowledge with the output from the linguistic processors in making a decision or response. In Fodor's terminology, the linguistic processors are vertically organized, whereas the GPS is horizontally organized, meaning that the GPS, unlike the linguistic modules, integrates information from many disparate domains.

Forster's theory postulates autonomous, nonpenetrable modules. The lexical processor is independent of the syntactic and message processors, and the syntactic processor is independent of the message processor. Furthermore, the entire linguistic system is independent of the general cognitive system, as Fodor (1983) suggests. This strictly serial and autonomous characterization of language processing means that word recognition and lexical access are not influenced in any way by higher-level knowledge sources and are exclusively bottom-up or data-driven processes. Forster (1979) attempts to explain all forms of context effects as post access, decisional or response biases. However, Forster (1990) posits that word frequency exerts an early effect on word recognition.

Forster's model is attractive because of its relative precision and the apparently testable claims it makes regarding the autonomy of processors. It also describes word recognition and lexical access in the context of sentence processing. In addition, it incorporates a specific mechanism of the word frequency effect—entries in the peripheral access files are organized according to frequency, and search proceeds from high- to low-frequency entries. This notion of the search mechanism lends itself well to empirical testing, although the majority of relevant data reported to date come from experiments in visual word recognition (Forster and Bednall, 1976; Andrews, 1989).

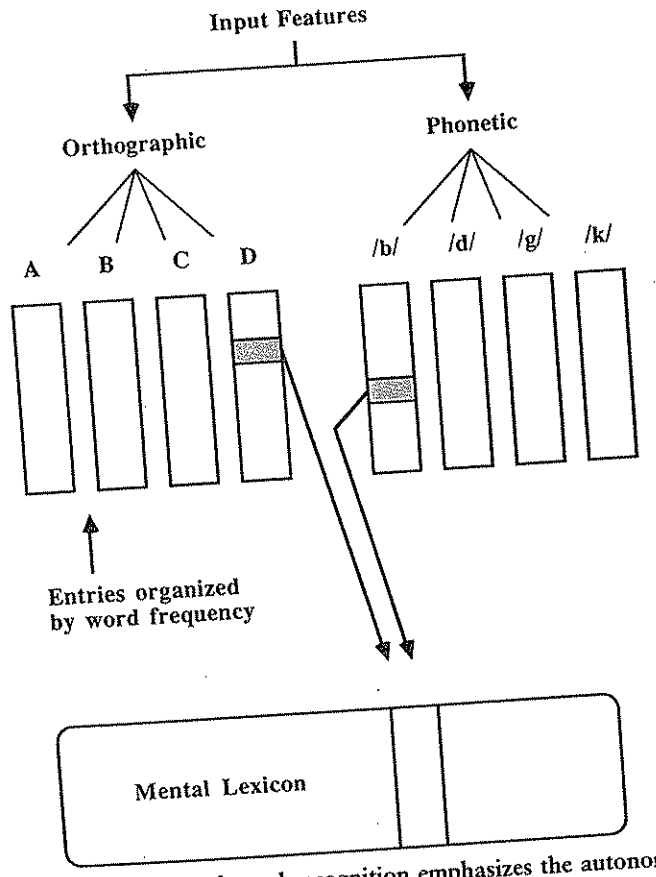


FIGURE 8-17 Forster's model of word recognition emphasizes the autonomy of the search processes and the role of a frequency-ordered search within specific access files. (Adapted from Forster, K. I. [1979]. Levels of processing and the structure of the language processor. In Cooper, W. E., & Walker, E. C. T. (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum [pp. 27-86].)

Neighborhood Activation Model

The neighborhood activation model of word recognition (Luce, 1986; Goldinger, Luce, and Pisoni, 1989; Cluff and Luce, 1990; Luce, Pisoni, and Goldinger, 1990) assumes that word recognition reduces to a selection of a best match from a pool of activated word candidates and is thus similar in important respects to both Morton's logogen theory and Marslen-Wilson's cohort theory. However, the neighborhood activation model makes important assumptions about the role of competition for recognition among activated items. Central to the model is the concept of the **similarity neighborhood** (Landauer and Streeter, 1973; Coltheart et al., 1977; Luce, 1986; Andrews, 1989), a collection of words resident in the mental lexicon that are phonetically similar to each other and to any given stimulus word presented for recognition. Similarity neighborhoods are characterized by two main structural characteristics: (1) **neighborhood density,** the number of words in the neighborhood and their degrees of confusability with the stimulus word, and (2) **neighborhood frequency,** the frequencies of the words in the neighborhood relative to the frequency of the stimulus word (see Fravenfeld et al., 1993).

In experiments on perceptual identification of words presented in noise, auditory lexical decision, and auditory word naming, Luce (1986) observed that these structural characteristics of similarity neighborhoods strongly affected the speed and accuracy of word recognition. Words from sparse neighborhoods were recognized faster and more accurately than words from dense neighborhoods, and words from low-frequency neighborhoods were recognized faster and more accurately than words from high-frequency neighborhoods. Indeed, neighborhood characteristics were more reliable predictors of word recognition than word frequency itself; in the auditory word-naming experiment,

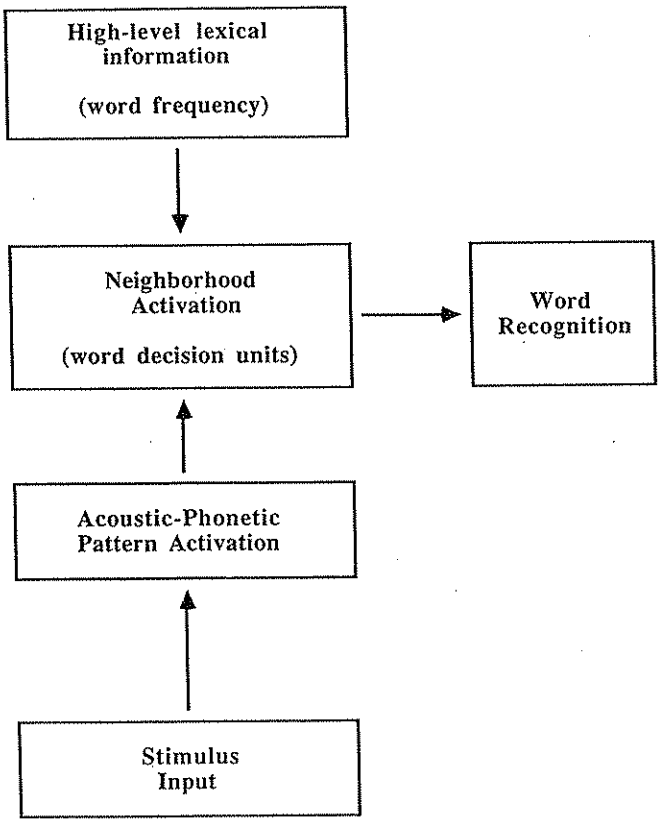


FIGURE 8-18 The neighborhood activation model of spoken word recognition. The model emphasizes the importance of similarity neighborhoods in isolating a single word candidate from the lexicon. Word frequency is assumed to bias the word decision units as they monitor activation patterns in the lexicon. (Adapted from Luce, P. A. [1986]. Neighborhoods of words in the mental lexicon. Unpublished doctoral dissertation. Indiana University.)

robust effects of neighborhood density were observed, but no effects of word frequency were evident (see Balota and Chumbley [1984, 1990] regarding the lability of word frequency effects).

In the neighborhood activation model, word recognition is much like that of both logogen theory and cohort theory, but with two basic modifications. The model (Fig. 8-18) assumes that upon stimulus input, a set of acoustic-phonetic patterns are activated in memory. The activation levels of these patterns are assumed to be a direct function of their phonetic similarity to the stimulus input. The activated phonetic patterns in turn activate a system of *word decision units,* conceptually similar to logogens. The word decision units are activated directly and autonomously from the bottom-up information provided by the signal, as in cohort theory. Once the word decision units are activated, they monitor a number of sources of information, especially the fluctuating activation levels of the acoustic-phonetic patterns. However, unlike processing in the system of logogens or the cohort, the word decision units also monitor the overall level of activity in the decision system itself, in a manner similar to the processing units in the TRACE model. Finally, the decision units are sensitive to higher-level lexical information, including word frequency. This information biases the decisions of the units by differentially weighting the activity levels of the words to which they respond. Word recognition occurs when the system of decision units selects a best match from the activated neighborhood, at which time all information about the word is made available to working memory.

The neighborhood activation model places much of the burden of spoken word recognition on discrimination and selection among similar acoustic-phonetic patterns corresponding to words. Accordingly, it can account for effects of similarity between stimulus words and their neighbors in the lexicon. In both logogen and cohort theory, it is explicitly assumed that word recognition is independent of the number of activated candidates. Therefore, these models do

not explain neighborhood density or neighborhood frequency effects. In addition, the model accounts for word frequency by assuming that frequency information biases the decisions of the word decision units. By assuming that frequency works in the late decision stage rather than in the early activation of the word units, the neighborhood activation model accounts for the common observation that word frequency effects vary across experimental tasks. Since different tasks introduce different decisional requirements, the neighborhood activation model predicts different effects of word frequency. Logogen theory, cohort theory, and Forster's search model propose that frequency is an integral and early contributor to word candidate activation or search order, and so these models are not well suited to account for experiments in which word frequency effects are attenuated or absent (e.g., Balota and Chumbley, 1984; Luce, 1986).

Despite the advantages of the neighborhood activation model over other models of word recognition, it does introduce several methodological difficulties. First, the concept of phonetic similarity among words in memory is difficult to quantify for empirical tests, and crude estimation methods, such as the *N* metric (Coltheart et al., 1977), are most commonly employed. Also, the concept of similarity depends on assumptions of representation. Similarity may be defined with respect to the speaker's phonetic repertoire or with respect to the listener's idealized phonetic representations, which are unavailable for inspection. Despite these difficulties, however, the concept of similarity is easily handled in theory, and the empiric effects of similarity neighborhoods are robust despite estimation. A second shortcoming of the model is the treatment of the temporal characteristics of word recognition. Unlike cohort theory, which explicitly accounts for the time course of word recognition, the neighborhood activation model offers no account of the recognition of multisyllabic words (although see Cluff and Luce, 1990).

Finally, we should mention another model to which the neighborhood activation model bears resemblance—the activation-verification model (Becker, 1976, 1979, 1980; Becker and Killion, 1977; Paap et al., 1982). In the activation verification framework, presentation of a stimulus word activates a pool of similar candidates selected by coarse sensory analysis. These candidates are subjected to *verification* in which each candidate word is compared with the stimulus until a best match is established. The verification process is similar to the search procedure in Forster's search model; candidates are submitted for verification in descending order of word frequency. By incorporating the concept of the verification set, which is much like a similarity neighborhood, the activation-verification model can account for the effects of set size and similarity among neighbors. However, the model's assumption of the frequency-ordered verification process reduces its flexibility in predictions of word frequency effects across tasks (Dobbs, Friedman, and Lloyd, 1985).

TRACE and other Connectionist Models

The TRACE model of speech perception* (Elman and McClelland, 1986; McClelland and Elman, 1986; Elman, 1989) is a nearly completely interactive system. Coming out of the growing connectionist movement and based on the interactive-activation model of visual word recognition (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982), TRACE advocates multiple levels of representation and rich feedforward and feedback connections between processing units. In addition, TRACE incorporates processes for both activation and inhibition of units in the network, as in the interactive-activation framework.

Figure 8-19 displays a section of a TRACE network. The functional units are simple, highly interconnected processing units called *nodes*. When information passes upward through the levels, nodes that collect sufficient confirmatory evidence to pass a threshold will fire and send activation along weighted links to their related nodes. In this manner, information consistent with the expectations of the early feature detectors is proliferated upward in the network to encourage recognition of the features' associated phonemes, and then recognition of the phonemes encourages recognition of the phonemes' associated words.

*Although most contemporary "models of speech perception" are clearly concerned with speech perception and most "models of word recognition" are concerned with word recognition, several models address both. LAFS is one of these. TRACE also accounts for phenomena from both the speech perception and word recognition literature. We recognize the contribution of TRACE and similar connectionist models to theories of speech perception. Our decision to discuss it in this section is simply an acknowledgment of its importance as a theory of word recognition.

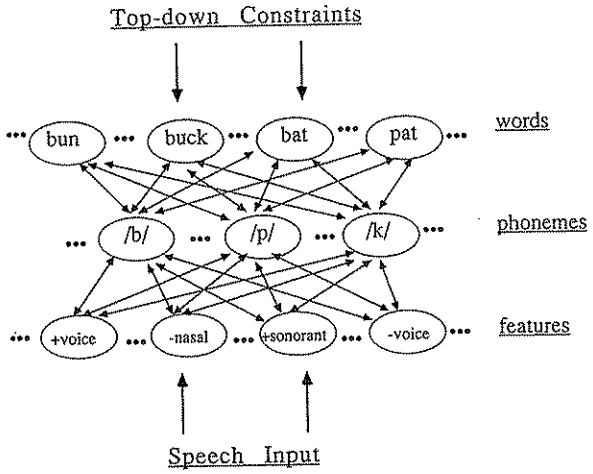


FIGURE 8-19 A section of a connectionist network with TRACE architecture. The network contains nodes corresponding to phonetic features, phonemes, and words. Information is provided to the network via the speech signal and top-down knowledge. Excitatory and inhibitory links among the nodes control perception and learning in the network.

A key property of TRACE is the organization of excitatory and inhibitory links between nodes and levels. All connections from one level to another are excitatory, i.e., activation of a node on one level will increase the activity of all connected nodes on adjacent levels. As an example, if the feature detector node for voicing encounters voicing cues consistent with /k/, then the node for /k/ will be activated and in turn will activate all words in the lexicon that contain /k/. Within levels, however, all nodes are connected by inhibitory links, so the model must quickly resolve ambiguity in the signal. For example, if the features for /k/ and /g/ are encountered simultaneously, the nodes corresponding to the features and phonemes for both possibilities not only become activated but also inhibit their nearest competitors. Computationally, the end result of this process is a winner-take-all form of perceptual decision (Elman and McClelland, 1986), meaning that the node that receives the most positive activation also receives the most veto power over its competitors. A final important property of the model is its use of perceptual feedback. That is, not only do activation and the flow of information in the model proceed from the early feature detection system to the lexicon, but the expectations at the lexical and phonemic levels can bias perception on the levels below (Gamong, 1980; Nygaard, 1993).

The interactive nature of TRACE offers much to theories of speech perception and word recognition. McClelland and Elman (1986) list nearly a dozen well-known phenomena that the model can simulate, ranging from categorical perception to trading relations, as well as findings from the word recognition literature, such as earliness of word recognition. As regards speech perception, TRACE does not treat coarticulatory speech as noise imposed on an idealized string of phonemes. Instead, Elman and McClelland (1984) call contextual variability *lawful variability,* a rich source of information in TRACE. (The authors say, "You can tell a phoneme by the company it keeps.") Although the model assumes segmental representations in speech, no explicit segmentation is imposed. Instead, phones and allophones are simply assumed in the model's architecture, so segmentation falls out naturally. As regards word recognition, the inhibitive links among nodes at the lexical level allow TRACE to account for neighborhood effects. In brief, by virtue of its simple assumptions of interacting units, TRACE demonstrates many of the attributes of theories of speech perception and word recognition in an integrated system without postulating or proliferating restrictive rules or specialized mechanisms.

However, like all models, TRACE has its problems. Many of them relate to the simplifying assumptions about speech input. Others are inherent to the model. Among the most serious problems are these: (1) It has no mechanism for predicting word frequency effects (although it is easy to imagine how a set of lexical level biases could be instantiated). (2) It has no obvious way of identifying a nonword. Judging lexical status is one of the most important abilities of word recognition (Forster, 1979) and should be in-

cluded in any model. TRACE could set criterial confidence values for outputs so that unfamiliar words would be judged as nonwords, but this would confound distinctions between degraded inputs and nonwords (a discrimination that human listeners make easily). (3) Perhaps the most important problem with TRACE arises from one of its most attractive features. It acknowledges and even exploits variability and coarticulation in its perceptual decisions, but it does not address other sources of variability common in natural language, such as talker idiosyncracies, changing speaking rates, stress assignments, and others. Even more troubling is the fact that TRACE demands a certain degree of invariance in its variability. It acknowledges that the cues for phonemes are not localized in specific segments, but at the same time it does require that all cues occur in a predetermined time window. While the problems of temporally distributed cues in speech are not easily resolved in the original TRACE model, it is hoped that the new breeds of *recurrent networks* (Jordan, 1986) may alleviate some of the difficulties of working with time windows in speech (for example, see Elman [1990, 1993]).

SUMMARY

This chapter identifies and elucidates several of the principal issues in research and theory on speech perception and auditory word recognition. Some are long-standing concerns in the field. Despite their long history as empirical and theoretical issues, problems such as the lack of acoustic-phonetic invariance and segmentation, the problem of perceptual normalization, and the specialization of speech perception remain vital and controversial areas of research today. And although innovative approaches to these issues have developed both in research and in theory, the fundamental complexity of speech perception continues to puzzle researchers. No comprehensive solutions to these problems are in the immediate future, but the trends are encouraging.

A particularly encouraging trend is the growing emphasis on considering speech perception and spoken word recognition as interacting stages of a unitary process. Most research on speech perception over the past 40 years has been concerned with the perception of isolated phonetic contrasts or phonemes in brief, meaningless syllables. Although the modularist approaches maintain that research and theory can proceed in a vacuum, the major current trend appears to be toward interactionism, bridging the gap that has traditionally separated the study of these different stages of speech comprehension. Already we have observed the development of several connectionist approaches to language processing, emphasizing the value of interaction between levels of representation. Perhaps the major insight of this approach has been the value of allowing models of speech perception and word recognition to constrain each other. As noted by Pisoni and Luce (1987), theorizing about one stage of language processing without regard for related stages may lead to theories that work well in artificial isolation, but if theories about one stage of processing are incompatible with our understanding of another stage, it is not clear what we have learned.

In short, we believe that the growing interest in the perception of spoken language, going beyond the level of the phoneme to the level of the word, reflects a healthful trend toward more comprehensive accounts of language perception. Of course, much research remains to be done on almost every level of spoken language understanding. The problems of speech perception and spoken word recognition, along with all aspects of language perception, promise to provide interesting and challenging research opportunities for at least another 40 years.

Acknowledgment

This work was supported by NIH Research Grant DC-00111-14 to Indiana University, Bloomington, Indiana.

Review Questions

1. What is the animorphism paradox and why has it remained a central concern in speech research?
2. Explain the problem of the information transfer rate in speech communication.
3. Discuss several possible conclusions that one might draw from studies comparing speech and nonspeech perception.
4. Why has duplex perception been cited as support for a theory of a phonetic module?

5. Does the evidence suggest that nonhuman animals perceive speech sounds in a manner similar to that of humans?

6. Explain the problem of perceptual constancy. Relate it to talker variability and changes in speaking rate.

7. What is the central claim of the motor theory of speech perception? What are the claims of information-processing models?

8. Describe three phenomena that all models of word recognition should be equipped to explain.

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Chicago, IL: Aldine.

Abramson, A. S., & Lisker, L. (1967). Discriminability along the voicing continuum: Cross language tests. In B. Hála, M. Romportl, & P. Janota (Eds.), *Proceedings of the Sixth International Congress of Phonetic Sciences.* Prague: Academia (pp. 569-573).

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 802-814.

Aslin, R. N. (1985). Effects of experience on sensory and perceptual development: Implications for infant cognition. In Mehler, J., & Fox, R. (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion.* Hillsdale, NJ: Erlbaum (pp. 157-183).

Aslin, R. N., & Pisoni, D. B. (1980). Some developmental processes in speech perception. In Yeni-Komshian, G., Kavanagh, J. F., & Ferguson, C. A. (Eds.), *Child phonology: Perception and production.* New York: Academic Press (pp. 67-96).

Bagley, W. C. (1900-1901). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology, 12,* 80-130.

Bailey, P. J., Summerfield, Q., & Dorman, M. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories status report on speech research SR-51/52,* 1-25.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 340-357.

Balota, D. A., & Chumbley, J. I. (1990). Where are the effects of frequency in visual word recognition tasks? Right where we said they were! Comment on Monsell, Doyle, and Haggard. *Journal of Experimental Psychology: General, 119,* 231-237.

Barclay, J. R. (1972). Noncategorical perception of a voiced stop: A replication. *Perception & Psychophysics, 11,* 269-273.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics, 44,* 395-408.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 556-566.

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 252-259.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition, 8,* 493-512.

Becker, C. A., & Killion, T. H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 389-401.

Bentin, S., & Mann, V. (1990). Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes. *Journal of the Acoustical Society of America, 88,* 64-74.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In Nusbaum H. & Goodman J. (Eds.), *The development of speech perception.* Cambridge, MA: MIT Press (pp. 167-224).

Best, C. T., MacRoberts, G. W., & Sithole, N. M. (1988). Examination of the perceptual reorganization for speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 245-260.

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics, 29,* 191-211.

Bever, T. G., Lackner, J., & Kirk, R. (1969). The underlying structures of sentences are the primary units of immediate speech processing. *Perception & Psychophysics, 5,* 225-231.

Bradley, D. C., & Forster, K. I. (1987). A reader's view of listening. *Cognition, 25,* 103-134.

Bregman, A. S. (1978). The formation of auditory streams. In Requin, J. (Ed.), *Attention and performance VII.* Hillsdale, NJ: Erlbaum (pp. 63-75).

Bregman, A.S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Broadbent, D. E. (1965). Information processing in the nervous system. *Science, 150,* 457-462.

Carr, P. B., & Trill, D. (1964). Long-term larynx-

excitation spectra. *Journal of the Acoustical Society of America, 36,* 2033-2040.

Carrell, T. D. (1984). *Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification.* Unpublished doctoral dissertation. Indiana University.

Carrell, T. D., Smith, L. B., & Pisoni, D. B. (1981). Some perceptual dependencies in speeded classification of vowel color and pitch. *Perception & Psychophysics, 29,* 1-10.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural language. In Luce, R. D., Bush, R., & Galanter, E. (Eds.), *Handbook of mathematical psychology,* Vol. 2. New York: Wiley (pp. 269-321).

Church, B., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 521-533.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 551-563.

Cohen, A., & Nooteboom, S. G. (Eds.) (1975). *Structure and process in speech perception.* Heidelberg: Springer-Verlag.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In Cole, R. A. (Ed.), *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum (pp. 133-163).

Cole, R. A., & Rudnicky, A. I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874-1946. *Psychological Review, 90,* 94-101.

Cole, R. A., & Scott, B. (1974a). The phantom in the phoneme: Invariant cues for stop consonants. *Perception and Psychophysics, 15,* 101-107.

Cole, R. A., & Scott, B. (1974b). Toward a theory of speech perception. *Psychological Review, 81,* 348-374.

Collier, R., & t'Hart, J. (1975). The role of intonation in speech perception. In Cohen, A., & Nooteboom, S.G. (Eds.), *Structure and process in speech perception.* Heidelberg: Springer-Verlag (pp. 107-123).

Coltheart, M., Davelaar, E., and Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In Dornic, S. (Ed.), *Attention and performance VI.* Hillsdale, NJ: Erlbaum (pp. 535-555).

Connine, C., Blasko, D., & Titone, D. (1993). Do the beginnings of words have a special status in auditory word recognition? *Journal of Memory and Language, 32,* 193-210.

Cooper, F. S. (1972). How is language conveyed by speech? In Kavanagh, J. F., & Mattingly, I. G. (Eds.), *Language by ear and by eye.* Cambridge, MA: MIT Press.

Cooper, W. E. (1976). Syntactic control of timing in speech production: A study of complement clauses. *Journal of Phonetics, 4,* 151-171.

Cooper, W. E., & Sorenson, J. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America, 62,* 683-692.

Creelman, C. D. (1957). The case of the unknown talker. *Journal of the Acoustical Society of America, 29,* 655.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics, 20,* 55-60.

Cutler, A. (1989). Auditory lexical access: Where do we start? In Marslen-Wilson, W. D. (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press (pp. 342-356).

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31,* 218-236.

Cutler, A., & Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception & Psychophysics, 29,* 217-224.

Cutler, A., & Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition, 7,* 49-59.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech, 20,* 1-10.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology, 24,* 381-410.

Cutting, J. E. (1978). There may be nothing peculiar to perceiving in a speech mode. In Requin, J. (Ed.), *Attention and performance VII.* Hillsdale, NJ: Erlbaum (pp. 229-244).

Cutting, J. E. (1987). Perception and information. *Annual Review of Psychology, 38,* 61-90.

Cutting, J. E., & Pisoni, D. B. (1978). An information-processing approach to speech perception. In Kavanagh, J. F., & Strange, W. (Eds.), *Speech and language in the laboratory, school, and clinic.* Cambridge, MA: MIT Press (pp. 38-72).

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. In Cohen, A., & Nooteboom, S. G. (Eds.), *Structure and process in speech perception.* Heidelberg: Springer-Verlag (pp. 178-194).

Darwin, C. J. (1976). The perception of speech. In Carterette, E. C. & Friedman, M. P. (Eds.), *Handbook of perception.* New York: Academic Press (pp. 175-216).

Day, R. S. (1968). Fusion in dichotic listening. Unpublished doctoral dissertation. Stanford University.

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics, 51,* 355-362.

Delattre, P. C., Liberman, A. M., & Cooper, F. S.

(1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769-773.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. H. (1952). An experimental study of the acoustic determinants of vowel color: Observations of one- and two-formant vowels synthesized from spectrographic patterns. *Word, 8,* 195-210.

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review, 93,* 283-321.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27,* 761-764.

Derr, M. A., & Massaro, D. W. (1980). The contribution of vowel duration, F_O contour, and frication duration as cues to the /juz/ - /jus/ distinction. *Perception & Psychophysics, 27,* 51-59.

Diehl, R. L. (1986). Coproduction and direct perception of phonetic segments: A critique. *Journal of Phonetics, 14,* 61-66.

Diehl, R. L., & Kleunder, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1,* 121-144.

Dobbs, A. R., Friedman, A., & Lloyd, J. (1985). Frequency effects in lexical decisions: A test of the verification model. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 81-92.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception and Psychophysics, 32,* 562-570.

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception and Psychophysics, 18,* 341-347.

Eimas, P.D., & Corbit, J. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4,* 99-109.

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science, 209,* 1140-1141.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303-306.

Elman, J. L. (1989). Connectionist approaches to acoustic/phonetic processing. In Marslen-Wilson, W. D. (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press (pp. 227-260).

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48,* 71-99.

Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In Lass, N. J. (Ed.), *Speech and Language: Advances in basic research and practice,* Vol. 102. New York: Academic Press (pp. 337-374).

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech waveform. In Perkell, J. S., & Klatt, D. H. (Eds.), *Invariance and variability in speech processing.* Hillsdale, NJ: Erlbaum (pp. 360-385).

Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos, 5,* 3-17.

Fant, G. (1973). *Speech sounds and features.* Cambridge, MA: MIT Press.

Fitch, H. L., Hawles, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics, 27,* 343-350.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fodor, J. A. (1985). Précis of *The modularity of mind. The Behavioral and Brain Sciences, 8,* 1-42.

Forster, K. I. (1976). Accessing the mental lexicon. In Wales, R. J., & Walker, E.C.T. (Eds.), *New approaches to language mechanisms.* Amsterdam: North Holland (pp. 257-287).

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In Cooper, W. E., & Walker, E.C.T. (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum (pp. 27-86).

Forster, K. I. (1989). Basic issues in lexical processing. In Marslen-Wilson, W. D. (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press (pp. 75-107).

Forster, K. I. (1990). Lexical processing. In Osherson, D. N., & Lasnik, H. (Eds.), *An invitation to cognitive science, Vol.* 1. Cambridge, MA: MIT Press (pp. 95-131).

Forster, K. I, & Bednall, E.S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition, 4,* 53-61.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3-28.

Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America, 88,* 1236-1249.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics, 55,* 597-610.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming of doors. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 742-754.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In Mattingly, I. G., & Studdert-Kennedy, M. (Eds.), *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Erlbaum (pp. 33-59).

Frauenfelder, U., Baayen, R., Hellwig, F., &

Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language, 32,* 781-804.

Fujisaki, H., & Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual report of the Engineering Research Institute,* Vol. 28. Tokyo: University of Tokyo (pp. 67-73).

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual report of the Engineering Research Institute,* Vol. 29. Tokyo: University of Tokyo (pp. 207-214).

Fujisaki, H., & Kawashima, T. (1971). A model of the mechanisms for speech perception: Quantitative analysis of categorical effects in discrimination. *Annual report of the Engineering Research Institute,* Vol. 30. Tokyo: University of Tokyo (pp. 59-68).

Gaitenby, J. H. (1965). The elastic word. *Haskins Laboratories status report on speech research, SR-2,* 3.1-3.12.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110-125.

Garner, W. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

Geiselman, R. E., & Bellezza, F. S. (1976). Long-term memory for speaker's voice and source location. *Memory & Cognition, 4,* 483-489.

Geiselman, R. E., & Bellezza, F. S. (1977). Incidental retention of speaker's voice. *Memory & Cognition, 5,* 658-665.

Geiselman, R. E., & Crawley, J. M. (1983). Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning and Verbal Behavior, 22,* 15-23.

Gerstman, L. H. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, au-16,* 78-80.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston, MA: Houghton-Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, MA: Houghton-Mifflin.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28,* 501-518.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). The nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152-162.

Goodman, J. C., & Huttenlocher, J. (1988). Do we know how people identify spoken words? *Journal of Memory and Language, 27,* 684-698.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologica, 9,* 317-323.

Green, K. P., Stevens, E. B., & Kuhl, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics, 55,* 249-260.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28,* 267-283.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics, 38,* 299-310.

Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition, 25,* 135-155.

Grunke, M. E., & Pisoni, D. B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics, 31,* 210-218.

Hall, M.D., & Pastore, R. E. (1992). Musical duplex perception: Perception of figurally good chords with subliminal distinguishing tones. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 752-762.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Effect of third formant transitions on the perception of the voiced stop consonants. *Journal of the Acoustical Society of America, 30,* 122-126.

Hawles, T., & Jenkins, J. J. (1971). Problem of serial order in behavior is not resolved by context-sensitive associative memory models. *Psychological Review, 78,* 122-129.

Hockett, C. (1955). *Manual of phonology.* Publications in Anthropology and Linguistics No. 11. Bloomington, Indiana: Indiana University Press.

Hoffman, H. S. (1958). Study of some cues in the perception of voiced stop consonants. *Journal of the Acoustical Society of America, 30,* 1035-1041.

Holmberg, T. L., Morgan, K. A., & Kuhl, P. K. (1977). Speech perception in early infancy: Discrimination of fricative consonants. *Journal of the Acoustical Society of America, 62,* S76 (Abstract).

Huggins, A. W. F. (1972). On the perception of temporal phenomena in speech. In Requin, J. (Ed.), *Attention and performance VII.* Hillsdale, NJ: Erlbaum (pp. 279-297).

Isenberg, D., & Liberman, A. M. (1978). Speech and non-speech percepts from the same sound. *Journal of the Acoustical Society of America, 64,* S20 (Abstract).

Jenkins, J. J. (1989). Is this the way to Camelot? *Contemporary Psychology, 5,* 451-452.

Johnson, K. (1990). The role of perceived speaker identity in F_0 normalization of vowels. *Journal of the Acoustical Society of America, 88,* 642-654.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America, 94,* 701-714.

Joos, M. A. (1948). Acoustic phonetics. *Language, 24,* (Suppl. 2), 1-136.

Jordan, M. I. (1986). Serial order: A parallel distributed processing approach. *Report 8604,* Institute for Cognitive Science, University of California, San Diego.

Jusczyk, P. W. (1985). On characterizing the development of speech perception. In Mehler, J., & Fox, R. (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion.* Hillsdale, NJ: Erlbaum (pp. 199-229).

Jusczyk, P. W. (1986). A review of speech perception research. In Kaufman, L., Thomas, J., & Boff, K. (Eds.), *Handbook of perception and performance.* New York: Wiley (pp. 27-57).

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. W. (1992). Some consequences of stimulus variability on speech processing by 2-month old infants. *Cognition, 43,* 253-291.

Jusczyk, P. W., Pisoni, D. B., Reed, M. A., Fernald, A., & Myers, M. (1983). Infants' discrimination of the duration of rapid spectrum changes in nonspeech signals. *Science, 222,* 175-177.

Jusczyk, P. W., Pisoni, D. B., Walley, A. C., & Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *Journal of the Acoustical Society of America, 67,* 262-270.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustic Society of America, 72,* 379-389.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 322-335.

Kewley-Port, D., & Luce, P. A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & Psychophysics, 35,* 353-360.

Klatt, D. H. (1974). The duration of [S] in English words. *Journal of Speech and Hearing Research, 17,* 51-63.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics, 3,* 129-140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208-1221.

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

Klatt, D. H. (1989). Review of selected models of speech perception. In Marslen-Wilson, W. D. (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press (pp. 169-226).

Klatt, D. H., & Cooper, W. E. (1975). Perception of segment duration in sentence context. In Cohen, A., & Nooteboom, S. G. (Eds.), *Structure and process in speech perception.* Heidelberg: Springer-Verlag (pp. 69-80).

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237,* 1195-1197.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66,* 1668-1679.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138-1141.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190,* 69-72.

Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63,* 905-917.

Kuhl, P. K., & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics, 31,* 279-292.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56,* 485-502.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98-104.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12,* 119-131.

Lea, W. A. (1973). An approach to syntactic recognition without phonemics. *IEEE Transactions on Audio and Electroacoustics, au-21,* 249-258.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

Liberman, A. M. (1970a). The grammars of speech and language. *Cognitive Psychology, 1,* 301-323.

Liberman, A. M. (1970b). Some characteristics of perception in the speech mode. In Hamburg, D. A. (Ed.), *Perception and its disorders: Proceedings of ARNMD.* Baltimore: Williams & Wilkins (pp. 238-254).

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist, 37,* 148-167.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431-461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. H. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs, 68,* 1-13.

Liberman, A. M., Harris, K. S., Hoffman, H. A., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358-368.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1-36.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243,* 489-494.

Liberman, A. M., Mattingly, I. G., & Turvey, M. T. (1972). Language codes and memory codes. In Melton, A. W., & Martin, E. (Eds.), *Coding processes in human memory.* New York: Winston (pp. 307-334).

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. *American Anthropology, 74,* 287-307.

Lightfoot, N. (1989). Effects of talker familiarity on serial recall of spoken word lists. *Research on speech perception, progress report no. 15.* Indiana University.

Lindblom, B. E. F., & Svensson, S. G. (1973). Interaction between segmental and non-segmental factors in speech recognition. *IEEE Transactions on Audio and Electroacoustics, au-21,* 536-545.

Lisker, L., & Abramson, A. S. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384-422.

Lisker, L., & Abramson, A. S. (1967). The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences. Prague: Academia.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89,* 874-886.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon.* Unpublished doctoral dissertation. Indiana University.

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio. *Journal of the Acoustical Society of America, 78,* 1949-1957.

Luce, P. A., & Pisoni, D. B. (1987). Speech perception: New directions in research, theory, and applications. In Winitz, H. (Ed.), *Human communication and its disorders.* Norwood, NJ: Ablex (pp. 1-87).

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In Altmann, G. (Ed.), *Cognitive models of speech processing.* Cambridge, MA: MIT Press (pp. 122-147).

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253-257.

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2,* 369-390.

MacKain, K. S., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science, 219,* 1347-1349.

Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition, 14,* 211-235.

Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. (1981). *Integration of time-varying cues and the effects of phonetic context.* Unpublished manuscript, Haskins Laboratories, New Haven, CT.

Marcus, S. M. (1984). Recognizing speech: On mapping from sound to meaning. In Bouma, H., & Bowhuis, D. G. (Eds.), *Attention and performance X: Control of language processes.* Hillsdale, NJ: Erlbaum (pp. 151-164).

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189,* 226-228.

Marslen-Wilson, W. D. (1980a). Optimal efficiency in human speech processing. Unpublished manuscript.

Marslen-Wilson, W. D. (1980b). Speech understanding as a psychological process. In Simon, J. C. (Ed.), *Spoken language generation and understanding.* Dordrecht, Holland: Reidel (pp. 39-67).

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71-102.

Marslen-Wilson, W. D. (1989). Access and integration: Projecting sound onto meaning. In Marslen-Wilson, W. D. (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press (pp. 3-24).

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G.T.M. Altman (Ed.), *Cognitive Models of Speech Processing.* Cambridge, MA: MIT Press (pp. 148-172).

Marslen-Wilson, W. D. (1993). Issues of process and representation in lexical access. In Altman, G. T. M. (Ed.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting.* Cambridge, MA: MIT Press (pp. 187-210).

Marslen-Wilson, W. D., & Tyler, L. K. (1975). Processing structure of sentence perception. *Nature, 257,* 784-785.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-71.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29-63.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 676-684.

Massaro, D. W. (1972). Preperceptual images, pro-

cessing time, and perceptual units in auditory perception. *Psychological Review, 79,* 124-145.

Massaro, D. W. (1986). A new perspective and old problems. *Journal of Phonetics, 14,* 69-74.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry. *The Behavioral and Brain Sciences, 12,* 741-794.

Massaro, D. W., & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/ - /si/ distinction. *Journal of the Acoustical Society of America, 60,* 704-717.

Massaro, D. W., & Cohen, M. M. (1977). The contribution of voice-onset time and fundamental frequency as cues to the /zi/ – /si/ distinction. *Perception & Psychophysics, 22,* 373-382.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 753-771.

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science, 1,* 55-63.

Massaro, D. W., & Cohen, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General, 122,* 115-124.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In Lass, N.J. (Ed.), *Speech and language: Advances in basic research and practice,* Vol. 3. New York: Academic Press (pp. 129-165).

Mattingly, I. G., & Liberman, A. M. (1988). Specialized perceiving systems for speech and other biologically-significant sounds. In Edelman, G., Gall, W., & Cohen, W. (Eds.), *Auditory function: The neurobiological bases of hearing.* New York: Wiley (pp. 775-793).

McClelland, J. L. (1979). On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287-330.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review, 88,* 375-405.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 621-638.

Miller, G. A. (1962). Decision units in the perception of speech. *IRE transactions on information theory, IT-8,* 81-83.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelley, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America, 60,* 410-417.

Miller, J. L. (1980). The effect of speaking rate on segmental distinctions: Acoustic variation and perceptual compensation. In Eimas, P. D., & Miller, J. L. (Eds.), *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum (pp. 39-74).

Miller, J. L. (1987). Mandatory processing in speech perception. In Garfield, J. L. (Ed.), *Modularity in knowledge representation and natural-language understanding.* Cambridge, MA: MIT Press (pp. 309-322).

Miller, J. L. (1990). Speech perception. In Osherson, D. N., & Lasnik, H. (Eds.), *An invitation to cognitive science,* Vol. 1. Cambridge, MA: MIT Press (pp. 69-93).

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semi-vowel. *Perception & Psychophysics, 25,* 457-465.

Miller, J. L., & Wayland, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics, 54,* 205-210.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9,* 283-303.

Monsen, R. B., & Engebretson, A. M. (1977). Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America, 62,* 981-993.

Morse, P. A., & Snowdon, C. T. (1975). An investigation of categorical speech discrimination by rhesus monkeys. *Perception & Psychophysics, 17,* 9-16.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165-178.

Morton, J. (1979). Word recognition. In Morton, J., & Marshall, J. D. (Eds.), *Psycholinguistics 2: Structures and processes.* Cambridge, MA: MIT Press (pp. 109-156).

Morton, J. (1982). Disintegrating the lexicon: An information processing approach. In Mehler, J., Walker, E. C. T., & Garrett, M. (Eds.), *On mental representation.* Hillsdale, NJ: Erlbaum (pp. 89-109).

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379-390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365-378.

Murdock, B. B., Jr. (1962). The serial position effect in free recall. *Journal of Experimental Psychology, 64,* 482-488.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America, 63,* 234-245.

Neisser, U. (1967). *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In Chase, W. G. (Ed.), *Visual information processing.* New York: Academic Press (pp. 283-308).

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700-708.

Nooteboom, S. G., Brokx, J. P. L., & de Rooij, J. J. (1978). Contributions of prosody to speech perception. In Levelt, W. J. M., & Flores d'Arcais, G. B. (Eds.), *Studies in the perception of language.* New York: Wiley (pp. 75-107).

Nusbaum, H. C. (1984). Possible mechanisms of duplex perception: "Chirp" identification versus dichotic fusion. *Perception & Psychophysics, 35,* 94-101.

Nusbaum, H. C., & Morin, T. (1992). Paying attention to differences among talkers. In Tohkura, Y., Vatikiotis-Bateson, E. & Sagisaka, Y. (Eds.), *Speech perception, production, and linguistic structure.* Tokyo: IOS Press.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In Schwab, E. C., & Nusbaum, H. C. (Eds.), *Perception of speech and visual form: Theoretical issues, models, and research.* New York: Academic Press (pp. 113-157).

Nusbaum, H. C., Schwab, E. C., & Sawusch, J. R. (1983). The role of "chirp" identification in duplex perception. *Perception & Psychophysics, 33,* 323-332.

Nygaard, L. C. (1993). Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 268-286.

Nygaard, L. C., & Eimas, P. D. (1990). A new version of duplex perception: Evidence for phonetic and nonphonetic fusion. *Journal of the Acoustical Society of America, 88,* 75-86.

Nygaard, L. C., Sommers, M.S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42-46.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172-191.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America, 54,* 1235-1247.

Oller, D. K., Eilers, R. E., & Ozdamar, O. (1990). A psychoacoustic model of the ba/wa boundary shift. *Journal of the Acoustical Society of America, 87,* S38 (Abstract).

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation-verification model for letter and word recognition: The word-superiority effect. *Psychological Review, 89,* 573-594.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309-328.

Pastore, R. E. (1981). Possible psychoacoustic factors in speech perception. In Eimas, P. D., & Miller, J. L. (Eds.), *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum (pp. 165-205).

Pastore, R. E., Schmeckler, M. A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception & Psychophysics, 33,* 469-474.

Peters, R. W. (1955a). The effect of length of exposure to speaker's voice upon listener reception. *Joint Project Report No. 44.* U.S. Naval School of Aviation Medicine, Pensacola, FL (pp. 1-8).

Peters, R. W. (1955b). The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise. *Joint Project Report No. 56.* U.S. Naval School of Aviation Medicine, Pensacola, FL (pp. 1-9).

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Peterson, L. J., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193-198.

Pisoni, D. B. (1971). On the nature of categorical perception of speech sounds. *Supplement to status report on speech research, SR-27.* New Haven, CT: Haskins Laboratories.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253-260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition, 3,* 7-18.

Pisoni, D. B. (1977). Identification and discrimination

of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America, 61,* 1352-1361.

Pisoni, D. B. (1978). Speech perception. In Estes, W. K. (Ed.), *Handbook of learning and cognitive processes,* Vol. 6. Hillsdale, NJ: Erlbaum (pp. 167-233).

Pisoni, D. B. (1991). Modes of processing speech and nonspeech signals. In Mattingly, I. G., & Studdert-Kennedy, M. (Eds.), *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Erlbaum (pp. 225-238).

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 297-314.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics, 34,* 314-322.

Pisoni, D. B., Logan, J. S., & Lively, S. E. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In Nusbaum, H. C., & Goodman, J. C. (Eds.), *Development of speech perception: The transition from recognizing speech sounds to spoken words.* Cambridge, MA: MIT Press (pp. 121-166).

Pisoni, D. B., & Luce, P. A. (1986). Speech perception: Research, theory, and the principal issues. In Schwab, E. C., & Nusbaum, H. C. (Eds.), *Perception of speech and visual form: Theoretical issues, models, and research.* New York: Academic Press (pp. 1-50).

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition, 25,* 21-52.

Pisoni, D. B., & Sawusch, J. R. (1975). Some stages of processing in speech perception. In Cohen, A., & Nooteboom, S. G. (Eds.), *Structure and process in speech perception.* Heidelberg: Springer-Verlag (pp. 16-34).

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 699-725.

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics, 52,* 37-52.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 421-435.

Port, R. F. (1977). *The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochu words.* Bloomington, Indiana: Indiana University Linguistics Club. Indiana University Press.

Porter, R. J., Jr. (1986). Speech messages, modulations, and motions. *Journal of Phonetics, 14,* 83-88.

Pruitt, J. S., Strange, W., Polka, L., & Aguilar, M. C. (1990). Effects of category knowledge and syllable truncation during auditory training on Americans' discrimination of Hindi retroflex-dental contrasts. *Journal of the Acoustical Society of America, 87,* S72 (Abstract).

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America, 55,* 678-680.

Remez, R. E. (1986) Realism, language, and another barrier. *Journal of Phonetics, 14,* 89-97.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101,* 129-156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947-950.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92,* 81-110.

Repp, B. H. (1983a). Categorical perception: Issues, methods, findings. In Lass, N. J. (Ed.), *Speech and language: Advances in basic research and practice,* Vol. 10. New York: Academic Press (pp. 243-335).

Repp, B. H. (1983b). Trading relations among acoustic cues in speech perception: Speech-specific but not special. *Haskins Laboratories status report on speech research, SR-76,* 129-132.

Repp, B. H. (1984). Against a role of "chirp" identification in duplex perception. *Perception & Psychophysics, 35,* 89-93.

Roberts, M., & Summerfield, Q. (1981). Audio-visual adaptation in speech perception. *Perception & Psychophysics, 30,* 309-314.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60-94.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89,* 43-50.

Samuel, A. G. (1986). The role of the lexicon in speech perception. In Schwab, E. C., & Nusbaum, H. C. (Eds.), *Perception of speech and visual form: Theoretical issues, models, and research.* New York: Academic Press (pp. 89-111).

Samuel, A. G., & Ressler, W. H. (1986). Attention within auditory word perception: Insights from the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 70-79.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior, 9,* 295-302.

Schacter, D. L., & Church, B. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 915-930.

Schouten, M. E. H. (1980). The case against a speech mode of perception. *Acta Psychologica, 44,* 71-98.

Schwab, E. C. (1981). *Auditory and phonetic processing for tone analogs of speech.* Unpublished doctoral dissertation. State University of New York at Buffalo.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing. In Bouma, H., & Bouwhis, D. G. (Eds.), *Attention and performance X: Control of language processes.* Hillsdale, NJ: Erlbaum (pp. 165-181).

Shankweiler, D. P., Strange, W., & Verbrugge, R. R. (1977). Speech and the problem of perceptual constancy. In Shaw, R., & Bransford, J. (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology.* Hillsdale, NJ: Erlbaum (pp. 315-345).

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In David, E. E., Jr., & Denes, P. B. (Eds.), *Human communication: A unified view.* New York: McGraw-Hill (pp. 51-66).

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358-1368.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In Eimas, P. D., & Miller, J. L. (Eds.), *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum (pp. 1-38).

Strange, W. (1972). *The effects of training on the perception of synthetic speech sounds: Voice onset time.* Unpublished doctoral dissertation. University of Minnesota.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics, 36,* 131-145.

Strange, W., & Jenkins, J. J. (1978). The role of linguistic experience in the perception of speech. In Pick, H. L., Jr., & Walk, R. D. (Eds.), *Perception and experience.* New York: Plenum (pp. 125-169).

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America, 60,* 213-221.

Studdert-Kennedy, M. (1974). The perception of speech. In Sebeok, T. A. (Ed.), *Current trends in linguistics,* Vol. 12. The Hague: Mouton (pp. 2349-2385).

Studdert-Kennedy, M. (1976). Speech perception. In Lass, N. J. (Ed.), *Contemporary issues in experimental phonetics.* New York: Academic Press (pp. 243-293).

Studdert-Kennedy, M. (1980). Speech perception. *Language and Speech, 23,* 45-66.

Studdert-Kennedy, M. (1982). On the dissociation of auditory and phonetic perception. In Carlson, R., & Granström, B. (Eds.), *The representation of speech in the peripheral auditory system.* Amsterdam: Elsevier (pp. 3-10).

Studdert-Kennedy, M. (1983). Perceiving phonetic events. *Haskins Laboratories: Status report on speech research, SR-74/75,* 53-69.

Studdert-Kennedy, M. (1986). Two cheers for direct realism. *Journal of Phonetics, 14,* 99-104.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77,* 234-249.

Studdert-Kennedy, M., & Shankweiler, D. P. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America, 48,* 579-594.

Summerfield, Q. (1975). Acoustic and phonetic components of the influence of voice changes and identification times for CVC syllables. *Report of speech research in progress, 2(4),* Queens University of Belfast (pp. 73-98).

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36,* 314-331.

Summerfield, Q. (1983). Audio-visual speech perception, lipreading and artificial stimulation. In Lutman, M. E., & Haggard, M. P. (Eds.), *Hearing Science and Hearing Disorders.* London: Academic Press.

Summerfield, Q., & Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of speech research in progress, 2(2),* Queens University of Belfast (pp. 12-23).

Sussman, H. M. (1989). Neural coding of relational invariance in speech: Human language analogs to the Barn Owl. *Psychological Review, 96,* 631-642.

Sussman, H. M. (1991). The representation of stop consonants in three-dimensional acoustic space. *Phonetica, 48,* 18-31.

Sussman, H. M., Hoemeke, K., & Ahmed, F. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *Journal of the Acoustical Society of America, 94,* 1256-1268.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America, 90,* 1309-1325.

Svensson, S. G. (1974). Prosody and grammar in speech perception. *Monographs from the Institute*

of Linguistics No. 2. Stockholm, Sweden: University of Stockholm, Institute of Linguistics.

Taft, M., & Hambly, G. (1986). Exploring the Cohort Model of word recognition. *Cognition, 22,* 259-282.

Tanenhaus, M. K., & Lucas, M. M. (1987). Context effects in lexical processing. *Cognition, 25,* 213-234.

Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America, 81,* 755-764.

Townsend, J. T. (1989). Winning "20 questions" with mathematical models. *The Behavioral and Brain Sciences, 12,* 775-776.

Tyler, L. K. (1984). The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics, 36,* 417-427.

Tyler, L. K., & Frauenfelder, U. H. (1987). The process of spoken word recognition: An introduction. *Cognition, 25,* 1-20.

Tyler, L. K., & Marslen-Wilson, W. D. (1982). Speech comprehension processes. In Mehler, J., Walker, E. C. T., & Garrett, M. (Eds.), *Perspectives on mental representation: Experimental and theoretical studies of cognitive processes and capacities.* Hillsdale, NJ: Erlbaum (pp. 169-184).

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters: Part I. Recognition of backward voices. *Journal of Phonetics, 13,* 19-38.

Van Lancker, D., Kreiman, J., & Wickens, T. D. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of rate-altered voices. *Journal of Phonetics, 13,* 39-52.

Verbrugge, R. R., & Rakerd, B. (1986). Evidence of talker-independent information for vowels. *Language and Speech, 29,* 39-57.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America, 60,* 198-212.

Vinegrad, M. D. (1972). A direct magnitude scaling method to investigate categorical versus continuous modes of speech perception. *Language and Speech, 15,* 114-121.

Walley, A. C., Pisoni, D. B., & Aslin, R. N. (1981). The role of early experience in the development of speech perception. In Aslin, R. N., Alberts, J., & Peterson, M. J. (Eds.), *The development of perception: Psychobiological perspectives.* New York: Academic Press (pp. 219-255).

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics, 41,* 262-275.

Warren, R. M. (1989). The use of mathematical models in perceptual theory. *The Behavioral and Brain Sciences, 12,* 776.

Waters, R. S., & Wilson, W. A., Jr. (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. *Perception & Psychophysics, 19,* 285-289.

Whalen, D., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science, 237,* 169-171.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76,* 1-15.

Wickelgren, W. A. (1976). Phonetic coding and serial order. In Carterette, E. C., & Friedman, M. P. (Eds.), *Handbook of perception,* Vol. 7. New York, NY: Academic Press (pp. 227-264).

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338-353.

Suggested Readings

Basic Research in Speech Perception

Borden, G. J., Harris, K. S. & Raphael, L. J. (1994). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech,* 3rd ed. Baltimore: Williams & Wilkins.

Cohen, A., & Nooteboom, S. G. (Eds.). (1975). *Structure and process in speech perception.* Heidelberg: Springer-Verlag.

Cole, R. A., & Rudnicky, A. I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874-1946. *Psychological Review,* 90, 94-101.

Eimas, P. D., & Miller, J. L. (Eds.). (1981). *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum.

Elman J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In Lass, N. J. (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10). New York: Academic Press, (pp. 337-374).

Gernsbacher, M. A. (Ed.), (1994). *Handbook of Psycholinguistics.* San Diego, CA: Academic Press.

Levelt, W. J. M., & Flores d'Arcais, G. B. (Eds.). (1978). *Studies in the perception of language.* New York: Wiley.

Pisoni, D. B. (1978). Speech perception. In Estes, W. K. (Ed.), *Handbook of learning and cognitive processes,* Vol. 6. Hillsdale, NJ: Erlbaum (pp. 167-233).

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In Lass, N. J. (Ed.), *Speech and language: Advances in basic research and practice.* Vol. 10. New York: Academic Press, (pp. 243-335).

Requin, J. (Ed.). (1978). *Attention and performance VII.* Hillsdale, NJ: Erlbaum.

Specialization of Speech Perception

Garfield, J. L. (Ed.). (1987). *Modularity in knowledge representation and natural-language understanding.* Cambridge, MA: MIT Press.

Mattingly, I. G., & Studdert-Kennedy, M. (Eds.). (1991). *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Erlbaum.

Variability in Speech Perception

Cole, R. A. (Ed.). (1980). *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum.

Perkell, J. S., & Klatt, D. H. (Eds.). (1986). *Invariance and variability in speech processing.* Hillsdale, NJ: Erlbaum.

Theories of Speech Perception

Altmann, G. (Ed.). (1990). *Cognitive models of speech processing.* Cambridge, MA: MIT Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Schwab, E. C., & Nusbaum, H. C. (Eds.). (1986). *Perception of speech and visual form: Theoretical issues, models, and research.* New York: Academic Press.

van den Broeke (Ed.). (1986). *Journal of Phonetics.* 14(1) (special issue).

Theories of Spoken Word Recognition

Bouma, H., & Bowhuis, D. G. (Eds.). (1984). *Attention and performance X: Control of language processes.* Hillsdale, NJ: Erlbaum.

Cooper, W. E., & Walker, E. C. T. (Eds.). (1979). *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum.

Frauenfelder, U. H., & Tyler, L. K. (Eds.). (1987). *Spoken Word Recognition.* Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (Ed.). (1989). *Lexical representation and process.* Cambridge, MA: MIT Press.