Annu. Rev. Psychol. 2004. 55:149–79
doi: 10.1146/annurev.psych.55.090902.142028

First published online as a Review in Advance on September 15, 2003

SPEECH PERCEPTION

Randy L. Diehl
Department of Psychology and Center for Perceptual Systems, University of Texas, Austin, Texas 78712-0187; email: diehl@psy.utexas.edu

Andrew J. Lotto
Boys Town National Research Hospital, Omaha, Nebraska 68131; email: lottoa@boystown.org

Lori L. Holt
Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; email: lholt@andrew.cmu.edu

Key Words auditory pattern recognition, categorical perception, phonetic context effects, perceptual learning, speech production

■ **Abstract** This chapter focuses on one of the first steps in comprehending spoken language: How do listeners extract the most fundamental linguistic elements—consonants and vowels, or the distinctive features which compose them—from the acoustic signal? We begin by describing three major theoretical perspectives on the perception of speech. Then we review several lines of research that are relevant to distinguishing these perspectives. The research topics surveyed include categorical perception, phonetic context effects, learning of speech and related nonspeech categories, and the relation between speech perception and production. Finally, we describe challenges facing each of the major theoretical perspectives on speech perception.

CONTENTS

INTRODUCTION

Over the past 50 years, researchers in speech perception have focused on the mapping between properties of the acoustic signal and linguistic elements such as phonemes and distinctive features. This mapping has turned out to be quite complex, and a complete explanation of how humans recognize consonants and vowels remains elusive. The search for an explanation has given rise to three main theoretical perspectives on speech perception that frame much of the empirical work. In this chapter, we briefly describe these perspectives and then review some of the research most relevant to evaluating them. We end by highlighting some of the main challenges facing each theoretical view.

MOTOR THEORY OF SPEECH PERCEPTION

Beginning in the early 1950s, Alvin Liberman, Franklin Cooper, Pierre Delattre, and other researchers at the Haskins Laboratories carried out a series of landmark studies on the perception of synthetic speech sounds (Delattre et al. 1951, 1952, 1955, 1964; Liberman 1957; Liberman et al. 1952, 1954, 1956). This work provided the foundation of what is known about acoustic cues for linguistic units such as phonemes and features and revealed that the mapping between speech signals and linguistic units is quite complex. In time, Liberman and his colleagues became convinced that perceived phonemes and features have a simpler (i.e., more nearly one-to-one) relationship to articulation than to acoustics, and this gave rise to the motor theory of speech perception.

The motor theory (MT) has undergone significant changes since its initial formulation (Liberman 1996), but every version has claimed that the objects of speech perception are articulatory events rather than acoustic or auditory events. More specifically, it was hypothesized that the articulatory events recovered by human listeners are neuromotor commands to the articulators (e.g., tongue, lips, and vocal folds)—also referred to as intended gestures—rather than more peripheral events such as actual articulatory movements or gestures (Liberman & Mattingly 1985, Liberman et al. 1967). This theoretical choice was guided by a belief that the objects of speech perception must be more-or-less invariant with respect to phonemes or feature sets and by a further belief that such a requirement was satisfied only by neuromotor commands. The process of speech production was characterized by Liberman et al. (1967) as a series of causal links between descriptive levels:

phonemes (or sets of distinctive features) → neuromotor commands → muscle contractions → vocal tract shapes → acoustic signals. Whereas phonemes (or feature sets) were assumed to stand approximately in one-to-one correspondence with neuromotor commands and with muscle contractions, the mapping between muscle contractions and vocal tract shapes was thought to be highly complex owing to the fact that adjacent vowels and consonants are coarticulated (i.e., produced with temporal and, to some extent, spatial overlap). Because the relation between vocal tract shapes and acoustic signals was assumed to be one-to-one, the complex mapping between phonemes and speech sounds was attributed mainly to the effects of coarticulation.

As an illustration of the complex mapping between phonemes and their acoustic realizations, Liberman et al. (1967) displayed spectrograms of synthetic two-formant patterns (shown in Figure 1) that are perceived by listeners as the syllables /di/ ("dee") and /du/ ("doo"). In these patterns, the steady-state formants correspond to the target values of the vowels /i/ and /u/, and the rapidly changing formant frequencies (formant transitions) at the onset of each syllable carry important information about the initial consonant. In particular, the rising first-formant (F1) transition of both syllables signals that the consonant is a voiced "stop" such as /b/, /d/, or /g/, whereas the rising second-formant (F2) transition of /di/ and the

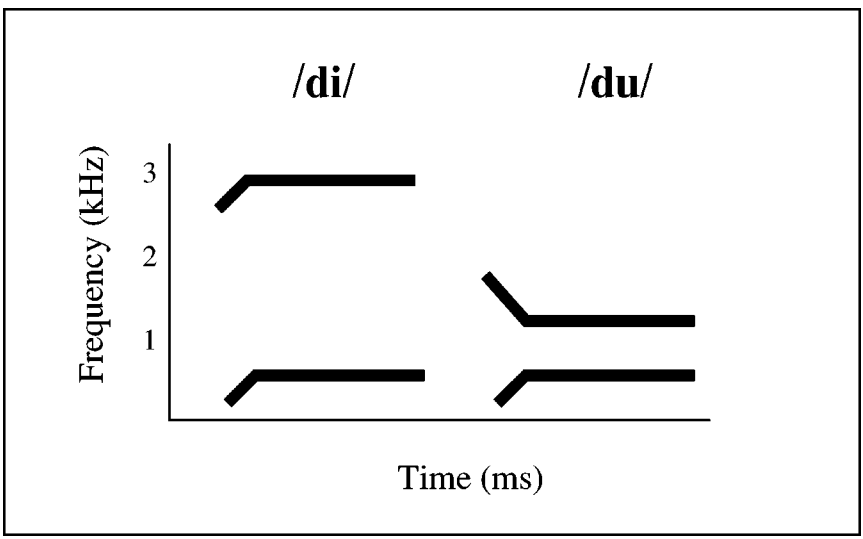

Figure 1 Formant patterns for simplified versions of /di/ and /du/. Note that the transition of the second formant (i.e., the one higher in frequency) differs dramatically for the two syllables. Nonetheless, the consonant in both cases is perceived as /d/. The first formant trajectory, which is equivalent in both syllables, is not informative about place of articulation and would be the same for /b/ and /g/ initial syllables. (Adapted from Delattre et al. 1952.)

falling F2 transition of /du/ provide critical information about place of articulation (i.e., that the consonant is /d/ rather than /b/ or /g/). That such different patterns of F2 transition could give rise to the same phonemic percept strongly suggested to the motor theorists that invariance must be sought at an articulatory rather than an acoustic level of description.

A second important claim of MT is that the human ability to perceive speech sounds cannot be ascribed to general mechanisms of audition and perceptual learning but instead depends on a specialized decoder or module that is speech-specific, unique to humans, and, in later versions of the theory (Liberman 1996, Liberman & Mattingly 1985), innately organized and part of the larger biological specialization for language. The speech decoder was hypothesized by Liberman et al. (1967) to operate by "somehow running the process [of speech production] backward" (p. 454). This claim was elaborated by Liberman & Mattingly (1985) as follows: "[T]he candidate signal descriptions are computed by an analogue of the production process—an internal, innately specified vocal-tract synthesizer. . .—that incorporates complete information about the anatomical and physiological characteristics of the vocal tract and also about the articulatory and acoustic consequences of linguistically significant gestures" (p. 26). Liberman and his colleagues argued that, among other theoretical advantages, MT is parsimonious inasmuch as the same mechanism is used for both speech production and speech perception.

DIRECT REALIST THEORY OF SPEECH PERCEPTION

Starting in the 1980s, an alternative to MT—referred to as the direct realist theory (DRT) of speech perception—was developed by Carol Fowler, also working at the Haskins Laboratories (Fowler 1981, 1984, 1986, 1989, 1994, 1996). Like MT, DRT claims that the objects of speech perception are articulatory rather than acoustic events. However, unlike MT, DRT asserts that the articulatory objects of perception are actual, phonetically structured, vocal tract movements, or gestures, and not events that are causally antecedent to these movements, such as neuromotor commands or intended gestures. DRT also contrasts sharply with MT in denying that specialized (i.e., speech-specific or human-specific) mechanisms play a role in speech perception. Instead, following the general theory of direct perception developed by James J. Gibson (1966, 1979), Fowler argues that speech perception can be broadly characterized in the same terms as, for example, visual perception of surface layout.

This view is elegantly summarized by Fowler (1996) in the following passage:

> Perceptual systems have a universal function. They constitute the sole means by which animals can know their niches. Moreover, they appear to serve this function in one way: They use structure in the media that has been lawfully caused by events in the environment as information for the events. Even though

> it is the structure in media (light for vision, skin for touch, air for hearing) that sense organs transduce, it is not the structure in those media that animals perceive. Rather, essentially for their survival, they perceive the components of their niche that caused the structure. (p. 1732)

Thus, according to DRT, a talker's gestures (e.g., the closing and opening of the lips during the production of /pa/) structure the acoustic signal, which then serves as the informational medium for the listener to recover the gestures. The term "direct" in direct realism is meant to imply that perception is not mediated by processes of inference or hypothesis testing; rather, the information in the acoustic signal is assumed to be rich enough to specify (i.e., determine uniquely) the gestures that structure the signal. To perceive the gestures, it is sufficient for the listener simply to detect the relevant information. The term "realism" is intended to mean that perceivers recover actual (physical) properties of their niche, including, in the case of speech perception, phonetic segments that are realized as sets of physical gestures. This realist perspective contrasts with a mentalistic view that phonetic segments are "internally generated, the creature of some kind of perceptual-cognitive process" (Hammarberg 1976, p. 355; see also Repp 1981).

Just as MT was motivated in part by a particular view of speech production (especially, the claim that coarticulation of consonants and vowels results in a complex mapping between phonemes and vocal tract shapes and hence between phonemes and acoustic signals), DRT was seen as broadly compatible with an alternative view of speech production (Fowler 1980, 1981; Fowler & Smith 1986). According to this view, the temporal overlap of vowels and consonants does not result in a physical merging or assimilation of gestures; instead, the vowel and consonant gestures are coproduced. That is, they remain, to a considerable extent, separate and independent events analogous to, say, a singer's vocal production and any temporally overlapping musical accompaniment. Because coproduced gestures are assumed to structure the acoustic signal in independent (albeit temporally overlapping) ways, the listener should, on the assumptions of DRT, have no difficulty recovering those gestures and their temporal sequencing. Fowler & Smith (1986) likened the perception of coproduced segments to a kind of "vector analysis" in which complex stimulus events are appropriately factored into separate components. For example, in the context of a following nasal consonant (e.g., /n/), a vowel tends to be nasalized, an effect known as anticipatory coarticulation. However, listeners appear not to hear the vowel as nasalized, instead attributing the nasalization to the following consonant alone (Krakow et al. 1988).

Because MT and DRT both claim that the objects of speech perception are gestures (intended in the case of MT, actual in the case of DRT), advocates of the two theories cite some of the same empirical findings as supporting evidence. Thus, for example, the fact that /di/ and /du/ (see Figure 1) are perceived as having the same initial consonant (despite their disparate F2 transitions) is explained both by Liberman et al. (1967) and by Fowler (1996) in terms of an assumed commonality of gestures in the two cases.

GENERAL AUDITORY AND LEARNING APPROACHES TO SPEECH PERCEPTION

In the mid 1970s, several new empirical findings posed a challenge to MT, the then-dominant account of human speech perception. Earlier work at Haskins Laboratories had found clear differences between perception of certain speech sounds and perception of nonspeech analogs of those speech stimuli (Liberman et al. 1961a,b; Mattingly et al. 1971). Because these results appeared to underscore the special nature of speech perception, they were interpreted as supporting MT (Liberman et al. 1967, 1972). However, Stevens & Klatt (1974), Miller et al. (1976), and Pisoni (1977) showed that in some instances perception of speech stimuli does parallel that of nonspeech stimuli provided they share critical temporal properties. The authors claimed that general auditory mechanisms were responsible for the observed similarities in perceptual performance. Even more surprising were demonstrations that nonhuman animals exhibit aspects of speech perceptual performance (Kuhl & Miller 1975, 1978) that were assumed by motor theorists to be unique to humans (Liberman et al. 1972). Some of these parallels between speech and nonspeech perception and between speech perception in humans and nonhumans are described later in more detail.

Stimulated by these and related findings, a number of speech investigators [e.g., Diehl 1987; Diehl & Kluender 1989a,b; Holt et al. 1998; Kingston & Diehl 1994, 1995; Kluender 1994; Kuhl 1986; Lotto 2000; Massaro & Oden 1980; Nearey 1990; Nearey & Hogan 1986; Ohala 1996; Pastore 1981; Sussman et al. 1998 (see Lane 1965 for an early critique of MT)] have explored alternatives to both MT and DRT, which will be referred to here as the general approach (GA). In contrast to MT, GA does not invoke special mechanisms or modules to explain speech perception. Rather, it assumes, as a working hypothesis, that speech sounds are perceived using the same mechanisms of audition and perceptual learning that have evolved in humans or human ancestors to handle other classes of environmental sounds. In contrast to MT and DRT, GA assumes that listeners' recovery of spoken messages from the acoustic signal (whether these messages are construed as distinctive features, phonemes, words, or some higher-level linguistic units) is neither equivalent to nor mediated by the perception of gestures.

Recall that the perceived equivalence of the consonant in /di/ and /du/ (despite varying acoustic patterns) was cited as supporting evidence for MT and DRT. A GA explanation for the perceptual equivalence would be based on the general ability of the perceiver to make use of multiple imperfect acoustic cues to categorize complex stimuli. In the same way that Brunswik (1956) proposed that object constancy in vision is the result of combining multiple attributes of varying ecological validity, the listener can maintain perceptual constancy in the face of structured variance in acoustics. For GA this constancy does not require the recovery of articulatory gestures or a special mode of perception. In support of this view, Kluender et al. (1987) demonstrated that birds could be trained to respond to natural /d/-initial

TABLE 1 Taxonomy of major theoretical approaches to speech perception

	Special mechanisms	**General mechanisms**
Gestural	Motor theory	Direct realism
Nongestural	Eclectic specializations	General approach

syllables versus /b/- and /g/-initial syllables. Despite the lack of any specialized mechanisms or experience producing speech, the birds were able to correctly respond to the same consonants in novel vowel contexts.

GA is labeled an approach rather than a theory because, as summarized in preceding paragraphs, it is quite abstract, defining itself mainly by its opposition to key claims of MT and DRT. At this level of abstraction, GA has too little content to be falsifiable. However, it does provide a general framework within which particular theoretical claims may be formulated and tested. Examples of such claims are reviewed in the following sections.

Table 1 presents a simplified taxonomy of the major theoretical approaches to speech perception based on the postulation of special versus general mechanisms and on the proposed objects of perception. The lower left quadrant corresponds to a possible claim that speech perception uses special mechanisms to recover a nongestural representation of linguistic elements. Although such a claim has not been developed into a coherent theory, there have been several proposals that specialized processes may work in concert with general perceptual mechanisms. For example, the ability of human infants to learn the phoneme categories of their native language has been attributed to specialized processes of categorization (Kuhl 1991, 1992, 1993) or to an attentional or learning bias for speech sounds (Jusczyk 1997). These are listed as "eclectic specializations" in the table.

CATEGORICAL PERCEPTION

An important early discovery at the Haskins Laboratories was an effect referred to as *categorical perception* (Liberman et al. 1957, 1961a,b). In a typical experiment, a series of synthetic consonant-vowel (CV) syllables varying in an acoustic parameter (e.g., the slope of the F2 transition) and ranging perceptually across several initial consonants (e.g., /bV/-/dV/-/gV/) were presented to listeners for phonemic labeling, or identification, and for discrimination of pairs of stimuli located near each other in the series. Two striking patterns were evident in the results. First, labeling functions exhibited abrupt boundaries between phoneme categories; second, discrimination accuracy was close to chance for stimulus pairs within a phoneme category but nearly perfect for stimulus pairs that straddled an identification boundary. These are the defining properties of categorical perception. They imply that in speech perception discriminability is closely related to the presence

or absence of functional (i.e., phonemic) differences between sounds. Because categorical discrimination functions were not found for certain nonspeech analogs of the speech stimuli (Liberman et al. 1961a,b), the motor theorists cited categorical perception as a hallmark of perception in the "speech mode" (Liberman et al. 1967).

This section focuses mainly on perception of voice distinctions in syllable-initial stop consonants, for example, /ba/ versus /pa/, /da/ versus /ta/, and /ga/ versus /ka/. Linguists commonly describe /b/, /d/, and /g/ as having the distinctive feature *+voice* and /p/, /t/, and /k/ as having the distinctive feature *–voice*, where the former but not the latter are produced with voicing, or vocal fold vibration. In a cross-language study of initial stop consonants, Lisker & Abramson (1964) identified a key phonetic correlate of voice contrasts, e.g., voice onset time (VOT), the interval between the release of articulatory occlusion (e.g., the opening of the lips) and the onset of voicing. Cross-linguistically, initial stops tend to occur in one of three ranges of VOT values: long negative VOTs (voicing onset leads the articulatory release by 50 ms or more); short positive VOTs (voicing lags behind the release by no more than 20 ms); and long positive VOTs (voicing onset lags behind the release by more than 25 ms). From these three phonetic types, languages usually choose two to implement their voice contrasts. For example, Spanish uses long negative VOTs to realize *+voice* stops and short positive VOTs to realize *−voice* stops; whereas English uses short positive VOTs to implement *+voice* stops and long positive VOTs to implement *–voice* stops.

Lisker & Abramson (1970, Abramson & Lisker 1970) next examined VOT perception among native speakers of English, Spanish, and Thai. All three language groups showed clear evidence of categorical perception. However, the locations of phoneme boundaries and the associated peaks in discriminability varied among the groups, reflecting differences in the way each language realizes voice distinctions. These results suggested that categorical perception of VOT arises from language experience, with listeners becoming more sensitive to phonetic differences that play a functional role in their language and/or less sensitive to differences that do not.

Complicating this language learning explanation were results of experiments performed with human infants. Eimas et al. (1971) reported that infants from an English-speaking environment discriminate differences in VOT for stimulus pairs that straddle the English /ba/-/pa/ boundary but show no evidence of discriminating equivalent VOT differences when the stimuli are from the same English category. Consistent with later versions of MT, the authors interpreted these results as evidence of an innate linguistic mode of perception in humans. Further supporting this view, Lasky et al. (1975) found that infants raised in a Spanish-speaking environment can discriminate differences in VOT if the stimuli straddled either the Spanish or the English voice boundary but show no evidence of discrimination otherwise. The discriminability of the English voice contrast by Spanish-learning infants suggested that language experience is not a necessary condition for categorical discrimination of VOT stimuli (see also Aslin et al. 1981).

Recall that categorical perception was claimed by motor theorists to be a hallmark of the speech mode of perception (Liberman et al. 1967). However, later studies (Miller et al. 1976, Pisoni 1977) yielded convincing evidence of categorical perception for several types of nonspeech analogs of VOT stimuli. In naturally produced stop-vowel syllables, negative VOTs correspond to a low-frequency "voice bar" that precedes the articulatory release, whereas positive VOTs are associated with a sharp attenuation of F1 before voicing onset (see Figure 2*A*.) VOT can thus be abstractly described as the relative onset time of low- versus high-frequency signal components. Pisoni (1977) created nonspeech analogs of VOT stimuli that consisted of a lower and a higher frequency tone with onsets varying from −50-ms tone onset time (TOT) to +50-ms TOT, where negative values indicate prior onset of the lower-frequency tone (see Figure 2*B*). After training in labeling selected stimuli, adult listeners displayed abrupt identification boundaries near −20 ms and +20 ms TOT values (analogous to the Spanish and English VOT boundaries) as well as peaks in discriminability near those boundaries. Similar bimodal discrimination performance for TOT stimuli was observed for infants (Jusczyk et al. 1980). The close parallel between categorical perception of VOT and TOT stimuli was attributed by Pisoni (1977) to a psychophysical threshold for detecting the temporal order of stimulus components (Hirsh 1959, Hirsh & Sherrick 1961). By this account, onset asynchronies of less than approximately 20 ms are judged as simultaneous, while those greater than that are judged as ordered in time. This yields three natural categories of onset asynchrony that correspond well to the three phonetic voice categories commonly used among the world's languages. Thus, in the case of VOT, languages appear to locate phoneme contrasts to exploit natural auditory boundaries, thereby enhancing distinctiveness, intelligibility, and perhaps learnability.

The finding that categorical perception is not unique to speech sounds weakened one of the empirical arguments for MT. An even more serious challenge was raised by results of experiments with nonhuman animals. Liberman et al. (1972, p. 324) had written:

> Presumably, they [animals] lack the special processor necessary to decode the speech signal. If so, the perception of speech must be different from ours. They should not hear categorically, for instance, and they should not hear the [di]-[du] patterns. . .as two segment syllables which have the first segment in common.

As for the /di/-/du/ example, it was pointed out earlier that Kluender et al. (1987) trained Japanese quail to respond to /d/-initial tokens, but to refrain from responding to /b/- and /g/-initial tokens, in various vowel contexts. With respect to the claim that categorical perception is a uniquely human ability, Kuhl and her colleagues (Kuhl 1981; Kuhl & Miller 1975, 1978; Kuhl & Padden 1982) presented strong evidence of categorical perception of human speech sounds by chinchillas and macaque monkeys. For example, Kuhl & Miller (1978) trained chinchillas to respond differently to two endpoint stimuli of a synthetic VOT series (/da/, 0 ms

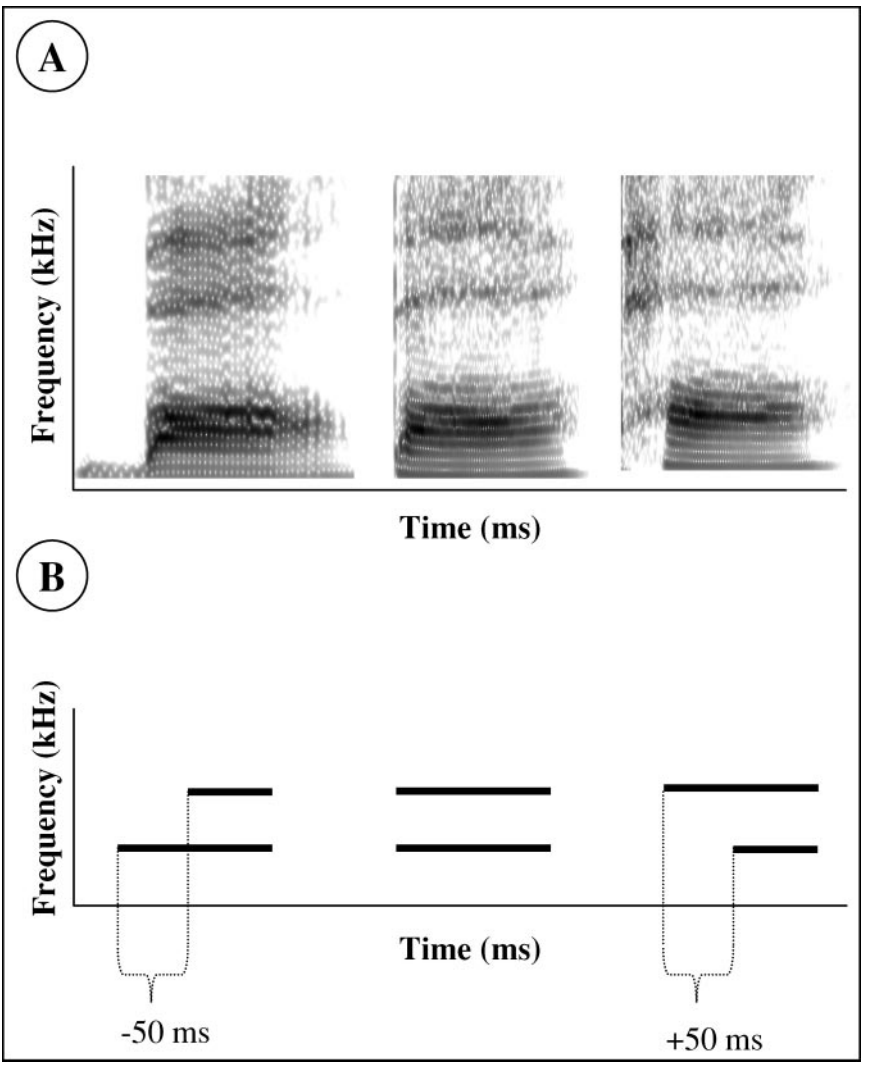

Figure 2 Spectrograms of (*A*) natural stop-vowel syllables with a voicing lead (negative VOT), short voicing lag (approximately 0 VOT), and long voicing lag (positive VOT) stop consonants. VOT is measured from the articulatory release (onset of formant transitions) to the onset of voicing, represented as low-frequency energy; and (*B*) corresponding TOT stimuli. These consist of two sine-wave segments that vary in relative onset time. (Adapted from Pisoni 1977.)

VOT; and /ta/, 80 ms VOT) and then tested the animals with stimuli at intermediate values. Their identification performance corresponded almost exactly to that of adult English-speaking listeners. Further generalization tests with labial (/ba/-/pa/) and velar (/ga/-/ka/) VOT stimuli, as well as tests of VOT discriminability (Kuhl 1981), also showed close agreement with the performance of English speakers.

Several neural correlates of categorical perception of VOT have been reported. Recording from a population of auditory nerve fibers in chinchilla, Sinex et al. (1991) found that cross-neuron variability in discharge patterns was reliably smaller for VOT stimuli near the English voice boundary than for stimuli located within either category. In a magnetoencephalographic study of the human primary auditory cortex, Simos et al. (1998) found that VOT stimulus pairs straddling the English voice boundary yielded differences in the location and amplitude of the peak response for native English-speaking listeners, whereas stimulus pairs drawn from the same category did not.

Although this discussion has focused on perception of VOT, other speech dimensions that are perceived categorically by humans appear to be perceived categorically by nonhumans as well. Kuhl & Padden (1983) reported that macaques show enhanced discriminability at phoneme boundaries for the feature place of articulation (/b/-/d/-/g/), and Dooling et al. (1995) found that budgerigars and zebra finches show enhanced discriminability at the English /r/-/l/ boundary. In both studies, discrimination performance of the animals closely matched that of human listeners.

The results of comparing speech and nonspeech perception and speech perception in humans and nonhumans strongly indicate that general auditory mechanisms (common to human adults and infants, other mammals, and even birds) contribute to the categorical perception of speech sounds. Evidently, however, language experience is also a significant factor in categorical perception. Lisker & Abramson (1970, Abramson & Lisker 1970) found cross-language differences in identification boundaries and discrimination peaks (see also Elman et al. 1977, Williams 1977). Although human infants exhibit heightened discriminability at both the Spanish and English voice boundaries, their language experience tends to maintain and perhaps enhance natural boundaries that coincide with phonemic boundaries and to downgrade natural boundaries that are linguistically nonfunctional (Werker & Tees 1984). In organizing their sound systems, languages exploit natural boundaries, but, within limits, they also modify them.

PHONETIC CONTEXT EFFECTS I: STIMULUS LENGTH EFFECT

The perceptual assessment of temporal cues for certain phoneme distinctions is known to depend on the duration of nearby regions of the acoustic signal (Diehl et al. 1980, Miller 1987, Summerfield 1981). For example, Miller & Liberman (1979) found that perception of the stop/glide distinction (e.g., /b/ versus /w/)

is influenced by the duration of the following vowel. Earlier research (Liberman et al. 1956) had demonstrated that the stop/glide distinction is reliably signaled by variation in the duration and slope of CV formant transitions, with shorter transitions specifying the stop category. The key result of Miller & Liberman (1979) was that a longer following vowel shifted the stop/glide boundary toward longer transition durations (i.e., more stops were perceived). Miller & Liberman explained this effect within the framework of MT: A longer vowel is evidence of a slower rate of articulation, and, to compensate perceptually, listeners accept a greater range of transition durations as compatible with the stop category.

Diehl & Walsh (1989) offered an alternative account of this stimulus length effect based on a putative general auditory factor referred to as durational contrast. According to this account, perceived length of an acoustic segment is affected contrastively by the duration of adjacent acoustic segments. Thus, a target segment will be judged as shorter next to a long segment than next to a short segment. Unlike the motor theoretic explanation of the stimulus length effect, the durational contrast hypothesis applies to both speech and nonspeech sounds.

To distinguish empirically between the two accounts, Diehl & Walsh compared labeling performance on several series of /ba/-/wa/ stimuli and on analogous nonspeech stimuli. The latter consisted of single sine-wave stimuli that mimicked the F1 trajectories and amplitude rise times of the speech stimuli. Listeners were asked to categorize the nonspeech items as having either abrupt or gradual onsets. When the stop/glide distinction was signaled by changes in transition duration, there was a reliable stimulus length effect like that observed by Miller & Liberman. A very similar effect was found for the corresponding nonspeech stimuli. Changes in rise time had only a small effect on identification of either the speech or nonspeech stimuli. Indeed, for seven of eight comparisons (main effects and interactions) involving the factor speech versus nonspeech, there were no significant differences in labeling performance. (The one exception was that when the stop/glide distinction was cued by rise time, there was a reliable stimulus length effect, but no such effect occurred for the corresponding nonspeech stimuli.) On balance, the parallel results between the speech and nonspeech conditions supported the durational contrast account of the stimulus length effect. (For a critique of this conclusion from the perspective of DRT, see Fowler 1990, 1991, and for a reply see Diehl et al. 1991.)

PHONETIC CONTEXT EFFECTS II: COMPENSATION FOR COARTICULATION

As described earlier, phonemes are coarticulated in running speech. Consider the production of the CV syllables /da/ and /ga/ in English. In isolation, /d/ is typically produced with an occlusion anterior in the vocal tract as the tongue tip makes contact with the roof of the mouth. In contrast, /g/ is produced with a posterior occlusion created by the tongue body. However, the place of articulation for these CVs changes when they are produced in the context of a preceding /al/ or /ar/

syllable (e.g., /al da/). The anterior articulation of /l/ leads to /d/ and /g/ occlusions that are closer to the front of the mouth, whereas the more posterior production of /r/ shifts the /d/ and /g/ occlusions in the opposite direction.

Because the acoustics of speech sounds are a function of the position and movement of the articulators, coarticulation results in context-sensitive acoustics for phonetic segments. For example, /da/ and /ga/ are differentiated in part by the onset frequency of F3; as a result of its more anterior place of articulation, /da/ typically has a higher-frequency F3 onset than /ga/. When produced following /al/, CVs will have a higher F3 frequency onset, due to the shift in place of articulation, than when produced following /ar/. With this in mind, consider the cases of /al ga/ and /ar da/ in which an anterior and a posterior production are paired. The /al/ raises the F3 onset frequency of /g/, whereas the /ar/ lowers the F3 onset frequency of /d/. The result is that the acoustics of the CVs in these two disyllables are quite similar. A recognition system that simply matched the formant transitions of the consonant to templates for /da/ and /ga/ would have trouble identifying these ambiguous consonants.

How do human listeners contend with the context-sensitive acoustics of phonemes? To answer this question, Mann (1980) presented listeners with a series of synthesized CVs varying in F3 onset frequency from a good /da/ to a good /ga/. These target syllables were preceded by natural productions of the context syllables /al/ or /ar/ (with a 50-ms intersyllabic silent gap). Listeners' identifications of the target CVs shifted depending on the preceding context. More /ga/ responses were made following /al/ than following /ar/. These context-moderated perceptions are in the opposite direction of the effects of coarticulation. In production, /al/ contexts lead to more anterior or /da/-like productions. In perception, /al/ contexts lead to more /ga/ identifications. Perception appears to compensate for the effects of coarticulation. Coarticulatory effects on acoustics and apparent perceptual compensation have also been demonstrated for consonant contexts and vowel targets (Holt et al. 2000, Lindblom & Studdert-Kennedy 1967, Nearey 1989), vowel contexts and consonant targets (Holt 1999, Mann & Repp 1981), and vowel contexts with vowel targets (Fowler 1981).

Whereas context-sensitive acoustics are problematic for accounts of speech perception that rely on acoustic pattern recognition of phonemes, these results support predictions of theories that propose gestures as the objects of speech perception. The mapping between acoustics and perception is not transparent, but the mapping from intended gesture and perception is straightforward. Intended and perceived gestures are consistent even though the acoustics are variable. According to MT, the intended gesture is recovered by accessing tacit knowledge of the acoustic consequences of the candidate articulatory motor commands (Mann 1980). According to DRT, acoustics are parsed into responsible gestures as a result of the sensitivity of the perceiver to the dynamics of articulation. From this view, effects of coarticulation serve as information for the identity of the context segment as opposed to obfuscators of the identity of the target segment. Regardless of the mechanism, the factoring of the speech stream into gestures appears to occur independent of

any linguistic representation. Mann (1986) demonstrated that the contexts /al/ and /ar/ can shift CV identification by Japanese speakers who cannot distinguish /r/ and /l/ (as both sounds are mapped to a single phoneme in Japanese). In addition, Fowler et al. (1990) found similar context-dependent shifts in responses of 4- to 5-month-old infants.

A GA account of perceptual compensation for coarticulation would rely on interactions between stimulus attributes in the auditory system or perceptual learning based on correlated features in the input. In support of a general auditory basis for these context effects, Lotto et al. (1997) demonstrated context-dependent responses to CVs in birds. Japanese quail were trained to peck a key when presented /da/ or /ga/. When ambiguous CV stimuli were presented following /al/ or /ar/, birds' responses shifted in the same manner as for humans. The extension of phonetic context effects to Japanese quail casts doubt on the necessity of specialized perceptual mechanisms. Lotto et al. (1997) suggested that the shifts in birds' responses were not due to the factoring of the signal into gestures but to general auditory interactions between the spectral components of the target and context. In particular, they proposed that the context effects are a result of spectral contrast. A redescription of the bird results in terms of acoustic components would be: Following a context with high-frequency F3 offset (/al/), more low-frequency F3 onset responses (/ga/) are obtained. Equivalently, following a context with low-frequency F3 offset (/ar/), more high-frequency F3 onset responses (/da/) are obtained. It should be noted that the proposed auditory contrast is between spectral patterns of higher versus lower energy in particular frequency regions, as opposed to changes in representations of frequency per se.

Evidence for spectral contrast has also been obtained with humans. Lotto & Kluender (1998) presented listeners with members of a /da/-/ga/ series preceded by nonspeech sounds that mimicked some of the important spectral content of the syllables /al/ and /ar/. The contexts were either tone glides modeling the offset transitions of F3 or steady-state tones set at F3 offset frequency. Despite a lack of articulatory or phonemic content, these contexts affected the identification of target CVs. Following high-frequency tones (based on /al/), more /ga/ identifications were obtained. The interaction of speech and nonspeech sounds runs counter to expectations of a modular approach to perception such as MT. Similar nonspeech context effects have been demonstrated for conditions that mimic consonant contexts with vowel targets (Holt et al. 2000) and vowel contexts with consonant targets (Holt 1999). Instead of proposing a special mechanism to handle the complexities of coarticulation, it may be that a general perceptual function allows (and even encourages) humans to coarticulate phonemes.

The spectral contrast account of phonetic context effects has been challenged by recent results of Fowler et al. (2000). They presented listeners with a /da/-/ga/ series preceded by a syllable that was perceptually ambiguous between /al/ and /ar/. The identity of this context was disambiguated by a synchronized video of a speaker producing /al da/ or /ar da/. The resulting identification of the target CV was a function of the visual input. Visual /al/ productions led to more /ga/ responses.

This result is consistent with gestural theories such as MT or DRT. The visual input provides information about the gestures involved in the context syllables and leads to appropriate compensation for coarticulation in perceiving the target. The results are inconsistent with an account that relies strictly on spectral contrast because there is no change in the acoustic makeup of the context in the two conditions. It should be noted that from the perspective of GA, it is quite reasonable to assume that humans learn correlations between visual and auditory input and that their perceptions are a result of a combination of these informational sources (Massaro 1987). That is, a generalist account does not require that all context effects be explained solely by spectral contrast or any purely auditory mechanism.

Despite the fact that visually moderated context effects would be consistent with all of the major theories of speech perception, the findings of Fowler et al. (2000) have recently been brought into question by new results. Stephens & Holt (2002) presented participants the ambiguous context and target CV sounds with aligned video /al/ and /ar/ and a blank screen during the CV. They failed to find a shift in CV identification as a function of context video. This raises the possibility that the effect reported by Fowler et al. was due to visual information associated with the target syllable rather than to the video aligned with the context. Although the videotaped speaker produced /da/ in both conditions, there were differences in the CV portion of the video for /al/ and /ar/ precursors. To examine the effect of these differences, Stephens & Holt (2002) presented the audio and video CV portions of the Fowler et al. stimuli with no auditory or visual context. The resulting identification functions resembled those originally obtained by Fowler et al. with boundary shifts as a function of whether the visual /da/ came from /al da/ or /ar da/. Thus, the identification shifts appear to be due to auditory and visual interactions during the target syllable and not due to visual moderation of the perceived context. These synchronized auditory-visual interactions are well known in speech perception as demonstrated by the McGurk effect (McGurk & McDonald 1976).

Whereas the results of Fowler et al. (2000) do not clearly indicate the existence of visually moderated context effects, there have been several demonstrations of lexically moderated context effects (Elman & McClelland 1988, Magnuson et al. 2003, Samuel & Pitt 2003). For example, Elman & McClelland (1988) presented context words that ended in an ambiguous fricative consonant. This consonant was disambiguated by lexical identity, being perceived as "s" in "copious_" and as "sh" in "Engli_." Despite the lack of acoustic change in this final consonant, identification of succeeding target consonants was shifted as a function of lexical context, from a /d/ following "English" to a /g/ following "copious." This result is difficult to reconcile with current accounts of phonetic context effects because the acoustic (and presumed corresponding gestural) characteristics of the precursor context stimuli are nearly identical across conditions. What remains unresolved is the type of representation on which lexicality asserts its effects. Cognitive models typically propose that lexical effects influence phonemic representations, but they could just as well be influencing gestural or auditory representations (or both).

The problem with phoneme representations here is that context effects have been demonstrated for listeners without the requisite linguistic representations (birds, infants, and Japanese listeners with /l/ and /r/ contexts). To fully account for all the results, future cognitive models of speech will need to incorporate richer auditory and/or gestural representations.

LEARNING SPEECH AND NONSPEECH CATEGORIES

So far, we have presented empirical and theoretical work concerning topics such as categorical perception and context effects that are relevant to the perception of the sounds of any language. However, one of the most important issues in speech perception is how listeners come to perceive sounds in a manner that is particular to their native language. In order to communicate proficiently, a listener must discriminate acoustic variance in the speech signal that is linguistically relevant and to generalize across variance that is irrelevant. Of course, what counts as relevant and irrelevant depends on the phoneme inventory of the specific language.

Before six months of age, infants have a well-documented ability to discriminate many (possibly most) of the sounds that are used contrastively in languages (e.g., Eilers 1977, Eimas 1974, Eimas et al. 1971, Miller & Eimas 1983). This includes the ability to tell apart sounds that are not phonemically distinctive in the infant's language environment (Best et al. 1988, Werker et al. 1981). Before the end of the first year of life, infants start to become perceptually "tuned" to their native language. That is, they respond to speech sounds in a language-specific manner; discriminating acoustic differences between phoneme categories of their language but no longer distinguishing sounds within those categories (Pegg & Werker 1997, Werker & Tees 1984). This change occurs before the development of a substantial lexicon.

In accord with MT, it has been suggested that these early speech perception abilities are indicative of "finely tuned linguistically relevant perceptual abilities" (Miller & Eimas 1983, p. 135) or even an "innately given, universal set of phonetic categories" (Eimas 1991, p.111). These proposals are analogous to the concept of a language acquisition device (LAD) offered by Chomsky (1965) for acquisition of syntax. Presumably, as with LAD, development would consist of retaining those phoneme contrasts that are used in the native language environment and discarding those that are not.

Most recent proposals on speech acquisition have tended to focus on the role of general perceptual learning rather than on innate knowledge and specialized perceptual mechanisms. It is hypothesized that infants acquire phoneme categories through the use of distributional properties of sounds in the language environment along with correlations between attributes of those sounds. This does not require specialized mechanisms, although speech may be a particularly salient signal for infants, and learning processes may be biased to pick up just the kind of information that is important for speech categories (Jusczyk 1993, 1997).

From the perspective of GA, the initial discriminative abilities of infants are a result of their well-developed auditory system, which provides sufficient temporal and frequency resolution (Werner & Bargones 1992). In addition, it seems reasonable to assume that languages tend to use contrasts that are readily distinguishable by language learners. Kuhl (1993) has proposed that much of the initial auditory space of the human infant (and other mammals) is segregated by natural boundaries that underlie many of the speech discrimination results.

Exposure to regularities within a natural language is proposed to lead to a reorganization of perception in terms of phoneme categories or equivalence classes. The information for these categories is present in the statistical properties of the input distributions. For many theorists, these categories provide mappings from acoustics to linguistic elements such as phonemes (e.g., Jusczyk 1993, 1997; Kluender et al. 1998; Kuhl 1993; Lotto 2000). From a DRT perspective, Best (1993, 1995) has offered the Perceptual Assimilation Model (PAM), according to which the initial discriminative abilities of infants are due to the direct recovery of the individual simple articulations that produce the sounds. With exposure to a language, infants begin to group coordinated gestures that are related to the phonemes of the language into equivalence classes.

Despite general agreement that perceptual learning depends on the distributional properties of stimuli, few attempts have been made to explore the actual mechanisms for auditory categorization of complex stimuli such as speech. One exception is a proposal by Kuhl (1991, 1993, 2000) that experience with speech sounds leads to the creation of category prototypes or high-density representations of exemplars that act as "perceptual magnets" that warp the perceptual space. However, the initial findings and subsequent predictions of the magnet model have not been supported (Frieda et al. 1999, Lively & Pisoni 1997, Lotto et al. 1998, Sussman & Lauckner-Morano 1995).

One difficulty of studying speech category learning in infants is a lack of control over the quality and quantity of language experience. In fact, there exists little information about typical speech input distributions for infants. In order to study general learning processes with known input distributions, Kluender et al. (1998) trained birds (starlings) to identify variants of the vowel in "heed" versus the vowel in "hid." The birds readily learned to peck a button when they heard one vowel category and to refrain for the complementary vowel category, and their responses generalized to novel variants of the vowels. Remarkably, the birds' peck rates were highest for those variants that human adult listeners judged as the best members of the vowel category. The correlation between bird responses and human "goodness" judgments was high across categories ($r = 0.99$) and within categories (average $r = 0.71$).

The bird and human data revealed two salient patterns. The first was a higher rate of responding (or higher goodness ratings) for stimuli located relatively far from the category boundary. The second was an increase in response near the centroid of the stimulus distribution used for training. This was the area of highest stimulus density during training, and the response pattern resembles a classic

prototype effect (e.g., Rosch 1978). This pattern indicates that the birds picked up information about the structure of the input distribution even though it was not necessary to perform the task. (Use of a linear boundary between the two categories would be sufficient for perfect performance.) Thus, perceptual systems may be quite sensitive to input distributions of experienced auditory stimuli, and this information may affect later categorization. In support of this conclusion, Maye et al. (2002) reported that the shape of previously experienced distributions could alter responses of human infants to speech sounds. They presented infants with either a bimodal distribution (resembling two categories) or unimodal distribution (resembling a single category) of VOT stimuli. In a subsequent discrimination task, infants with bimodal experience discriminated endpoints of the series (as if they belonged to separate categories) whereas infants with unimodal experience showed poorer discrimination. Taken together with the animal work and experiments on the categorization of nonspeech sounds (Guenther et al. 1999, Lotto 2000), these results are part of a growing literature on the ability of listeners to extract information about the statistics of input distributions. These studies likely will play a substantial role in our understanding of phoneme acquisition.

In summary, from the perspective of GA, the data on infant speech perception can be explained by an interaction between the operating characteristics of the auditory system and processes of perceptual learning. This audition-learning interaction is exemplified in a recent study by Holt et al. (2003). They took advantage of the natural boundaries that have been demonstrated for temporal processing to examine the formation of nonspeech categories. Participants were presented TOT stimuli similar to those used by Pisoni (1977; see Figure 2) and asked to label them as belonging to the experimenter-defined categories *A* or *B*, with the correct answer indicated by feedback. For half the participants, the category distributions were separated by the natural TOT boundary of +20 ms. For the other participants, the distributions were separated by a TOT boundary of +40 ms. For this condition, the natural boundary fell within the *A* category.

Two findings were noteworthy. The first was that participants whose experimenter-defined boundary was consistent with the natural boundary learned the categories much more readily. They required fewer than half as many trials to reach a criterion of 90% correct than subjects with experimenter-defined boundaries that were inconsistent with the natural boundary. That is, because the auditory system provided an initial parsing of the distributions, listeners had little difficulty learning the proper labels for each stimulus. Similarly, separating those categories by a natural boundary may facilitate the task of learning the voice categories of a language. The fact that languages tend to use these natural boundaries may be due to an advantage of learnability.

The second finding of Holt et al. (2003) was that participants assigned to the unnatural boundary condition did eventually learn to categorize the stimuli with high accuracy (greater than 90%). That is, the natural temporal order boundary is not essential to the categorization of TOT stimuli and, by extension, VOT stimuli. Learning processes are flexible enough to overcome some natural auditory

biases. The results of learning can be seen in the discrimination responses of adult speakers of Spanish, who show a much larger peak at the Spanish voice boundary than at the English voice boundary (Williams 1977). The results demonstrate the potential for perceptual learning studies to explain patterns of speech category acquisition.

RELATION BETWEEN SPEECH PRODUCTION AND PERCEPTION

Both MT and DRT assume that there exists a very close relationship between speech production and perception: Talkers produce gestures and listeners perceive them (or, in the case of MT, they perceive the intended gestures). Accordingly, regularities of speech production (e.g., context dependencies in the realization of successive phonemes) should be highly correlated with listeners' perceptual judgments. A wealth of data assures us that such a correlation exists, and on this point there is no serious disagreement among theorists (Diehl & Kluender 1987, Fowler 1986, Liberman 1996). However, GA differs from MT and DRT on how the correlation is to be explained.

GA offers two general accounts of the correlation between speech production and perception, which are, simply stated: Production follows perception, and perception follows production. The first of these is meant to subsume cases in which the need for auditory distinctiveness of phonemes shapes production. For example, as described earlier, languages tend to locate +voice and −voice phonemes so that they are separated by natural auditory boundaries along the VOT dimension. More generally, the sound systems of languages tend to satisfy a principle of dispersion, whereby interphoneme distances are maximized within the available phonetic space to promote intelligibility of utterances even under unfavorable listening conditions. In simulation experiments, the dispersion principle has been shown to predict the structure of common vowel inventories quite accurately, especially when realistic auditory models are used to define a measure of auditory distance (Diehl et al. 2003, Liljencrants & Lindblom 1972).

How is the dispersion principle implemented by talkers? A general answer to this question is provided by the auditory enhancement hypothesis (Diehl & Kluender 1989a,b; Diehl et al. 2001; Kingston & Diehl 1994, 1995), which states that the gestural correlates of individual phonemes are selected to yield mutually enhancing auditory effects. Consider, for example, the vowel /u/, which occurs in most of the world's languages (Maddieson 1984). The acoustic property that distinguishes /u/ from all other vowels is a low-frequency F2. When talkers are required to speak clearly (for example, in the presence of background noise), they typically produce /u/ by retracting and raising the tongue body, enlarging the pharynx by moving the tongue root forward, raising the velum (and thus blocking airflow from the mouth through the nasal cavities), lowering the larynx, protruding the lips, and constricting the lip orifice. Every one of these gestures independently

contributes to lowering the frequency of F2; together they create a maximally distinctive /u/, that is, one that is acoustically (and hence auditorily) most distant from other vowels (Diehl & Kluender 1989a,b). The dispersion principle and the auditory enhancement hypothesis are the main content of the claim that production follows perception.

The other claim of GA is that perception follows production. According to GA, listeners do not recover gestures, but they do perceive the acoustic consequences of gestures. Any regularities of speech production (e.g., context dependencies) will be reflected in the acoustic signal, and, through general mechanisms of perceptual learning, listeners come to make use of the acoustic correlates of these production regularities in judging the phonemic content of speech signals.

An implication of this discussion is that, by itself, the high correlation between speech production and perception is uninformative with respect to the debate between MT, DRT, and GA. All three predict that such a correlation must exist. Distinguishing them empirically requires other kinds of data including (but not restricted to) speech and nonspeech comparisons or human and animal comparisons. To illustrate this point, we consider the McGurk effect (McGurk & MacDonald 1976), where visual speechreading information may actually override inconsistent auditory information in determining the identification of a phoneme. In normal speech communication, visual and auditory cues are consistent, and listeners use both in making phoneme judgments. Both motor theorists (Liberman & Mattingly 1985) and direct realists (Fowler 1986, 1996) have claimed that the McGurk effect and, more generally, the use of auditory and visual information in speech perception, support a gestural account of perception. As Fowler (1996) puts it, "[L]isteners perceive gestures, and some gestures are specified optically as well as acoustically" (p. 1733). However, from the perspective of GA both acoustic and visual cues map perceptually onto phonemes, and the link between these cues can be attributed to perceptual learning. It is worth noting that biologically plausible computational models have demonstrated unsupervised learning of cross-modal categories (e.g., de Sa & Ballard 1998). Thus, a GA account appears to be no less compatible with results such as the McGurk effect than a gestural account.

Appealing to results outside the realm of normal speech perception may break this theoretical impasse. Diehl & Kluender (1989a) noted that a GA account explains not only the integration of auditory and visual information for phonemes but other forms of cue integration as well. For example, when certain acoustic properties of speech are artificially transduced into vibrotactile patterns on the skin, perceivers can learn to use this information along with correlated auditory and visual cues to identify phonemes (see, e.g., Sparks et al. 1978). Because the vibrotactile patterns cannot meaningfully be said to specify gestures, neither MT nor DRT appear to be able to accommodate the result without invoking assumptions similar to those of GA.

In a different attempt to distinguish between gestural and GA accounts of the McGurk effect, Fowler & Dekle (1991) asked listeners to identify a series of synthetic syllables ranging from /ba/ to /ga/ while concurrently viewing a printed

version of either syllable. The authors reasoned that literate perceivers have extensive experience seeing printed words and hearing the words spoken, and therefore if simple association learning is responsible for the McGurk effect, then an analogous result should be observed in their experiment. In fact, the printed syllables had no effect on identification of the synthetic syllables, and Fowler & Dekle concluded that this result was incompatible with an associationist account of the McGurk effect.

A problem with this conclusion is that GA does not view the process of perceptual learning as equivalent to simple associative learning per se. To learn the auditory, visual, or even vibrotactile correlates of a phoneme is to learn what kinds of stimulus properties serve as information (i.e., as perceptual cues) for that phoneme. The relation between a perceptual cue and the object/event that is signaled by the cue is correctly referred to as an association, but it is a very specific kind of association. It is quite unlike, for example, the links between semantic associates (e.g., doctor:nurse, dog:cat, and oak:maple), between objects and their names (dog: "dog"), or between phonemes and their orthographic representations (/b/:"B"). Concerning the latter kind of association, no amount of experience reading aloud is likely to establish an informational relationship between letters and phonemes such that "B" signals that /b/ is now occurring.

In a different analog of the McGurk & MacDonald (1976) experiment, Fowler & Dekle (1991) had listeners identify synthetic /ba/-/ga/ stimuli while concurrently touching the mouth of a talker producing the syllables /ba/ or /ga/. No visual information about the talker was available to the participants. As with the visual version of the McGurk & MacDonald experiment, this haptic version yielded reliable evidence of cross-modal effects on phoneme judgments. According to Fowler & Dekle, these results support a gestural account of speech perception (with both optical and haptic information specifying the gestures), while ruling out a perceptual learning account on the grounds that participants would not have had previous experience perceiving speech haptically.

Below we discuss reasons for denying the claim of DRT that humans use acoustic information to perceive gestures. However, no one would deny that at least some gestures (e.g., lip closing and opening) are visually accessible or that such visual information typically plays a useful role in speech perception. Nor is it surprising that humans can tell whether a talker is closing and opening the lips (as in the production of /ba/) merely by touching the talker's lips. Haptic speech perception may be unusual, but humans have abundant haptic experience with shapes and surface contours in general, ensuring likely success for this special case. GA would certainly not discount the use of gestural information to recognize phonemes in those cases where gestures are perceptually accessible.

As mentioned earlier, GA is not a theory as such but rather a general framework within which particular theoretical claims are formulated and tested. These claims may include competing explanations for the same phenomenon, as in the following example. A well-known correlate of the voice distinction is variation in fundamental frequency (f0): vowels immediately following +voice consonants

tend to have lower f0 values than vowels following −voice consonants. Correspondingly, a lower f0 tends to shift the perceived voice boundary toward higher values of VOT (i.e., more stimuli are judged as +voice). Diehl (1991, Diehl & Molis 1995) claimed that f0 is controlled by talkers as part of a strategy of auditory enhancement of the voice distinction: Voicing during the consonant and a low f0 and F1 near the consonant all contribute to the low frequency periodicity that, by hypothesis, is a main distinctive acoustic correlate of +voice consonants. In this view, f0 affects voice perception for auditory reasons (e.g., integration of low frequency energy) and not because f0 is a learned cue for the voice distinction. However, such a perceptual learning account is clearly compatible with GA. To test the two competing claims, Holt et al. (2001) trained Japanese quail to respond to one of three series of VOT stimuli: one in which VOT and f0 varied in the natural way (shorter VOT, lower f0), one in which the pattern was reversed (shorter VOT, higher f0), and one in which the relation between VOT and f0 was random. For birds trained in the random condition, there was no effect of f0 on responses to novel VOT stimuli, while for the other two groups, responses followed the learned pattern of stimulus covariation. These findings strongly support the perceptual learning account, and appear to rule out the auditory (low-frequency integration) account, of the influence of f0 on VOT perception.

CONCLUDING REMARKS: CHALLENGES TO MT, DRT, AND GA

In this concluding section, we describe what we think are the main challenges to each of the three main theoretical perspectives on speech perception.

Motor Theory

We argue above that a high correlation between measures of speech production and perception is by itself uninformative theoretically because all major perspectives predict such a correlation. Accordingly, the empirical case for MT must ultimately rest on demonstrations of patterns of performance that are specific to speech perception by humans. During the last four decades, motor theorists have described a variety of empirical phenomena that they believed satisfied the condition of speech- and/or human-specificity (Liberman 1996, Liberman et al. 1967, Liberman & Mattingly 1985). In preceding sections, we examined some of these phenomena, including categorical perception and several phonetic context effects, and concluded that they were not, in fact, unique to speech or to human listeners.

Another phenomenon claimed to be diagnostic of perception in the speech mode is duplex perception. When all of a synthetic /da/ or /ga/ syllable except for the F2 transition (which specifies place of articulation) is presented to one ear of a listener and the F2 transition alone is presented in proper temporal alignment to the other ear, the listener experiences two percepts: a nonspeech “chirp” (corresponding to the F2 transition alone) and a full /da/ or /ga/ syllable. Thus, the same

acoustic property is perceived in two very different ways, reflecting, according to Liberman & Mattingly (1985), the operation of both a speech and a nonspeech module that use the same input signal to create representations of distinct sound sources. However, Fowler & Rosenblum (1991) demonstrated an analog of duplex perception for the sound of a slamming door, with the high-frequency portion of the signal presented to one ear and the rest of the signal presented to the other ear. Because it is unlikely that humans have evolved specialized modules for slamming doors, Fowler & Rosenblum concluded that duplex perception does not provide clear evidence for MT.

A main challenge for motor theorists, therefore, is to offer more compelling evidence of genuine speech- and human-specific perceptual performance.

Direct Realist Theory

A core assumption of DRT is that perceivers recover the actual environmental events that structure informational media such as light or sound. Plainly, some environmental properties are perceptually accessible. Among them are the visually and haptically accessible layout of surfaces in the environment and the auditorily accessible location of sound sources. However, certain other environmental properties that structure light or sound are not similarly accessible. For example, organisms that are limited to two or three types of cone photopigments cannot unambiguously recover the spectral distribution of reflected light because every pattern of cone responses is compatible with an infinite set of hypothetical surface reflectances. In order for any environmental property to be perceptually recoverable in principle, there must be information available to the perceiver that uniquely specifies that property. This essential condition is met in the case of visual perception of surface layout [assuming some general constraints such as rigidity (Ullman 1984)] and in the case of auditory perception of sound location (Grantham 1995), but the condition is not met in the case of visual detection of surface reflectance.

The question of interest here is, Do acoustic signals uniquely specify the properties of sound sources such as the vocal tract? The answer appears to be no. Even if one restricts the discussion to anatomically possible vocal tract shapes, there are many different ways to produce a given speech signal. For example, approximately the same formant pattern can be achieved either by rounding the lips, lowering the larynx, or doing a little of both (Riordan 1977). Also, one can produce the distinctively low-frequency F3 of the American English vowel contained in the word "her" by making vocal tract constrictions at the lips, midpalate, or the midpharynx, or at some combination of these places (Ohala 1985, Lindau 1985). Additional evidence that different gestures can yield similar acoustic patterns is presented in Ladefoged et al. (1972), Nearey (1980), and Johnson et al. (1993).

Acoustic ambiguity of source specification also holds outside the domain of speech. For example, in a paper titled "One Cannot Hear the Shape of a Drum," Gordon et al. (1992) proved mathematically that quite different drum shapes can produce identical acoustic signals. Also, the same resonant sound can be initiated by air pressure sources generated by piston-like compression, bellows-like

compression, or by a heat-induced pressure increase in a fixed container. Examples of such source ambiguity appear to be pervasive.

In attempting to solve the "inverse problem" (i.e., mapping speech signals onto vocal tract shapes that produced them), speech-processing engineers have found it necessary to assume various facts such as length or certain other characteristics of the vocal tract (Atal & Hanauer 1971, McGowan 1994). Without such assumptions, the inverse problem appears to be intractable. In principle, this is not a problem for MT, which assumes that the speech module reflects the coevolution in humans of both production and perception. In this view, the human perceiver of speech sounds has implicit knowledge of the speech production apparatus, which presumably can be applied to solve the inverse problem. However, DRT does not have this recourse. According to Fowler (1989), nonhuman listeners can perceive human speech gestures just as humans can (which is why parallels between human and nonhuman perceptual performance are not viewed as surprising from the perspective of DRT). Clearly, Japanese quail and macaque monkeys do not have the implicit knowledge of the human vocal tract that might constrain the solution to the inverse problem.

Thus, a major challenge for DRT is to offer a credible solution to the inverse problem that does not rely on prior knowledge of the human vocal tract.

General Approach

We have described GA as a general framework within which specific hypotheses are formulated and tested. Some of these hypotheses (e.g., threshold of temporal ordering, dispersion and auditory enhancement, spectral and durational contrast, and covariance learning) were discussed in light of relevant findings, and the overall approach seems promising. Nevertheless, the challenges facing GA are daunting. They fall into two general categories, reflecting the dual emphasis of GA on auditory processing and perceptual learning. Our knowledge of mammalian auditory processing is large and growing, but detailed and accurate models are still largely restricted to the auditory periphery. Some of the hypotheses described within the GA framework (e.g., durational contrast) are not independently justified on the basis of known mechanisms of auditory processing and are therefore rather ad hoc. Related to this, there are not yet principled grounds for precisely predicting the conditions under which such hypotheses apply. For example, evidence for durational contrast has been reliably found in some conditions (Diehl & Walsh 1989) but not in others (Fowler 1992). We need to know a great deal more about auditory processing, especially beyond the auditory nerve, to properly constrain our models of speech perception. Particular attention must be focused on the role of neural plasticity at higher levels of the auditory pathway.

Current knowledge about how humans learn speech categories is even more limited. As reviewed earlier, we are beginning to understand how listeners respond to various statistical properties of stimuli when experimenters control the input distributions. However, we lack comprehensive measurements of the statistical properties of natural speech sounds in the listener's environment. Without such

measurements, it is impossible to formulate models of natural language learning with good predictive power.

Therefore, a major challenge for GA is to develop hypotheses based on far more accurate information about the auditory representation of speech and the statistical properties of natural speech.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by NIH Grant R01 DC00427–14 to RLD and NIH Grant R01 DC04674–02 to AJL and LLH. We thank Sarah Sullivan for her help in preparing the manuscript.

The *Annual Review of Psychology* is online at http://psych.annualreviews.org

LITERATURE CITED

Abramson AS, Lisker L. 1970. Discriminability along the voicing continuum: cross language tests. In *Proc. Int. Congr. Phonet. Sci., 6th, Prague, 1967*, pp. 569–73. Prague: Academia

Aslin RN, Pisoni DB, Hennessy BL, Perey AJ. 1981. Discrimination of voice onset time by human infants: new findings and implications for the effects of early experience. *Child Dev.* 52:1135–45

Atal BS, Hanauer SL. 1971. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* 50:637–55

Best CT. 1993. Emergence of language-specific constraints in perception of nonnative speech: a window on early phonological development. In *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, ed. B de Boysson-Bardies, S de Schonen, P Jusczyk, P MacNeilage, J Morton, pp. 289–304. Norwell, MA: Kluwer Acad.

Best CT. 1995. A direct realist view of cross-language speech perception. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed. W Strange, pp. 171–206. Baltimore, MD: New York Press

Best CT, McRoberts GW, Sithole NM. 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *J. Exp. Psychol.: Hum. Percept. Perform.* 14:345–60

Brunswik E. 1956. *Perception and the Representative Design of Psychological Experiments.* Berkeley: Univ. Calif. Press

Chomsky N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press

Delattre PC, Liberman AM, Cooper FS. 1951. Voyelles synthétiques à deux formants et voyelles cardinales. *Le Maitre Phon.* 96:30–37

Delattre PC, Liberman AM, Cooper FS. 1955. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* 27:769–73

Delattre PC, Liberman AM, Cooper FS. 1964. Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. *Stud. Linguist.* 18:104–21

Delattre PC, Liberman AM, Cooper FS, Gerstman LJ. 1952. An experimental study of the acoustic determinants of vowel color. *Word* 8:195–210

de Sa VR, Ballard DH. 1998. Category learning through multimodal sensing. *Neurol. Comput.* 10:1097–117

Diehl RL. 1987. Auditory constraints on speech perception. In *The Psychophysics of Speech*

Perception, ed. MEH Schouten, 39:210–19. Dordrecht: Martimus-Nihboff

Diehl RL. 1991. The role of phonetics within the study of language. *Phonetica* 48:120–34

Diehl RL, Kluender KR. 1987. On the categorization of speech sounds. In *Categorical Perception*, ed. S Harnad, pp. 226–53. London: Cambridge Univ. Press

Diehl RL, Kluender KR. 1989a. On the objects of speech perception. *Ecol. Psychol.* 2:121–44

Diehl RL, Kluender KR. 1989b. Reply to commentators. *Ecol. Psychol.* 1:195–225

Diehl RL, Lindblom B, Creeger CP. 2003. Increasing realism of auditory representations yields further insights into vowel phonetics. In *Proc. Int. Congr. Phon. Sci., 15th, Barcelona.* 2:1381–84. Adelaide: Causal Publications

Diehl RL, Molis MR. 1995. Effects of fundamental frequency on medial [voice] judgments. *Phonetica* 52:188–95

Diehl RL, Molis MR, Castleman WA. 2001. Adaptive design of sound systems: some auditory considerations. In *The Role of Perceptual Phenomena in Phonological Theory*, ed. K Johnson, E Hume, pp. 123–39. San Diego: Academic

Diehl RL, Souther AF, Convis CL. 1980. Conditions on rate normalization in speech perception. *Percept. Psychophys.* 27:435–43

Diehl RL, Walsh MA. 1989. An auditory basis for the stimulus-length effect in the perception of stops and glides. *J. Acoust. Soc. Am.* 85:2154–64

Diehl RL, Walsh MA, Kluender KR. 1991. On the interpretability of speech/nonspeech comparisons: a reply to Fowler. *J. Acoust. Soc. Am.* 89:2905–9

Dooling RJ, Best CT, Brown SD. 1995. Discrimination of synthetic full-formant and sinewave /ra-la/ continua by budgerigars (*Melopsittacus undulatus*) and zebra finches (*Taeniopygia guttata*). *J. Acoust. Soc. Am.* 97:1839–46

Eilers RE. 1977. Context-sensitive perception of naturally produced stop and fricative consonants by infants. *J. Acoust. Soc. Am.* 61:1321–36

Eimas PD. 1974. Auditory and linguistic processing of cues for place of articulation by infants. *Percept. Psychophys.* 16:513–21

Eimas PD. 1991. Comment: some effects of language acquisition on speech perception. In *Modularity and the Motor Theory of Speech Perception*, ed. IG Mattingly, M Studdert-Kennedy, pp. 111–16. Hillsdale, NJ: Erlbaum

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J. 1971. Speech perception in infants. *Science* 171:303–6

Elman JL, Diehl RL, Buchwald SE. 1977. Perceptual switching in bilinguals. *J. Acoust. Soc. Am.* 62:971–74

Elman JL, McClelland JL. 1988. Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *J. Mem. Lang.* 27:143–65

Fowler CA. 1980. Coarticulation and theories of extrinsic timing. *J. Phon.* 8:113–33

Fowler CA. 1981. Production and perception of coarticulation among stressed and unstressed vowels. *J. Speech Hear. Res.* 24:127–39

Fowler CA. 1984. Segmentation of coarticulated speech in perception. *Percept. Psychophys.* 36:359–68

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

Fowler CA. 1989. Real objects of speech perception: a commentary on Diehl and Kluender. *Ecol. Psychol.* 1:145–60

Fowler CA. 1990. Sound-producing sources as objects of perception: rate normalization and nonspeech perception. *J. Acoust. Soc. Am.* 88:1236–49

Fowler CA. 1991. Auditory perception is not special: We see the world, we feel the world, we hear the world. *J. Acoust. Soc. Am.* 89: 2910–15

Fowler CA. 1992. Vowel duration and closure duration in voiced and unvoiced stops: There are no contrast effects here. *J. Phon.* 20:143–65

Fowler CA. 1994. Speech perception: direct

realist theory. In *The Encyclopedia of Language and Linguistics*, ed. RE Asher, pp. 4199–203. Oxford: Pergamon

Fowler CA. 1996. Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.* 99:1730–41

Fowler CA, Best CT, McRoberts GW. 1990. Young infants' perception of liquid coarticulatory influences on following stop consonants. *Percept. Psychophys.* 48:559–70

Fowler CA, Brown JM, Mann VA. 2000. Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *J. Exp. Psychol.: Hum. Percept. Perform.* 26:877–88

Fowler CA, Dekle D. 1991. Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol.: Hum. Percept. Perform.* 17:816–28

Fowler CA, Rosenblum LD. 1991. The perception of phonetic gestures. In *Modularity and the Motor Theory of Speech Perception*, ed. IG Mattingly, M Studdert-Kennedy, pp. 33–59. Hillsdale NJ: Erlbaum

Fowler CA, Smith MR. 1986. Speech perception as "vector analysis": an approach to the problems of invariance and segmentation. In *Invariance and Variability in Speech Processes*, ed. JS Perkell, DH Klatt, pp. 123–39. Hillsdale, NJ: Erlbaum

Frieda EM, Walley AC, Flege JE, Sloane ME. 1999. Adults' perception of native and nonnative vowels: implications for the perceptual magnet effect. *Percept. Psychophys.* 61:561–77

Gibson JJ. 1966. *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin

Gibson JJ. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin

Gordon C, Webb DL, Wolpert S. 1992. One cannot hear the shape of a drum. *Bull. Am. Math. Soc.* 27:134–38

Grantham DW. 1995. Spatial hearing and related phenomena. In *Hearing*, ed. B Moore, pp. 297–345. San Diego, CA: Academic

Guenther FH, Husain FT, Cohen MA, Shinn-Cunningham BG. 1999. Effects of categorization and discrimination training on auditory perceptual space. *J. Acoust. Soc. Am.* 106:2900–12

Hammarberg R. 1976. The metaphysics of coarticulation. *J. Phon.* 4:353–63

Hirsh IJ. 1959. Auditory perception of temporal order. *J. Acoust. Soc. Am.* 31:759–67

Hirsh IJ, Sherrick CE. 1961. Perceived order in different sense modalities. *J. Exp. Psychol.* 62:423–32

Holt LL. 1999. *Auditory constraints on speech perception: an examination of spectral contrast*. PhD thesis. Univ. Wis. 132 pp.

Holt LL, Lotto AJ, Diehl RL. 2003. Perceptual discontinuities and categorization: Implications for speech perception. *J. Acoust. Soc. Am.* 113:2255

Holt LL, Lotto AJ, Kluender KR. 1998. Incorporating principles of general learning in theories of language acquisition. In *Chicago Linguistic Society: The Panels*, ed. M Gruber, CD Higgins, KS Olson, T Wysocki, 34:253–68. Chicago, IL: Chicago Linguist. Soc.

Holt LL, Lotto AJ, Kluender KR. 2000. Neighboring spectral content influences vowel identification. *J. Acoust. Soc. Am.* 108:710–22

Holt LL, Lotto AJ, Kluender KR. 2001. Influence of fundamental frequency on stop-consonant voicing perception: a case of learned covariation or auditory enhancement? *J. Acoust. Soc. Am.* 109:764–74

Johnson K, Flemming E, Wright R. 1993. The hyperspace effect: phonetic targets are hyperarticulated. *Language* 69:505–28

Jusczyk PW. 1993. From general to language-specific capacities: the WRAPSA model of how speech perception develops. *J. Phon.* 21:3–28

Jusczyk PW. 1997. *The Discovery of Spoken Language*. Cambridge, MA: MIT Press

Jusczyk PW, Pisoni DB, Walley A, Murray J. 1980. Discrimination of relative onset time of two-component tones by infants. *J. Acoust. Soc. Am.* 67:262–70

Kingston J, Diehl RL. 1994. Phonetic knowledge. *Language* 70:419–54

Kingston J, Diehl RL. 1995. Intermediate properties in the perception of distinctive feature

values. In *Phonology and Phonetics: Papers in Laboratory Phonology*, ed. B Connell, A Arvaniti, 4:7–27. London: Cambridge Univ. Press

Kluender KR. 1994. Speech perception as a tractable problem in cognitive science. In *Handbook of Psycholinguistics*, ed. MA Gernsbacher, pp. 173–217. San Diego, CA: Academic

Kluender KR, Diehl RL, Killeen PR. 1987. Japanese quail can learn phonetic categories. *Science* 237:1195–97

Kluender KR, Lotto AJ, Holt LL, Bloedel SL. 1998. Role of experience for language-specific functional mappings of vowel sounds. *J. Acoust. Soc. Am.* 104:3568–82

Krakow RA, Beddor PS, Goldstein LM, Fowler CA. 1988. Coarticulatory influences on the perceived height of nasal vowels. *J. Acoust. Soc. Am.* 83:1146–58

Kuhl PK. 1981. Discrimination of speech by nonhuman animals: basic auditory sensitivities conducive to the perception of speech-sound categories. *J. Acoust. Soc. Am.* 95:340–49

Kuhl PK. 1986. Theoretical contributions of tests on animals to the special mechanisms debate in speech. *J. Exp. Biol.* 45:233–65

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

Kuhl PK. 1992. Speech prototypes: studies on the nature, function, ontogeny and phylogeny of the "centers" of speech categories. In *Speech Perception, Production and Linguistic Structure*, ed. Y Tohkura, E Vatikiotis-Bateson, Y Sagisaka, pp. 239–64. Tokyo, Jpn: Ohmsha

Kuhl PK. 1993. Innate predispositions and the effects of experience in speech perception: the native language magnet theory. In *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, ed. B de Boysson-Bardies, S de Schonen, P Jusczyk, P MacNeilage, J Morton, pp. 259–74. Dordrecht, Netherlands: Kluwer Acad.

Kuhl PK. 2000. Language, mind, and the brain: Experience alters perception. In *The New Cognitive Neurosciences*, ed. MS Gazzaniga, pp. 99–115. Cambridge, MA: MIT Press

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

Kuhl PK, Miller JD. 1978. Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. *J. Acoust. Soc. Am.* 63:905–17

Kuhl PK, Padden DM. 1982. Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Percept. Psychophys.* 32:542–50

Kuhl PK, Padden DM. 1983. Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J. Acoust. Soc. Am.* 73:1003–10

Ladefoged P, DeClerk J, Lindau M, Papcun G. 1972. An auditory-motor theory of speech production. *UCLA Work. Pap. Phon.* 22:48–75

Lane H. 1965. The motor theory of speech perception: a critical review. *Psychol. Rev.* 72:275–309

Lasky RE, Syrdal-Lasky A, Klein RE. 1975. VOT discrimination by four to six and a half month old infants from Spanish environments. *J. Exp. Child Psychol.* 20:215–25

Liberman AM. 1957. Some results of research on speech perception. *J. Acoust. Soc. Am.* 29:117–23

Liberman AM. 1996. Introduction: some assumptions about speech and how they changed. In *Speech: A Special Code.* Cambridge, MA: MIT Press

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. 1967. Perception of the speech code. *Psychol. Rev.* 74:431–61

Liberman AM, Delattre P, Cooper FS. 1952. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.* 65:497–516

Liberman AM, Delattre PC, Cooper FS, Gerstman LJ. 1954. The role of consonant-vowel

transitions in the stop and nasal consonants. *Psychol. Monogr.* 68:1–13

Liberman AM, Delattre PC, Gerstman LJ, Cooper FS. 1956. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. Exp. Psychol.* 52:127–37

Liberman AM, Harris K, Eimas P, Lisker L, Bastian J. 1961a. An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. *Lang. Speech* 4:175–95

Liberman AM, Harris KS, Hoffman HS, Griffith BC. 1957. The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54:358–68

Liberman AM, Harris KS, Kinney JA, Lane H. 1961b. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *J. Exp. Psychol.* 61:379–88

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

Liberman AM, Mattingly IG, Turvey MT. 1972. Language codes and memory codes. In *Coding Processes in Human Memory*, ed. AW Melton, E Martin, pp. 307–34. Washington, DC: Winston

Liljencrants J, Lindblom B. 1972. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48:839–62

Lindau M. 1985. The story of /r/. In *Phonetic Linguistics*, ed. V Fromkin, pp. 157–68. Orlando, FL: Academic

Lindblom BEF, Studdert-Kennedy M. 1967. On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am.* 42:830–43

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

Lisker L, Abramson AS. 1970. The voicing dimension: some experiments in comparative phonetics. In *Proc. Int. Congr. Phon. Sci., 6th, Prague, 1967*, pp. 563–67. Prague: Academia

Lively SE, Pisoni DB. 1997. On prototypes and phonetic categories: a critical assessment of the perceptual magnet effect in speech perception. *J. Exp. Psychol.: Hum. Percept. Perform.* 23:1665–79

Lotto AJ. 2000. Language acquisition as complex category formation. *Phonetica* 57:189–96

Lotto AJ, Kluender KR. 1998. General contrast effects of speech perception: effect of preceding liquid on stop consonant identification. *Percept. Psychophys.* 60:602–19

Lotto AJ, Kluender KR, Holt LL. 1997. Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *J. Acoust. Soc. Am.* 102:1134–40

Lotto AJ, Kluender KR, Holt LL. 1998. The perceptual magnet effect depolarized. *J. Acoust. Soc. Am.* 103:3648–55

Maddieson I. 1984. *Patterns of Sound*. London: Cambridge Univ. Press

Magnuson JS, McMurray B, Tanenhaus MK, Aslin RN. 2003. Lexical effects on compensation for coarticulation: the ghost of Christmash past. *Cogn. Sci.* 27:285–98

Mann VA. 1980. Influence of preceding liquid on stop-consonant perception. *Percept. Psychophys.* 28:407–12

Mann VA. 1986. Distinguishing universal and language-dependent levels of speech perception: evidence from Japanese listeners' perception of English "l" and "r". *Cognition* 24:169–96

Mann VA, Repp BH. 1981. Influence of preceding fricative on stop consonant perception. *J. Acoust. Soc. Am.* 69:548–58

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

Massaro DW, Oden GC. 1980. Evaluation and integration of acoustic features in speech perception. *J. Acoust. Soc. Am.* 67:996–1013

Mattingly IG, Liberman AM, Syrdal AK, Halwes T. 1971. Discrimination in speech and nonspeech modes. *Cogn. Psychol.* 2:131–57

Maye J, Werker JF, Gerken L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82:B101–11

McGowan RS. 1994. Recovering articulatory

movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary results. *Speech Commun.* 14:19–49

McGurk H, MacDonald J. 1976. Hearing lips and seeing voices. *Nature* 264:746–47

Miller JD, Wier CC, Pastore RE, Kelly WJ, Dooling RJ. 1976. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: an example of categorical perception. *J. Acoust. Soc. Am.* 60:410–17

Miller JL. 1987. Rate-dependent processing in speech perception. In *Progress in the Psychology of Language*, ed. A Ellis, pp. 119–57. Hillsdale, NJ: Erlbaum

Miller JL, Eimas PD. 1983. Studies on the categorization of speech by infants. *Cognition* 12:135–65

Miller JL, Liberman AM. 1979. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Percept. Psychophys.* 25:457–65

Nearey TM. 1980. On the physical interpretation of vowel quality: cinefluorographic and acoustic evidence. *J. Phon.* 8:213–41

Nearey TM. 1989. Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 85:2088–113

Nearey TM. 1990. The segment as a unit of speech perception. *J. Phon.* 18:347–73

Nearey TM, Hogan J. 1986. Phonological contrast in experimental phonetics: relating distributions of production data to perceptual categorization curves. In *Experimental Phonology*, ed. JJ Ohala, J Jaeger, pp. 141–61. New York: Academic

Ohala JJ. 1985. Around flat. In *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, ed. VA Fromkin, pp. 223–41. Orlando, FL: Academic

Ohala JJ. 1996. Speech perception is hearing sounds, not tongues. *J. Acoust. Soc. Am.* 99: 1718–25

Pastore RE. 1981. Possible psychoacoustic factors in speech perception. In *Perspectives on the Study of Speech*, ed. PD Eimas, JL Miller, pp. 165–205. Hillsdale, NJ: Erlbaum

Pegg JE, Werker JF. 1997. Adult and infant perception of two English phones. *J. Acoust. Soc. Am.* 102:3742–53

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

Repp BH. 1981. On levels of description in speech research. *J. Acoust. Soc. Am.* 69: 1462–64

Riordan CJ. 1977. Control of vocal-tract length in speech. *J. Acoust. Soc. Am.* 62:998–1002

Rosch EH. 1978. Principles of categorization. In *Cognition and Categorization*, ed. E Rosch, B Lloyd, pp. 28–48. Hillsdale, NJ: Erlbaum

Samuel AG, Pitt MA. 2003. Lexical activation (and other factors) can mediate compensation for coarticulation. *J. Mem. Lang.* 48: 416–34

Simos PG, Diehl RL, Breier JI, Molis MR, Zouridakis G, Papanicolaou AC. 1998. MEG correlates of categorical perception of a voice onset time continuum in humans. *Cogn. Brain Res.* 7:215–19

Sinex DG, McDonald LP, Mott JB. 1991. Neural correlates of nonmonotonic temporal acuity for voice onset time. *J. Acoust. Soc. Am.* 90:2441–49

Sparks DW, Kuhl P, Edmonds AE, Gray GP. 1978. Investigating the MESA (Multipoint Electrotactile Speech Aid): the transmission of segmental features of speech. *J. Acoust. Soc. Am.* 63:246–57

Stephens JD, Holt LL. 2002. Are context effects in speech perception modulated by visual information? *43rd Annu. Meet. Psychon. Soc., Kansas City, MO.*

Stevens KN, Klatt DH. 1974. Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Am.* 55:653–59

Summerfield Q. 1981. On articulatory rate and perceptual constancy in phonetic perception. *J. Exp. Psychol.: Hum. Percept. Perform.* 7:1074–95

Sussman HM, Fruchter D, Hilbert J, Sirosh J. 1998. Linear correlates in the speech signal: the orderly output constraint. *Behav. Brain Sci.* 21:241–99

Sussman JE, Lauckner -Morano VJ. 1995. Further tests of the "perceptual magnet effect" in the perception of [i]: identification and change-no-change discrimination. *J. Acoust. Soc. Am.* 97:539–52

Ullman S. 1984. Maximizing rigidity: the incremental recovery of 3-D structure from rigid and nonrigid motion. *Perception* 13:255–74

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

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

Werner LA, Bargones JY. 1992. Psychoacoustic development of human infants. In *Advances in Infancy Research*, ed. C Rovee-Collier, L Lipsitt, 7:103–45. Norwood, NJ: Ablex

Williams L. 1977. The perception of stop consonant voicing by Spanish-English bilinguals. *Percept. Psychophys.* 21:289–97