CHAPTER 1

Speech Perception

Carol A. Fowler and James S. Magnuson

Speech perception refers to the means by which acoustic and sometimes visual or even haptic speech signals are mapped onto the language forms (words and their component consonants and vowels) that language users know. For the purposes of this review, we will address three aspects of the language user's perceptual task. We identify as *phonetic perception* the task of extracting information from stimulation about language forms. Next we address how perceivers cope with or even exploit the enormous variability in the language forms that talkers produce. Finally, we address issues associated with lexical access.

1 Phonetic perception

For spoken messages to have their intended effect, minimally listeners/observers have to recognize the language forms, especially the words, that talkers produce. Having accomplished that, they can go on to determine what speakers mean or intend by what they say. The requirement that listeners characteristically successfully identify speakers' language forms has been called the *parity requirement* (Liberman and Whalen, 2000), and a benchmark by which a theory of phonetic perception may be evaluated is its ability to explain parity achievement. In this section, we focus specifically on how listeners extract information from acoustic or other-modal stimulation to identify language forms. In later sections, we address other sources of information that listeners may use.

Listeners encounter acoustic speech signals and often the facial speech gestures of the speaker. A task for speech perception researchers is to determine what is immediately perceived that allows perceptual recovery of language forms. One idea is that, because the main source of information that listeners receive is acoustic, they perceive some auditory transformation of an acoustically represented word. For example, Klatt (1979) suggested that words in the lexicon were, among other representations, represented as sequences of spectra that might be matched to spectra of input words.

This view may be short-sighted, however. Language is a generative system, and its

generativity depends on its compositionality. At the level of relevance here, consonants and vowels combine systematically into words, enabling language users to know, coin, produce, and perceive many tens of thousands of words. Accordingly, words, consonants, and vowels, among other linguistic units, are components of their language competence. Spontaneous errors of speech production occur in which individual consonants and vowels move or are substituted one for the other (e.g., Shattuck-Hufnagel, 1979; but see later in this chapter for a qualification), so we know that, as speakers, language users compose words of consonants and vowels. The need for parity in spoken communications suggests that listeners typically recover the language forms that talkers produce. Here, we will assume that listeners to speech perceive, among other linguistic units, words, consonants, and vowels.

What are consonants and vowels? In one point of view, they are cognitive categories that reside in the minds of speakers/hearers (e.g., Pierrehumbert, 1990). In another, they are actions of the vocal tracts of speakers (e.g., Goldstein and Fowler, 2003). This theoretical disagreement is important.

From the former perspective, speakers do not literally produce language forms. Among other reasons, they do not because they coarticulate when they speak. That is, they temporally overlap actions to implement one consonant or vowel with actions to implement others. The overlap distorts or destroys the transparency of the relation between acoustic signal and phonological segment. Accordingly, the acoustic signal at best can provide cues to the consonants and vowels of the speaker's message. Listeners perceive the cues and use them as pointers to mental phonological categories. Coarticulation creates the (lack of) invariance problem – that the same segment in different contexts can be signaled by different acoustic structures. It also creates the segmentation problem, that is, the problem of recovering discrete phonetic segments from a signal that lacks discrete acoustic segments.

The second point of view reflects an opinion that, in the course of the evolution of language, the parity requirement shaped the nature of language, and, in particular, of language forms. In consequence, language forms, being the means that languages provide to make linguistic messages public, optimally should be things that can be made public without being distorted or destroyed. In short, language forms should be vocal tract actions (phonetic gestures; e.g., Goldstein and Fowler, 2003). Coarticulation and, in particular, resistance to it when its effects would distort or destroy defining properties of language forms, does not distort or destroy achievement of gestures (e.g., Fowler and Saltzman 1993).

In the remainder of this chapter, we discuss current knowledge about the information that supports phonetic perception by way of a brief historical review of the key acoustic and perceptual discoveries in speech research, and we consider how these discoveries motivated past and current theories of speech perception. Next, we address how variability contributes to the lack of invariance problem – the apparent lack of an invariant mapping from the speech signal to phonetic percepts – and discuss challenges to current theories. Then, we discuss the interface of speech perception with higher levels of linguistic processing. We will close the chapter with a discussion of what we view to be the most pressing questions for theories of speech perception.

1.1 *What information supports phonetic perception?*

In the early years of research on phonetic perception at Haskins Laboratories (for a historical overview see Liberman, 1996), researchers used the sound spectrograph to represent acoustic speech signals in a way that made some of its informative structure visible. In addition, they used a Pattern Playback, designed and built at Haskins, to transform schematic spectrographic displays into sound. With these tools, they could guess from spectrographic displays what acoustic structure might be important to the identification of a syllable or consonant or vowel, preserve just that structure by

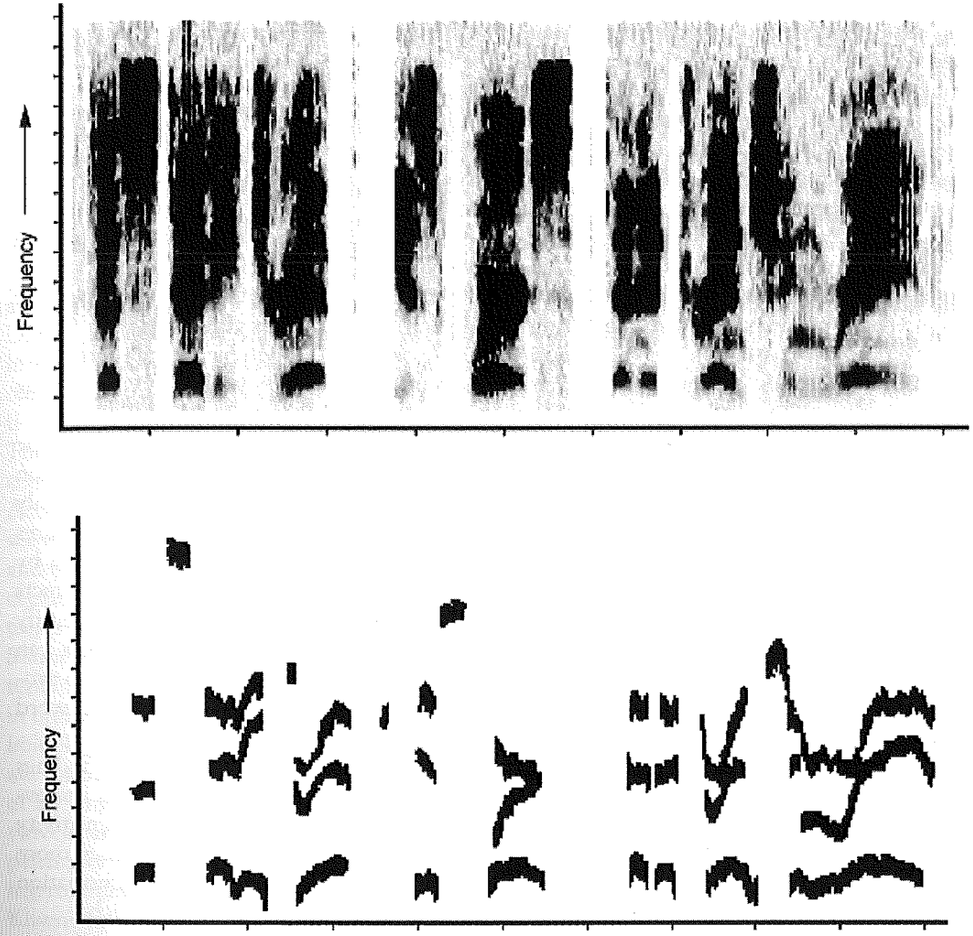

Figure 1.1. A comparison of spectrographic displays of normal (top) and sinewave speech. The sentence depicted in both instances is: "The steady drip is worse than a drenching rain." For more examples, including audio, see http://www.haskins.yale.edu/research/sws.html. Used with the permission of Philip Rubin, Robert Remez, and Haskins Laboratories.

producing a schematic representation of it, and ask whether, converted to sound by the Playback, the schematic representation preserved the phonetic properties of the speech. This research showed them the extent of the acoustic consequences of coarticulation. Acoustic speech signals do not consist of sequences of discrete phone-sized segments, and the acoustic structure that provides information about consonants and vowels is everywhere highly context sensitive. Haskins researchers made an effort to catalogue the variety of acoustic cues that could be used to identify consonants and vowels in their various coarticulatory contexts.

In recent years, researchers have found that the cues uncovered in that early research, acoustic reflections of *formants* – resonant frequencies of the vocal tract that show up as dark horizontal bands in spectrographic displays such as in Figure 1.1 – formant transitions, noise bursts for stops,

intervals of noise for fricatives, and so forth, do not exhaust what serves as information for listeners. For example, in sinewave speech, center frequencies of formants are replaced by single sinewaves and even frication noise is represented by a sinewave (see Figure 1.1). These signals are caricatures of speech signals, and they lack most traditional speech cues (e.g., Remez et al., 1981). That is, they lack a fundamental frequency and harmonics: The sinewaves lack the bandwidth of formants; they lack frication noise, stop bursts, and virtually all distinctive cues proposed to support phonetic perception. They leave more or less intact information signaling dynamic change. They sound bizarre, but they can be highly intelligible, permitting phonetic transcription and even identification of familiar speakers (Remez, Fellowes, and Rubin, 1997).

Other radical transformations of the acoustic signal quite different from the sinewave transformation also permit phonetic perception. For example, in noise-vocoded speech, the fine structure of an acoustic speech signal (effectively, the source) can be replaced with noise while the speech envelope (the filter) is retained. If this transformation is accomplished with as few as four frequency bands, speech is highly intelligible (Smith, Delgutte, and Oxenham, 2002). Smith et al. also obtained intelligible speech with chimaeric speech made in the complementary way, with fine structure preserved and envelope replaced by that of another sound.

A conclusion from the findings that these radical transformations of the acoustic signal, so unlike speech and so unlike each other, yield intelligible signals must be that there is massive redundancy in natural speech signals. Signals are informationally rich, not impoverished as implied by early research on speech.

We also know that phonetic information is conveyed by the face. Speech is better identified in noise if perceivers can see the face of the speaker (Sumby and Pollack, 1954). And it can be tracked more successfully in the context of competing speech emanating from the same location in space if the speaker's face, spatially displaced from the sound source, is visible to the perceiver (Driver, 1996). The much-studied McGurk effect (e.g., McGurk and MacDonald, 1976) also shows that perceivers extract phonetic information from the face. In that phenomenon, a face mouthing one word or syllable, say /da/, is dubbed with a different word or syllable, say /ma/. With appropriate selection of stimuli, perceivers often report hearing a word or syllable that integrates information from the two modalities. A typical percept given the example of visible /da/ and acoustic /ma/ is /na/, which has the place of articulation of the visible syllable, but the voicing and nasality of the acoustic syllable.

Cross-modal integration of phonetic as well, indeed, as indexical information occurs even when the information is highly impoverished. Even when facial gestures are provided by point lights[1] and speech by sinewaves, listeners show McGurk effects (Rosenblum and Saldana, 1998), can identify speakers (Rosenblum et al., 2002) and can determine which visible speaker of two produced a given acoustically presented word (Lachs, 2002; Kamachi et al., 2003).

In summary then, although early findings suggested that the acoustic signal is impoverished in the sense that context sensitivity precludes invariance and transparency, more recent findings suggest that the information available to the perceiver is very rich.

1.2 *Theories of phonetic perception*

Theories of phonetic perception partition into two broad categories. One class of theories (e.g., Diehl and Kluender, 1989; Sawusch and Gagnon, 1995) holds that auditory systems pick out cues in the acoustic speech signal and use the cues to identify mental phonological categories. Another class of theories (e.g., Fowler, 1986; Liberman and Mattingly, 1985) holds that listeners to speech use acoustic structure as information about its causal source, the linguistically

1 In this procedure, light reflecting patches are placed on the face and speakers are filmed in the dark so that only the patches can be seen.

significant vocal tract actions of the speaker. Those vocal tract actions are phonological categories (e.g., Goldstein and Fowler, 2003) or else point to them (Liberman and Mattingly, 1985). Gesture theories differ with respect to whether they do (the motor theory of Liberman and colleagues) or do not (the direct realist theory of Fowler and colleagues) invoke a specialization of the brain for speech perception.

An example of an auditory theory is the auditory enhancement theory proposed by Diehl and Kluender (1989). In that theory, as in all theories in this class, identification of consonants and vowels is guided by acoustic cues as processed by the auditory system. The auditory cues are used to identify the consonant or vowel conveyed by the speaker. According to Diehl and Kluender, we can see evidence of the salience of acoustic cues and of auditory processing in phonetic perception in the nature of the sound inventories that language communities develop. Sound inventories in languages of the world tend to maximize auditory distinctiveness in one way or another. For example, approximately ninety-four percent of front vowels are unrounded in Maddieson's (1984) survey of 317 languages; a similar percentage of back vowels are rounded. The reason for that pairing of frontness/backness and unrounding/rounding, according to Diehl and Kluender, is that both the backing gesture and the rounding gesture serve to lengthen the front cavity of the vocal tract; fronting without rounding keeps it short. Therefore, the two gestures conspire, as it were, either to lower (backing and rounding) or to raise the second formant, making front and back vowels acoustically more distinct than if the rounding gesture were absent or, especially, if the pairing were of rounding and fronting rather than backing.

Evidence seen as particularly compatible with auditory theories are findings in some studies that nonhuman animals appear to perceive speech as humans do (see the next section for a more detailed discussion). For most auditory theorists, it is patent that animals are incapable of perceiving human speech gestures. Therefore their perceptions must be guided by acoustic cues mapped neither to gestures nor to phonological categories. Moreover, nonhuman animals do not have a specialization of the brain for human speech perception; therefore, their perception of human speech must be an achievement of their auditory system. Parallel findings between human and nonhuman animals imply that humans do not perceive gestures either and do not require a specialization for speech. Compatibly, findings suggesting that nonspeech signals and speech signals are perceived in parallel ways are seen to contradict the ideas that gestures are perceived and that a specialization of the brain for speech achieves speech perception.

The first impetus for development of gesture theories was a pair of complementary findings. One finding (Liberman, Delattre, and Cooper, 1952) was that, in synthetic syllables, the same stop burst, centered at 1440 Hz placed before steady state formants for /i/ or /u/, led perceivers to hear /p/. Placed before /a/, they heard /k/. The second finding (Liberman et al., 1954), was that two-formant synthetic /di/ and /du/ had remarkably different second formant transitions. That for /di/ was high and rising; that for /du/ was low and falling. Yet the second formant transition was the information that identified the consonants in those synthetic syllables as /d/.

Together these two findings appear to tell a clear story. In the first, to produce a burst at 1440 Hz requires that a labial constriction gesture coarticulate with /i/ or /u/; to produce the same burst before /a/ requires coarticulation of a velar constriction gesture with a gesture or gestures for the vowel. Coarticulation also underlies the second finding. The same alveolar constriction released into a vocal tract configuration for the vowel /i/ will produce a high rising second formant; released into the configuration for /u/, it will produce a low falling second formant. As Liberman put it in 1957, "when articulation and sound wave go their separate ways, which way does perception go? The answer so far is clear. The perception always goes with articulation" (p. 121).

These findings led to the development of the motor theory of speech perception (e.g., Liberman 1957; Liberman et al., 1967; Liberman and Mattingly 1985; Liberman and Whalen, 2000; for a recent evaluation of the motor theory, see Galantucci, Fowler, and Turvey, 2006). In the 1967 version of that theory, coarticulation is proposed to be essential for the efficient transmission of speech. However, it creates difficulties for the perceiver that a specialization of the brain, unique to humans, evolved to handle. The specialization, later identified as a *phonetic module* (Liberman and Mattingly, 1985), was for both production of coarticulated speech and its perception. The evidence that, in the view of Liberman and his colleagues, revealed that listeners perceive speech gestures, suggested to them that the phonetic module involved the speech motor system in the act of perception, using a process of analysis by synthesis.

In a different theory of gesture perception inspired by Gibson's (e.g., 1966; 1979) more general perceptual theory, Fowler (1986; 1994) proposed that listeners perceive linguistically significant actions of the vocal tract (phonetic gestures) because acoustic signals, caused by the gestures, provide information about them. In this direct realist account, speech perception was proposed to be like perception of every other sort (contra occasional claims that direct realism requires special-purpose mechanisms for gesture perception). Perceivers' sense organs are stimulated by proximal stimuli that provide information for their causal source in the environment. Just as perceivers see objects and events rather than reflected light, and just as they feel object properties rather than the skin deformations that inform about them, they hear sounding events, not the acoustic signals that they cause.

The motor theory and direct realism are equally supported or challenged by most relevant evidence. For example, they are equally supported by the findings of Liberman, et al. (1952; 1954) described earlier. They are equally challenged, for example, by certain comparisons of speech and nonspeech perception. They can be differentiated, however, by research, some of which is described later in this chapter, that addresses the existence of a specialization for speech perception.

Following are some of the research findings that all theories of phonetic perception are required to explain.

1.2.1 CATEGORICAL PERCEPTION

Categorical perception was an early finding in the history of the study of speech perception by experimental psychologists (Liberman et al., 1957). When listeners were asked to identify members of an acoustic continuum of syllables varying in the F2 transition that ranged from /be/ to /de/ to /ge/, instead of showing a gradual shift in responses, they showed abrupt shifts, shown schematically in Figure 1.2. This occurred despite the fact that there was an equivalent acoustic change at every step along the continuum. A second hallmark of categorical perception, also shown in Figure 1.2, is that discrimination was considerably worse for pairs of syllables labeled as the same syllable than for syllables labeled differently. An early interpretation of this pair of findings was that it indexed a special way of perceiving speech. According to the motor theory of speech perception, listeners do not perceive the acoustic signal, but rather the articulatory gestures that produced the signal. Categorically distinct vocal tract gestures produce /b/, /d/, and /g/. Accordingly, they are perceived categorically as well. Identification functions are sharp, by this early account, because continuum members with the lowest frequency second formant onsets are perceived as bilabial (on the left side of Figure 1.2). Eventually, a syllable is encountered that cannot have been produced by lip closure, and it and the next few syllables are perceived as alveolar; final syllables all must have been produced by the tongue body, and are perceived as velar. Discrimination is near chance within these categories, according to the account, because all category members are perceived as equally bilabial (or alveolar or velar). It is only when one stimulus, say, is perceived as bilabial and one as alveolar that

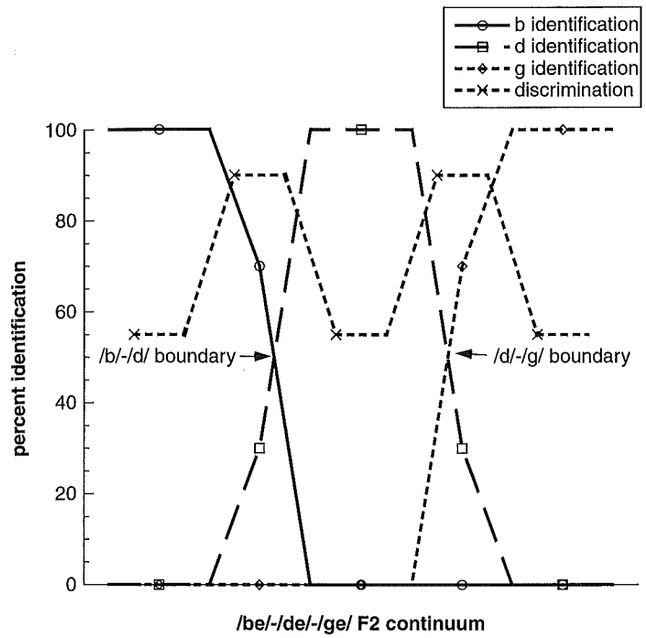

Figure 1.2. Schematic depiction of categorical perception findings. Identification functions are sharp, rather than gradual, and discrimination is poor within as compared to between consonant categories.

discrimination is possible. The categorical nature of speech perception has also been challenged by the findings considered next.

1.2.2 INTERNAL CATEGORY STRUCTURE

The claim (Studdert-Kennedy et al., 1970) that listeners to speech only discriminate syllables that they categorize differently was challenged early. Pisoni and Tash (1974) asked listeners to make same–different judgments of syllables along a /ba/ to /pa/ continuum. "Same" responses to acoustically different syllables were made with longer latencies than "same" responses to identical syllables. McMurray and colleagues have extended this finding by presenting subjects with word–word VOT continua (e.g., bear–pear) and a display of four pictures and asking subjects to click on the picture corresponding to the word they hear. The time course of lexical activation is estimated from eye movements subjects make as they hear the continuum items. Adults show gradient sensitivity within categories (that is, the farther the stimulus is from the category boundary, the faster they fixate the target picture; McMurray, Tanenhaus, and Aslin, 2002). Infants show similar gradient sensitivity in a head turn preference procedure (McMurray and Aslin, 2005). Accordingly, at least briefly, differences are perceived among syllables ultimately identified as the same syllable.

In fact, they are not perceived only briefly. Miller and colleagues (e.g., Miller and Volaitis, 1989; Allen and Miller, 2001) have shown that listeners give differential goodness ratings to syllables along, for example, a /b/ to /p/ continuum in an unspeeded task.

Kuhl has shown more about internal category structure. Listeners discriminate differentially within a category (Kuhl, 1991; Kuhl and Iverson 1995; but see Lively and Pisoni, 1997). They discriminate stimuli from the best category exemplar more poorly than from a poor category exemplar. Because

categories are language-specific, this suggests that a kind of warping of perceptual space occurs in the course of language learning.

1.2.3 DUPLEX PERCEPTION

When all of a syllable that is ambiguous between /da/ and /ga/ is presented to the left ear, and the disambiguating third formant transition is presented to the right ear, listeners hear two things at once (e.g., Mann and Liberman, 1983). They hear /da/ or /ga/ depending on which third formant transition has been presented, and they hear the transition as such, as a chirp that either rises or falls in pitch and that is distinct from the phonetic percept. Mann and Liberman interpreted this as showing that there are two auditory perceptual systems. Otherwise how could the same third formant transition be heard in two ways at the same time? One perceptual system renders a phonetic percept of /d/ or /g/. The other hears the transition literally as a fall or rise in pitch. This interpretation has been challenged, but not entirely successfully, by showing that perception of slamming doors can meet most, but not all, criteria for duplexity (Fowler and Rosenblum, 1990). If slamming door parts can be perceived in two ways at the same time, it cannot be because two perceptual systems, a door-perceiving system and the auditory system, underlie the percepts.

1.2.4 PARSING

Listeners behave as if they are sensitive to coarticulatory information in a speech signal. For example, in a classic finding by Mann (1980), listeners identified more syllables along a /da/ to /ga/ continuum as /da/ in the context of a precursor /ar/ than /al/ syllable. The pharyngeal tongue gesture of /r/ should pull the alveolar gesture of /d/ back, and listeners behave as if they recognize that. For intermediate syllables along the continuum, they behave as if they understand why the syllables are acoustically too far back for /d/ – coarticulation with the /r/ pulled the place of articulation of /d/ back. These findings show that listeners parse the coarticulatory effects of /r/ from acoustic information for /d/, and recent evidence shows that the parsed information is used as information for the coarticulating segment (e.g., Fowler, 2006). This pair of findings suggests close tracking by listeners of what talkers do.

There are many compatible findings. Listeners compensate for carryover (that is, left-to-right) coarticulation, for example, in the research by Mann (1980). They also compensate for anticipatory coarticulation (e.g., Mann and Repp, 1980). And they compensate for coarticulation that is not directional. For example, different speech gestures have converging effects on fundamental frequency (Fo). Other things equal, high vowels, such as /i/, have higher Fo than low vowels such as /a/, a phenomenon known as *intrinsic Fo*. Another use of Fo is to realize intonational accents. In research by Silverman (1987), two intonational accents, one on a high vowel and one on a low vowel, sounded equal in pitch when the accent on /i/ was higher in Fo than that on /a/. Listeners parse Fo that they ascribe to intrinsic Fo from Fo that they ascribe to an intonational accent. They do not ignore intrinsic Fo that they parse from an intonational accent. They use it as information for vowel height (Reinholt Peterson, 1986).

One interpretation of these findings is that listeners behave as if they are extracting information about speech gestures and are sensitive to acoustic effects of gestural overlap (Mann, 1980). In an /al/ context, listeners parse the /l/ coloring (fronting) that /l/ should cause from continuum members and hear more /ga/s. Parsing /r/ coloring (backing) leads them to hear more /da/s. However, another interpretation invokes a very general auditory process of spectral contrast (e.g., Lotto and Kluender, 1998). Again with respect to the Mann (1980) example, /al/ has a high ending F_3 that is higher in frequency than the onset F_3 of all members of the /da/ to /ga/ continuum. An /ar/ has a very low F_3 that is lower in frequency than the onset of F_3 of all continuum members. If a high-frequency sound exerts a contrastive effect, it makes following lower frequencies sound even lower than they are. This makes continuum members sound more /ga/-like. A low frequency ending F_3 should have the opposite effect, making continuum members sound more /da/-like.

1.2.5 MULTIMODAL SPEECH

As noted previously, perceivers of speech use more than acoustic information if they encounter it. They perceive speech in noise better if they can see the face of a speaker than if they cannot (Sumby and Pollack, 1954). Moreover, they integrate acoustic and optical speech information in the McGurk effect (McGurk and MacDonald, 1976). If haptic information replaces optical information, a McGurk effect also occurs (Fowler and Dekle, 1991). These effects may occur because listeners hear and see speech gestures. They integrate across the modalities, because the gestures specified by the two modalities of information should be from the same speech event. Alternatively (excepting perhaps the findings of Fowler and Dekle, 1991), the effects may occur because listeners/observers have a lifetime of experience both seeing and hearing speech, so they know what it looks like when various acoustic speech signals are produced. Seeing the speaking face, then, helps them to identify what was said.

Remarkably, Munhall and colleagues (Munhall et al., 2004) have shown that perceivers can extract phonetic information from the visible head movements of a speaker, such that speech is more intelligible in noise when natural head movements as well as facial phonetic gestures are visible to a speaker. Perceivers of speech are information omnivores. This finding awaits interpretation from either a gestural or an auditory theoretical perspective.

1.2.6 PERCEPTION OF SPEECH BY ANIMALS; PERCEPTION OF NONSPEECH BY HUMANS

Two lines of research, often conducted with the intention of challenging either or both of two theoretical claims about speech perception, are to test perception of speech by nonhuman animals (including monkeys, chinchillas, and birds), and to test perception of nonspeech by humans. The challenges are to the motor theory of speech perception, which claims that speech perception is accomplished by a specialization of the brain for phonetic perception, and both the motor theory and direct realism, which claim that listeners perceive vocal tract gestures. The logic of both lines of research is the same. It is that, if response patterns to speech by nonhuman animals or to nonspeech by humans are qualitatively like those to speech by humans, then most likely, perceptual processes applied to the signals are the same. If response patterns are the same, then, both a specialization for phonetic perception by humans and perception of gestures can be ruled out.

There are striking findings in both domains. For example, chinchillas have been found to have boundaries along a voice-onset time continuum similar to those of English listeners, including a shift in the boundary to longer values for farther back places of articulation (Kuhl and Miller, 1978). Japanese quail show compensation for coarticulation like that shown for Mann's (1980) human listeners described earlier (Kluender, Lotto, and Holt, 1997).

As for nonspeech, Lotto and Kluender (1998) showed an effect qualitatively like that of Mann (1980) when precursor syllables /ar/ and /al/ were replaced by tones at the ending F_3s of those syllables. They interpreted their finding as evidence of a contrast effect. The high tones caused the onset F_3s of members of the /da/–/ga/ continuum to sound effectively lower and so more /ga/-like; the low tones caused onset F_3s to sound higher and more /da/-like. There are many like findings (Holt, Lotto, and Kluender, 2000; Holt et al., 2001; Holt and Lotto, 2002). More recently Stephens and Holt (2003) showed a complementary effect. In their research, /al/ and /ar/ syllables were followed by tones tracking F_2 and F_3 transitions of a /da/ to /ga/ continuum. Listeners performed an AX discrimination task on syllable-transition pairs designed either to enhance or diminish discrimination if precursor syllables affected perception of the transitions. On enhanced trials, /al/ syllables, with a high ending F_3, preceded low frequency F_3 transitions, whereas /ar/ syllables, with a low ending F_3,

preceded high frequency F_3 transitions. On diminished trials the pairing was opposite. As predicted, discrimination performance on enhanced trials exceeded that on diminished trials, demonstrating a contrastive influence of the speech syllables on the nonspeech transitions.

How is this collection of findings to be interpreted? One way is that offered earlier. The findings disconfirm either or both of the claims of the motor theory and the central claim of the direct realist theory. No specialization for speech is required to explain basic speech perception findings. Moreover, animals are not supposed to perceive phonetic gestures; accordingly, when, for example, they show compensation for coarticulation qualitatively like humans do, these findings imply that humans do not perceive gestures. Further, in this view, when nonspeech signals exert context effects qualitatively like context effects exerted by speech signals, one can rule out gesture perception from nonspeech signals, and by implication, from speech signals.

These interpretations can be challenged and have been (e.g., Fowler, 1990). First, they are based on the weak logic that qualitative similarity in behavior implies identity of underlying perceptual processing. Fowler (1990) showed that behavioral responses to the sounds of ball bearings traversing two different kinds of ramps were qualitatively like, in the case of one ramp type, and qualitatively opposite, in the case of the other, to speech-rate normalization responses found by Miller and Liberman (1970) to syllables identified as /ba/ or /wa/. It is unlikely that processing of either sound type was like that for perception of /ba/ and /wa/.

As for the findings with nonhuman listeners, Best (1995) remarks that they do not disconfirm direct realism. Nonhuman animals also perceive sounding events directly, in terms of the forces structuring the acoustics. Nothing in the theory requires perfect perception, or that an organism have a particularly deep understanding of the specific physical objects and forces structuring the environment (consider, for example, our ability to identify general characteristics that cause the acoustic patterns that accompany bursting bubbles of various sorts of liquids). Therefore, when the speech categories tested involve rather large changes in causes (e.g., in place of articulation, or voicing) it is not surprising that nonhuman animals can perceive (or learn to perceive) the distinction. More crucial cases would involve categories that involve more subtle distinctions, and especially those that require substantial experience to acquire even by humans.

As for findings comparing perception of speech to nonspeech, a recent finding (Holt, 2005) suggests another perspective on the large majority of findings that are contrastive in direction. In Holt's experiments, she preceded members of a /da/ to /ga/ continuum with "acoustic histories." The histories in her first experiment consisted of twenty-one tones. The final tone in each history had a frequency of 2300 Hz. That tone and its twenty predecessors had an average frequency of 1800 Hz, 2300 Hz, or 2800 Hz. Holt found that the acoustic histories had a contrastive impact on /da/–/ga/ identifications even when as much as 1,300 ms or thirteen repetitions of the 2300 Hz tones intervened between the histories and the syllables. Histories with a higher average frequency were associated with more /ga/ responses than those with a lower average.

Holt's results support an interpretation that Warren (1985) offered for the ubiquity of contrast effects. Warren proposed that perceivers continuously update perceptual criteria to calibrate them to recently encountered stimuli. For example, perceivers who find themselves in a setting in which they encounter many high-pitched sounds effectively recalibrate what counts as high-pitched, judging as lower in pitch sounds that in other contexts they would judge high. This account invokes higher-level, cognitive sources of influence on speech processing than either auditory or gestural accounts of speech perception have anticipated.

1.3 *Variability, normalization, and phonetic constancy*

Speech is characterized by variability. The production and acoustic realization of speech sounds depend on context – segmental (Liberman et al., 1952), prosodic (Fougeron and Keating, 1997), discourse (old/new: Fowler and Housum, 1987; Fowler, Levy, and Brown, 1997; Nooteboom and Kruyt, 1987), the physical characteristics of talkers (Dorman, Studdert-Kennedy, and Raphael, 1977; Peterson and Barney, 1952), speaking rate (Miller, 1981), and acoustic environment. Phonetic perception is resilient to all of these, and to many other perturbations of speech including novel accents and wide ranges of acoustic environments and conditions (parking garages, anechoic chambers, listening to a child speak with her mouth full).

This *phonetic constancy* despite a *lack of invariance* (the many-to-many mapping between acoustics and phonetic categories) defies easy explanation and poses significant challenges to all theories of speech perception. The flip side to variability, though, is that speech is rich with information beyond phonetic cues.

For example, the listener can glean a tremendous amount of *indexical* information from the speech signal (Ladefoged and Broadbent, 1957), including talker identity (Van Lancker, Kreiman, and Emmorey, 1985), physical characteristics (e.g., sex, age, height, and weight; Krauss, Freyberg, and Morsella, 2002), socioeconomic status (Ellis, 1967), and emotional state (e.g., Murray and Arnott, 1993; Streeter et al., 1983). Some of these qualities obviously co-vary with the elements of the speech signal assumed to carry phonetic information (e.g., differences in dialect specifying the realization of *pen* as /pEn/ or /pIn/; see Niedzielski, 1999 for evidence that *expectations* about dialect influence vowel perception), while others may have greater or lesser effect on phonetic information. An increase in Fo or amplitude in an arousal state has less impact on the realization of phonetic categories than do changes in speaking rate that may accompany an arousal state (e.g., Streeter et al., 1983). Similarly, changes in voice quality associated with aging within adulthood (e.g., Caruso, Mueller, and Shadden, 1995) may have little impact on the realization of speech sounds.

So variability is a double-edged sword: It creates a complex mapping between acoustics and percepts, but carries a tremendous amount of important information, and indexical information in particular. For several decades, there were two strands of talker variability research. Research focused on phonetic information largely took a negative view of variability, and sought to avoid or eliminate it by finding invariant cues or finding a way to transform speech signals to make the acoustic-phonetic mapping invariant. Research focused on indexical variability viewed variability as a rich source of useful information, and largely ignored phonetic perception.

Recently, this separation has been reconsidered in light of evidence that phonetic and indexical variability is preserved in memory and influences perception as well as memory of words. Exemplar and nonanalytic theories motivated by these findings present the possibility that subcategorical phonetic information and even nonlinguistic variability in the speech signal may help solve the lack of invariance problem in understanding phonetic constancy. We will briefly review the two major conventional approaches to phonetic constancy (the search for invariant cues and normalization theories) and then discuss the potential of exemplar theories.

1.4 *Finding invariants*

One possibility is that a larger window of analysis would provide a less variable mapping from acoustics to phonetic categories. Candidates have included syllables (Massaro, 1972), overlapping context-sensitive units (Wickelgren, 1969), or whole phrases (Miller, 1962), but none of these result in an invariant mapping.

Another possibility is that the speech signal contains invariant cues to phonetic categories, but we have not discovered them yet. For example, some evidence suggested context-invariant cues might be found by integrating across time scales not captured in typical acoustical analyses (Kewley-Port, 1983; Stevens and Blumstein, 1978). However, candidate cues have thus

far turned out not to provide an invariant mapping. Furthermore, when both candidate invariant and variable cues are present, listeners focus attention on the variable cues (Blumstein, Isaacs, and Mertus, 1982; Walley and Carrell, 1983).

A third possibility is that articulatory actions provide the basis for phonetic categorization. Although the mapping from articulators to vocal tract shapes is indeterminate (e.g., Gopinath and Sondhi, 1970), it does not follow that phonetic gestures cannot be recovered from acoustic signals by listeners. Motor theorists invoke innate mechanisms, special to speech, that provide knowledge (about coarticulation's acoustic effects) that may permit selection among the possible vocal tract configurations consistent with an acoustic signal. However, there is a more fundamental reason why gestures may be recovered from indeterminate acoustic signals: Proofs of indeterminacy are only proofs that not *every* detail of vocal tract configurations is recoverable. However, phonetic gestures are proposed to be coarse-grained vocal tract actions that create and release constrictions in equifinal ways. For example, lip closures for /b/, /p/, and /m/ are achieved by a variety of patterns of jaw and lip movement and, because tongue configuration is irrelevant to that gesture, with a variety of tongue configurations. To detect the lip gesture, listeners only have to detect that the lips have closed. The precise lip, jaw, and tongue configurations need not be recovered.

In the case of variability, then, direct realism presents a promissory note along the following lines. The acoustics researchers extract from speech (spectra, formants, etc.) do not capture the underlying causal structure listeners perceive. The similarity of different tokens of the same gesture is greater than that of the acoustics as we currently analyze them, such that the same lawful relations hold between two speakers' gestures (distal stimuli) and the resulting acoustic structure (proximal stimuli). The heavy lifting of identifying the precise structure remains, but making a distinction between perfect recovery versus good-enough, coarse-grained recovery of articulation may lead to progress.

It may not be obvious how this approach can handle the sensitivity of listeners to far more in the signal than constriction locations and degrees (such as indexical information and speaking rate). Some indexical information (e.g., differences in vocal tract morphology) may be specified by the acoustic information for vocal tract shape just considered. However, much indexical information as well as information about rate is expected to be specified by acoustic information for vocal tract dynamics.[2]

1.5 *Normalization*

A more common approach for addressing the variability in speech signals is to suppose that phonetic categorization depends on auditory, perceptual, and/or cognitive transformations of speech signals that map them onto speaker- and rate-independent categories. This sort of approach dates at least to Joos (1948), who proposed that either the incoming speech or internal phonetic categories must be warped to bring the two into registration when a talker change is detected. In addition to the descriptive problem that normalization addresses – the apparent lack of invariance – there is behavioral evidence suggesting a perceptual cost to a change in talker. Mullennix, Pisoni, and Martin (1989) found that identification of words presented in noise was less accurate and slower when words in a series of trials were produced by multiple talkers compared to when blocks of trials were produced by a single talker. Nusbaum and Morin (1992) compared phoneme, syllable, and word monitoring in blocked and mixed talker conditions, and found a constant talker variability cost of about thirty ms. The descriptive lack of acoustic–phonetic invariance and perceptual costs motivate *talker normalization theories.*

Normalization theories typically focus on vowels (although similar problems

2 We thank Gordon Ramsay for a tutorial on acoustic–articulation indeterminacy.

apply to consonants; Dorman et al., 1977), and hold either that the basis for phonetic constancy comes from *stimulus intrinsic* or *stimulus extrinsic* information (Ainsworth, 1975; Nearey, 1989). Stimulus intrinsic theories hold that each sample of speech (e.g., any vowel) contains sufficient information for accurate classification. For example, Syrdal and Gopal (1986) proposed a normalization algorithm in which Fo and F3 are used to rescale F1 and F2. This works fairly well but does not result in an invariant mapping (see J. D. Miller, 1989 for a review of formant ratio approaches). Shankweiler, Strange, and Verbrugge (1977) hypothesized that dynamic information from consonant–vowel transitions provide talker-independent phonetic cues; even when vowel nuclei were excised and syllabic onsets and offsets from talkers of different genders were paired, subjects were able to identify vowels with high accuracy (but not perfectly).

However, listeners do not rely solely on information in individual samples. Instead, vowel perception depends on recent context. In the classic experiment by Ladefoged and Broadbent (1957), identical /b/-vowel-/t/ words were identified by listeners following context sentences produced by synthetic speech representing distinct vowel spaces. Perception of a target word depended on the vowel space specified in the context phrase (e.g., with an identical utterance identified as *bit* following one talker and as *bet* following another).[3] Such findings motivate *stimulus extrinsic* theories, in which the listener builds up a talker-specific mapping over multiple samples of speech (possibly initially with mechanisms like those proposed under stimulus intrinsic accounts). Joos (1948) speculated that listeners build up a map of a talker's vowel space from the first few words they hear, gradually improving the mapping as lexical cues (e.g., in conventional greetings) disambiguate acoustic patterns. Algorithmic approaches in this vein (e.g., classification relative to point vowel measures; Gerstman, 1968) proved quite accurate for isolated samples of speech.

But passive mechanisms such as Gerstman's fail to account for details of behavioral and neural evidence. For example, Nusbaum and Morin (1992) found that the effect of talker variability (blocked versus mixed) in their monitoring tasks interacted with that of cognitive load, a hallmark of active processing (Schneider and Shiffrin, 1977). Wong, Nusbaum, and Small (2004) examined neural activity using fMRI with the blocked/mixed monitoring paradigm. In the mixed talker condition, they found increases in activity in areas associated with speech (posterior superior temporal gyrus) and areas associated with shifts of attentional processing (superior parietal cortex). Magnuson and Nusbaum (2007; see also Nusbaum and Magnuson, 1997) used the monitoring paradigm with two sets of synthetic speech that differed slightly in average Fo (with some small collateral changes in vowel space). They told one group of subjects they were hearing two talkers and another group that they were hearing one talker; a third group received no special instructions. The latter two groups showed no difference between blocked and mixed conditions, while the group expecting two talkers showed the effect typical of mixing two genuinely different talkers – they were significantly slower in the mixed talker condition. Similarly, Johnson, Strand, and D'Imperio (1999) used instructions to manipulate expectations about talker sex, which caused a shift in vowel category boundaries like those caused by manipulations of Fo or visual cues to sex. These results suggest that adapting to changes in talker characteristics is an active, resource demanding process subject to cognitive penetration. Nusbaum and Morin (1992) proposed a variant of stimulus extrinsic normalization, *contextual tuning*, which assumes there are active control structures that retune phonetic mappings when talker changes are detected (or even just expected, under certain conditions).

3 For a demonstration, see https://engineering.purdue.edu/~malcolm/interval/1997-056/VowelQuality.html.

1.6 *Exemplar theories*

A radically different proposal is that talker differences do not require normalization – but not because there is invariant information for phonetic categories. Rather, *exemplar theories* (Goldinger, 1998; Johnson, 1997) propose that holistic exemplar traces of speech are stored in memory and form the basis for linguistic perception.[4] As in exemplar models of categorization more generally, the theory is that the exemplars statistically sample the space of talker characteristics (including putatively linguistic and nonlinguistic dimensions). The argument is that phonetic constancy has been misconstrued. The acoustic correlates of talker variability are not random, and this lawful variability (cf. Elman and McClelland, 1986) may constrain rather than complicate the problem of phonetic constancy (and vice versa; cf. Remez et al., 1997). Evidence for this view includes findings that recognition memory improves when talker characteristics are preserved across new and old items[5] (e.g., Goldinger, 1996) as does performance in processing tasks like shadowing (Goldinger, 1998).

Goldinger (1998) presented promising simulations with a computational implementation of an exemplar model in which phonetic categorization was accomplished without explicit normalization. However, the inputs to the model were presegmented (exemplars were words) and prenormalized – separate units represented talker and phonetic information. Further tests of the model with more realistic input (in which phonetic information is conditioned by talker characteristics, as in real speech) are needed to evaluate the claim that exemplar models do not require normalization. As Goldinger and Azuma (2003) discuss, *adaptive resonance theory* (e.g., Grossberg, 2003) may provide a framework capable of handling more realistic inputs while incorporating the key principles of Goldinger's *episodic lexicon theory* (see Samuel and Sumner, this volume, for more discussion of the theory).

There is also evidence for *contingency* of linguistic and indexical variability. Task-irrelevant variability in either dimension slows processing (Mullennix and Pisoni, 1990), and training on talker identification facilitates phonetic perception of speech produced by trained-on talkers (Nygaard and Pisoni, 1998; Nygaard, Sommers, and Pisoni, 1994), suggesting that perceptual learning occurs for both types of information no matter which is attended, or, more strongly, that the two dimensions are perceived integrally and nonanalytically – that is, as holistic exemplar traces (Nygaard and Pisoni, 1998).

However, contingency is not always found. Luce and Lyons (1998) found specificity effects in explicit recognition, but not in an implicit priming task. McLennan and Luce (2005) found that specificity effects depend on processing time. When listeners are forced to respond quickly, the advantage for preserving nonlinguistic characteristics of a spoken word disappear. These results challenge strong integrality, but are consistent with the parallel-contingent explanation of Mullennix and Pisoni (1990), who argued that phonetic identification is hierarchically dependent on analyses of talker characteristics.

1.7 *Summary*

Auditory/cognitive accounts assume that phonetic constancy depends on finding one of the following: invariant acoustic cues to phonetic categories; transformations of the speech signal that render talker- (or other context-) specific mappings; or active control structures that, for example, hypothesize mappings and refine them based on success in mapping to phonetic, lexical, or even more complex categories. The motor theory and direct realist theory assume gestures as fundamental units (inferred in the former, directly perceived in the latter based on coarse-grained acoustic information for constriction locations and degrees). *Episodic/exemplar/nonanalytic* theories predict that holistic memory traces of speech will provide a good enough basis for classification to approximate an invariant mapping, and are the only theories to account for specificity effects (e.g., benefits to preserving nonphonetic details of utterances). Each of these approaches continues to hold promise for explaining phonetic constancy. However, none of them can yet specify how information in the speech signal is mapped to phonetic categories in order to achieve phonetic constancy despite variation in talkers, rate, acoustic environment, and other contextual dimensions.

2 Speech perception and lexical access

Psycholinguists tend to classify speech perception and spoken word recognition (and other levels of description) as distinct aspects of spoken language understanding, with most theories postulating distinct levels of representation for each division. Even in theories that posit distinct phonetic-perceptual and lexical levels, the interface between these levels is of great importance. We will focus on three key issues: the lexical segmentation problem, interface representations, and the modularity/interaction debate.

2.1 *Segmentation and interface representations*

The *lexical segmentation problem* (in contrast to the phonemic segmentation problem mentioned earlier) is that fluent speech provides few acoustic cues to word boundaries, and none of those are very reliable (Cole and Jakimik, 1980; Saffran and Sahni, this volume). This fact has helped justify the simplifying assumption that a speech perception mechanism provides a phoneme readout of the speech signal. Thus, phonemes form the interface between perception and a presumably more cognitive word recognition system. This interface assumption simplifies in two directions: Speech perception is given a concrete goal, and word recognition is freed from, for example, the lack of invariance problem. However, it is also a potentially complicating assumption.

Consider the *embedding problem* in spoken word recognition. Assuming that the bottom-up input is a string of phonemes with no cues to word boundaries poses a tremendous parsing challenge. McQueen et al. (1995) estimated that eighty-four percent of polysyllabic words have at least one shorter word embedded within them, and many have more than one (*catalog* has *cat*, *cattle*, *at*, *a*, and *log*) and most short words are embedded within longer words. The problem is compounded by embeddings that straddle word boundaries and phonological processes that create additional ambiguities (e.g., in fluent speech, phonological assimilation can make the realization of *right berry* nearly identical to that of *ripe berry*; Gaskell and Marslen-Wilson, 1996; Gow, 2001). Theories have typically relied on lexical competition to solve the segmentation and embedding problems (e.g., McClelland and Elman, 1986; Norris, 1994).

However, the embedding problem is significantly reduced (but does not go away completely) if we assume that prosodic and subphonemic information is available for word recognition. For example, employing

4 Goldinger (1998) claimed his simulations with an exemplar model proved that an exemplar mechanism could solve the lack of invariance problem without normalization. However, the input to his model consisted of presegmented word-length samples of separate units representing phonetic and indexical characteristics; it was prenormalized. In real speech, phonetic information is conditioned by talker characteristics. The episodic model may provide phonetic constancy without normalization, but tests with closer analogs to real speech (or real speech) are needed to evaluate the claim.

5 One argument proponents of exemplar models sometimes make is that the preservation of surface detail falsifies normalization theories, which they claim require that all information aside from abstract phonetic categories be discarded. There are two problems with this claim. First, normalization does not require that subcategorical detail be discarded, and few accounts of normalization explicitly or implicitly discard it. All it entails is bringing the signal and phonetic categories into registration. Second, preservation of surface detail would not falsify an abstractionist account of normalization, as it would be consistent with multiple parallel representations of different aspects of the signal at varying levels of abstraction, as is found in the visual and auditory systems more generally (Bushara et al., 1999; Mishkin, Ungerleider, and Macko, 1983). Indeed, Belin, Zatorre, Lafaille, Ahad, and Pike (2000) present evidence supporting their model on which a dorsal pathway is specialized for extracting linguistic content, while a ventral pathway is specialized for talker identification.

the *metrical segmentation strategy* (Cutler and Norris, 1988) – positing word boundaries before all strong syllables – would correctly segment about ninety percent of (frequency-weighted) words (Cutler and Carter, 1987), but would not completely resolve the segmentation and embedding problems. However, other cues are available. Building on evidence that subcategorical mismatches (misleading coarticulatory cues introduced by splicing parts of two words together, for example, splicing the stop burst of /b/ in *job* onto the *jo* of *jog* so that formant transitions at the end of the vowel signal a forthcoming velar, rather than bilabial stop) influence lexical access (Dahan et al., 2001; Marslen-Wilson and Warren, 1994), Salverda, Dahan, and McQueen (2003) used eyetracking to measure the time course of lexical activation when the first syllable of a word like *hamster* was replaced by a production of an embedded word like *ham* (which was about fifteen ms longer in duration than the *ham* of *hamster*). Not only were listeners sensitive to the manipulation (as revealed by initial biases to look to *ham* rather than *hamster*), they showed differential sensitivity depending on the prosodic context in which the short word was recorded (see Davis, Marslen-Wilson, and Gaskell, 2002 for converging offline evidence from priming). Thus, there is growing evidence that much more detail is available at lexical access than categorical phonemic abstractions (though whether such detail makes direct contact with the lexicon [Gaskell and Marslen-Wilson, 1997] or is mediated by a phonemic or gestural level of representation capable of delivering continuous levels of activation, as in TRACE [McClelland and Elman, 1986; Dahan et al., 2001] or direct realism [Fowler, 1986] remains an open question).

2.2 *Interaction or autonomy?*

A central debate in speech perception and spoken word recognition is whether the influence is bi-directional – whether there is *feedback interaction* between higher and lower levels of representation, or whether representational organization is strictly *modular* (or *autonomous*).

It is clear that sublexical decisions can be influenced by lexical knowledge. For example, Ganong (1980) spliced identical /t/ to /d/ continua onto following contexts "-ash" and "-ask," creating nonword–word and word–nonword continua, respectively. Lexical status shifted the identification function, such that more steps on each continuum were identified as the stop consistent with a word. Another well-known example is *phoneme restoration*, in which listeners report that they hear a segment that has been completely replaced by noise (e.g., Warren, 1970).

The interpretation of these effects is controversial. They are easily explained by lexical-to-phoneme feedback in an interactive model like TRACE (McCelland and Elman, 1986). But others argue that all effects consistent with feedback could occur prelexically (based on, e.g., transitional probability information stored at a phonemic level) or at a postlexical decision stage in a feedforward, autonomous system (Norris, McQueen, and Cutler, 2000). What is needed to distinguish between the two accounts is evidence for a lexical influence on perception that cannot be attributed to postlexical decisions about phonemes. An ideal task would demonstrate sublexical, perceptual *consequences* of lexical information, ideally after the lexical context has been heard (as in lexically mediated compensation for coarticulation; Elman and McClelland, 1988; Magnuson et al., 2003; Pitt and McQueen, 1998; Samuel and Pitt, 2003), or would not require an explicit phonemic decision (e.g., selective adaptation based on restored phonemes; Samuel, 2001). However, repeated demonstrations of lexical influence on phonemes in these paradigms have not convinced everyone concerned (see Norris, McQueen, and Cutler, 2003, for thoughtful arguments and discussion, and Magnuson, Mirman, and Harris, this volume, for further theoretical and computational details). Thus, there are strong arguments on both sides, but the jury is still out on the empirical evidence.

3 Avenues to progress

We have selectively reviewed the key phenomena and theoretical issues in speech perception. We will close with a discussion of two questions for future research we see as vital for advancing our understanding of speech perception. The issues we have chosen reflect the two main sections of our chapter.

3.1 *What is the basis for phonetic perception?*

The key debate here is between general auditory and gesture theories. Can evidence distinguish the theories in convincing ways? There are some barriers to this. First, neither account is sufficiently developed in critical domains for evidence to make a clear distinction. Theorists in both domains have considerable wiggle room that they will use at least until the accounts are further sharpened or until a superior alternative to both appears. The following, then, is a guess about the kinds of evidence that theorists will seek to confirm or disconfirm either account.

For gesture accounts to be viable, it will be necessary for researchers to address convincingly the proofs of indeterminacy in the mapping from acoustics to articulation. A promising avenue may be the one outlined earlier. The proofs are that not every aspect of vocal tract configuration can be recovered from its acoustic correlates. However, gesture theories do not claim that listeners recover that level of detail. Rather, listeners recover information about gestures, coarse-grained actions that create and release constrictions. Needed now is evidence whether, for the gestures that theorists identify as linguistically relevant across languages of the world, acoustic information can specify them. Proofs that even that level of detail cannot be recovered would disconfirm the direct realist theory, but not necessarily the motor theory, which invokes innate knowledge of coarticulation and its acoustic consequences to assist in gesture recovery.

Neuropsychological evidence that would disconfirm auditory accounts and correspondingly confirm the motor theory would be convincing evidence for a specialization of the brain for speech perception. Such a specialized system should not activate when acoustic signals similar to speech, but not perceived to have phonetic properties, are presented to listeners. The specialized system should also activate when speech is produced or silently mouthed. It might be expected to activate when speech gestures are seen as well as heard. Neuropsychological evidence that would confirm auditory accounts would be evidence that brain areas active for speech signals are also active for acoustically similar nonspeech signals.

As for behavioral evidence, unfortunately it is not obvious what strategies not already adopted would be informative. Strategies already adopted have so far led to inconclusive outcomes.

Direct realists, but not motor theorists or auditory theorists, predict that when nonspeech signals have been caused by real-sounding events (as opposed to being, for example, tones) behavioral response patterns to them should resemble those to speech when distal properties of the speech and nonspeech events are similar in relevant ways. When they are different in relevant ways, response patterns should differ. These predictions hold, in principle, regardless of similarity in acoustic structure. For auditory theorists, the predictions are opposite. Response patterns should be similar when acoustic patterns are similar in relevant ways regardless of the similarity of distal event properties. This distinction in predictions is sharp, but likely difficult to cash out experimentally. Similar event properties are likely most of the time (but not always) to generate similar acoustic patterns.

3.2 *What is the basis for phonetic constancy?*

We described three major approaches to phonetic constancy: auditory/cognitive,

gestural invariance, and episodic/exemplar/nonanalytic. None of the approaches provides a satisfactory answer. Truly invariant acoustic cues to phonetic categories have yet to be identified. Passive normalization algorithms proposed in auditory/cognitive approaches are not capable of human levels of accuracy, nor can they account for evidence for active control in talker adaptation (and note that while talker characteristics have been our focus, rate variability poses similar challenges). Contextual tuning explains evidence for active control (Nusbaum and Magnuson, 1997) and may provide a unitary account of talker and rate adaptation, but as yet, concrete proposals for mechanisms are lacking.

A full account of gestural invariance (the basis for phonetic constancy in motor and direct realist theories) will require identifying the characteristics of the proximal stimulus (acoustics) that map lawfully onto the underlying distal causes (gestures). The possibility that coarse-grained gestural information good enough for phonetic categorization is carried in the acoustics of speech acoustic is promising, but this approach is in very early stages and is thus untested.

Episodic/exemplar/nonanalytic models (Goldinger, 1998) parsimoniously account for effects of linguistic and nonlinguistic variability, and hold promise for the phonetic constancy problem, but a truly convincing theory would require strong support for the claim that exemplar models do not require normalization, including tests with (more) realistic inputs than Goldinger used in his 1998 simulations, and an account of why the impact of linguistic and nonlinguistic information can be dissociated (McLennan and Luce, 2005). As Goldinger and Azuma (2003) discuss, adaptive resonance theory (Grossberg, 2003) may provide a unifying framework that incorporates the key theoretical principles of the exemplar approach as well as mechanisms suited to addressing the challenges of real speech. We are eager to see whether the promise of this approach can be realized.

The truth likely lies somewhere between the theories we have discussed. As Remez (2005) puts it, after reviewing evidence that phonetic perception is resilient against the removal or perturbation of virtually any hypothetical cue to phonetic perception, "perceptual organization [is] attuned to a complex form of regular if unpredictable spectrotemporal variation within which the specific acoustic and auditory elements matter far less than the overall configuration they compose" (2005, pp. 42–3).

References

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In Fant, G. & Tatham M. (Eds.) *Auditory analysis and perception of speech* (pp. 103–13). London: Academic Press.

Allen, J. S. & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, 63, 798–810.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403, 309–12.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In Strange, W. (Ed.) *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 167–200). Timonium MD: York Press.

Blumstein, S. E., Isaacs, E., & Mertus, J. (1982). The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 72, 43–50.

Bushara, K. O., Weeks, R. A., Ishii, K., Catalan, M., Tian, B., & Hallett, M. (1999). Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. *Nature Neuroscience*, 2, 759–66.

Caruso, A., Mueller, P., & Shadden, B. B. (1995). Effects of aging on speech and voice. *Physical and Occupational Therapy in Geriatrics*, 13, 63–80.

Cole, R. & Jakimik, J. (1980). A model of speech perception. In Cole, R. (Ed.) *Perception and Production of Fluent Speech* (pp. 133–63). Hillsdale NJ: Erlbaum.

Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133–42.

Cutler, A. & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–21.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 16, 507–34.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 218–44.

Diehl, R. & Kluender, K. (1989). On the objects of speech perception. *Ecological Psychology*, 1, 121–44.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, 22, 109–22.

Driver, J. (1996). Enhancement of selective listening of illusory mislocation of speech sounds due to lip-reading. *Nature*, 381, 66–8.

Ellis, D. S. (1967). Speech and social status in America. *Social Forces*, 45, 431–7.

Elman, J. & McClelland, J. (1988). Exploiting lawful variability in the speech wave. In Perkell, J. S. & Klatt, D. H. (Eds.) *Invariance and variability in speech processes* (pp. 360–80). Hillsdale, NJ: Erlbaum.

Fougeron, C. A. & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 101, 3728–40.

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

(1990). Sound-producing sources as objects of speech perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, 88, 1236–49.

(1994). The direct-realist theory of speech perception. *The Encyclopedia of Language and Linguistics* 8 (pp. 4199–203). Oxford: Pergamon Press.

(2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, 68, 161–77.

Fowler, C. A. & Dekle, D. J. (1991). Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 816–28.

Fowler, C. A. & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26, 489–50.

Fowler, C. A., Levy, E. T., & Brown, J. M. (1997). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language*, 37, 24–40.

Fowler, C. A. & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 742–54.

Fowler, C. A. & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, 36, 171–95.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, 13, 361–77.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, 6, 110–25.

Gaskell, M. G. & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, 22, 144–58.

Gaskell, M. G. (1997). Integrating form and meaning: A distributed model of speech perception. *Language & Cognitive Processes*, 12, 613–56.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio Electroacoustics, AU-16*, 78–80.

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

Gibson, J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 1166–83.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251–79.

Goldinger, S. D. & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, 31, 305–20.

Goldstein, L. & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language

use. In Schiller, N. & Meyer, A. (Eds.) *Phonetics and phonology in language comprehension and production: Differences and similarities.* (pp. 159–207) Berlin: Mouton de Gruyter.

Gopinath, B. & Sondhi, M. M. (1970). Determination of the shape of the human vocal tract from acoustical measurements. *Bell Systems Technical Journal, 49,* 1195–214.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics, 31,* 423–45.

Holt, L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16,* 305–12.

Holt, L. & Lotto, A. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research, 167,* 156–69.

Holt, L., Lotto, A., & Kluender, K. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America, 108,* 710–22.

(2001). Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *Journal of the Acoustical Society of America, 109,* 764–74.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K. & Mullennix, J. W. (Eds.) *Talker Variability in Speech Processing* (pp. 145–66). San Diego: Academic Press.

Johnson, K., Strand, E. A., & D'imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 24,* 359–84.

Jones, S. (1996). Imitation or exploration: Young infants' matching of adults' oral gestures. *Child Development, 67,* 1952–69.

Joos, M. (1948). *Acoustic phonetics.* Baltimore, MD: Linguistic Society of America.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13,* 1709–14.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 322–35.

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

Kluender, K., Lotto, A., & Holt, L. (1997). Perceptual compensation for coarticulation by Japanese quail (coturnix coturnix japonica). *Journal of the Acoustical Society of America, 102,* 1134–40.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38,* 618–25.

Kuhl, P. (1991). Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50,* 93–107.

Kuhl, P. & Iverson, P. (1995). Linguistic experience and the perceptual magnet effect. In Strange, W. (Ed.) *Speech perception and linguistic experience, Issues in cross-language research.* (pp. 121–54) Baltimore, MD: York Press.

Kuhl, P. & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63,* 905–17.

Lachs, L. (2002). *Vocal tract kinematics and cross modal speech information.* Ph.D Dissertation, Indiana University.

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

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29,* 117–23.

(1996). *Speech, A special code.* Cambridge, MA: Bradford Books.

Liberman, A., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. G. (1967). Perception of the speech code. *Psychological Review, 74,* 431–61.

Liberman, A., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology, 65,* 497–516.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358–68.

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

Liberman, A. M. & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4,* 187–96.

Lively, S. & Pisoni, D. B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1665–79.

Lotto, A. & Kluender, K. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602–19.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 26,* 708–15.

Maddieson, I. (1984). *Patterns of sounds.* Cambridge: Cambridge University Press.

Magnuson, J. S., Mirman, D., & Harris, H. D. (this volume). Computational models of spoken word recognition.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 391–409.

Magnuson, J., McMurray, B., Tanenhaus, M., & Aslin, R. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmas past. *Cognitive Science, 27,* 285–98.

Mann, V. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28,* 407–12.

Mann, V. &. Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition, 14,* 211–35.

Mann, V. & Repp, B. (1980). Influence of vocalic context on perception of the [s]-[s] distinction. *Perception & Psychophysics, 28,* 213–28.

Marslen-Wilson, W. & Warren, P. (1994). Levels of perceptual representations and process in lexical access: Words, phonemes, features. *Psychological Review, 101,* 653–75.

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, 79,* 124–45.

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

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–48.

McMurray, B. & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95,* B15-B26.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33-B42.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes, 10,* 309–31.

Miller, G. A. (1962). Decision units in the perception of speech. *IRE Transactions on Information Theory, IT-8,* 81–3.

Miller, J. L. (1981). Phonetic perception: Evidence for context-dependent and context-independent processing. *Journal of the Acoustical Society of America, 69,* 822–31.

Miller, J. L. & Volaitis, L. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46,* 505–12.

Mishkin, M., Ungerleider, L., & Macko, K. (1983). Object vision and spatial vision: Two central pathways. *Trends in Neuroscience, 6,* 414–17.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379–90.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365–78.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15,* 133–6.

Murray, I. R. & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, *Journal of the Acoustical Society of America, 93,* 1097–108.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–113.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language & Social Psychology, 18,* 62–85.

Nooteboom, S. G. & Kruyt, J. G. (1987). Accent, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America, 82,* 1512–24.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., McQueen, J. M, & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204–38.

Nusbaum, H. C. & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy

as a cognitive process. In Johnson, K. & Mullennix, J. W. (Eds.) *Talker Variability in Speech Processing* (pp. 109–32). San Diego: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Tohkura, Y., Sagisaka, Y., & Vatikiotis-Bateson, E. (Eds.) *Speech Perception, Speech Production, and Linguistic Structure*. (pp. 113–34). Tokyo: OHM.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42–6.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60, 355–76.

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, 24, 175–84.

Pierrehumbert, J. (1990). Phonological and phonetic representations. *Journal of Phonetics*, 18, 375–94.

Pisoni, D. B. & Tash, J. (1974). Reaction times to comparisons within and across phonetic boundaries. *Perception & Psychophysics*, 15, 285–90.

Pitt, M. A. & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory & Language*, 39, 347–70.

Reinholt Peterson, N. (1986). Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations. *Phonetica*, 43, 31–42.

Remez, R. E. (2005). The perceptual organization of speech. In Pisoni, D. B. & Remez, R. E. (Eds.) *The Handbook of Speech Perception*, (pp. 28–50). Oxford: Blackwell.

Remez, R., Fellowes, J., & Rubin. P. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 651–66.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science*, 212, 947–50.

Rosenblum, L. & Saldana, H. (1998). Time-varying information for speech perception. In Campbell, R., Dodd, B., & Burnham, D. (Eds.). *Hearing by eye II, Advances in the Psychology of speechreading and auditory-visual speech*. (pp.61–81) East Sussex, UK: Psychology Press.

Rosenblum, L. D., Yakel, D., Baseer, N., Panchal, A., Nodarse, B., & Niehus, R. (2002). Visual speech information for face recognition. *Perception & Psychophysics*, 64, 220–9.

Saffran, J. R. & Sahni, S. D. (this volume). Learning the sounds of language.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12, 348–51.

Samuel, A. G. & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language*, 48, 416–34.

Samuel, A. G. & Sumner, M. (this volume). Current directions in research on spoken word recognition.

Sawusch, J. & Gagnon, D. (1995). Auditory coding, cues and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 635–52.

Schneider, W. & Shiffrin, R. (1977). Controlled and automatic human information processing: 1. Detection, search and attention. *Psychological Review*, 84, 1–66.

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In Shaw, R. & Bransford, J. (Eds.) *Perceiving, acting, and knowing* (pp. 315–45). Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In Cooper, W. & Walker, E. (Eds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. (pp. 295–342) Hillsdale, NJ: Lawrence Erlbaum.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours*, Ph.D. Dissertation, Cambridge University.

Smith, Z., Delgutte, B., & Oxenham, A. (2002). Chimearic sounds reveal dichotomies in auditory perception. *Nature*, 416, 87–90.

Stephens, J. & Holt, L. (2003). Preceding phonetic context affects perception of non-speech sounds. *Journal of the Acoustical Society of America*, 114, 3036–9.

Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64, 1358–68.

Streeter, L. A., Macdonald, N. H., Apple, W., Krauss, R. M., & Galotti, K. M. (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, 73, 1354–60.

Studdert-Kennedy, M. & Goldstein, L. (2003). Launching language: The gestural origins of discrete infinity. In Christiansen, M. & Kirby, S. (Eds.) *Language evolution*. (pp. 235–54). Oxford: Oxford University Press.

Studdert-Kennedy, M. G., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). The motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, 77, 234–49.

Sumby, W. H. & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–15.

Syrdal, A. K. & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79, 1086–100.

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

Walley A. D. & Carrell T. D. (1983). Onset spectra and formant transition in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 73, 1011–22.

Warren, R. M. (1970). Restoration of missing speech sounds. *Science*, 167, 392–3.

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

(1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, 92, 574–84.

Wong, PCM, Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, 16, 1173–84.

Author Note

Preparation of this manuscript was supported by NICHD grant HD-01994 to Haskins Laboratories. We thank Philip Rubin, Robert Remez, and Yvonne Manning-Jones for making Figure 1 available to us from Haskins Laboratories' Website.