Section 1

SPEECH PERCEPTION

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

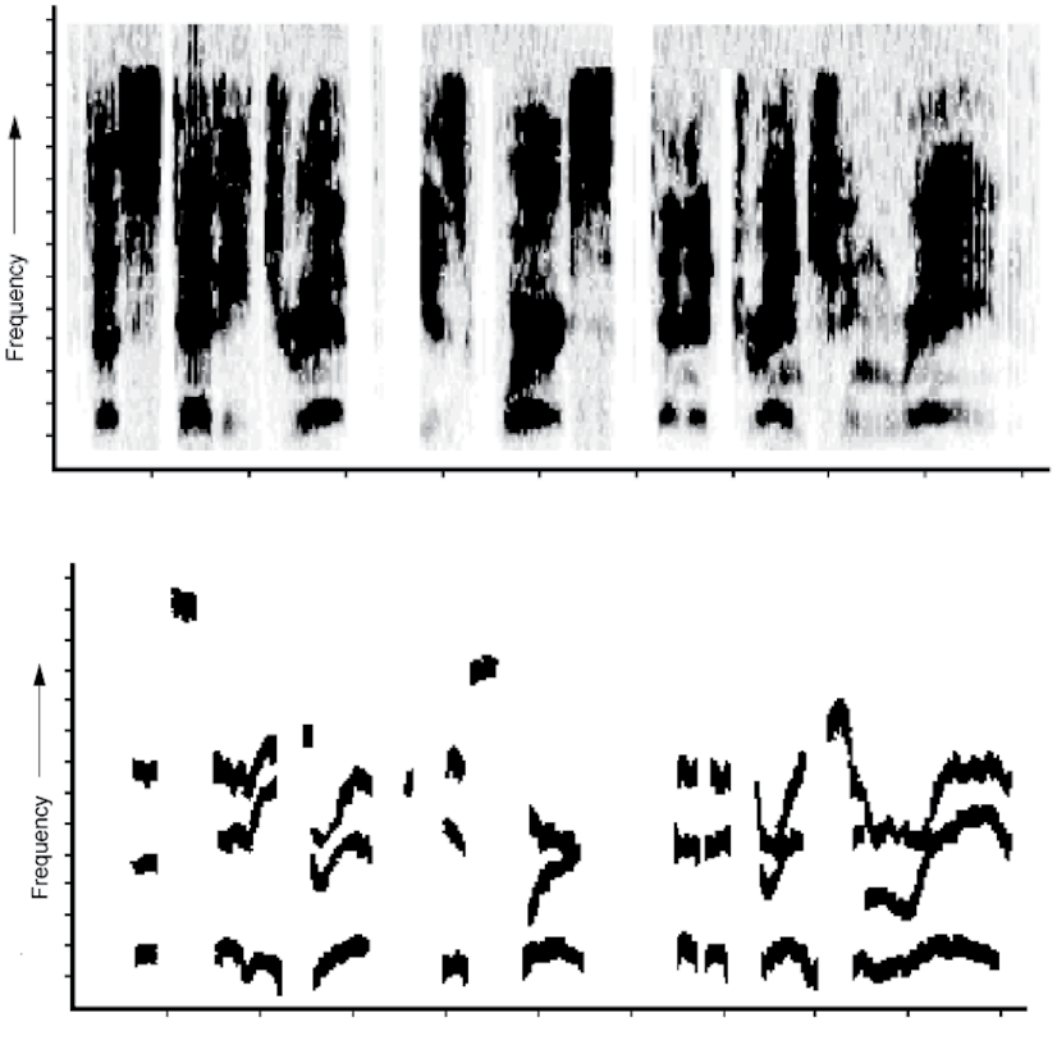

Figure 1.1. A comparison of spectrographic displays of normal (top) and sinewave speech. The sentence depicted in both instances is: "The steady drip is worse than a drenching rain." For more examples, including audio, see http://www.haskins.yale.edu/research/sws.html. Used with the permission of Philip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

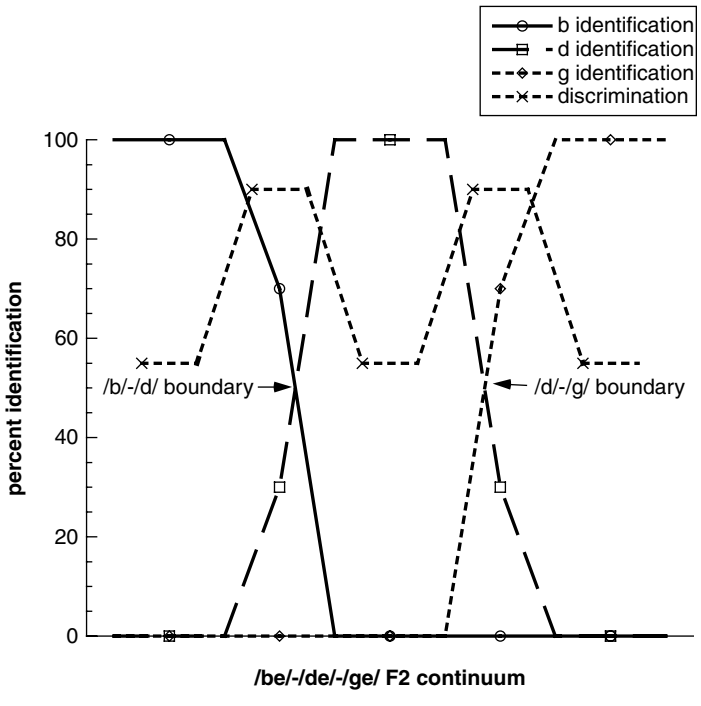

Figure 1.2. Schematic depiction of categorical perception findings. Identification functions are sharp, rather than gradual, and discrimination is poor within as compared to between consonant categories.

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 (F0). Other things equal, high vowels, such as /i/, have higher F0 than low vowels such as /a/, a phenomenon known as *intrinsic F0*. Another use of F0 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 F0 than that on /a/. Listeners parse F0 that they ascribe to intrinsic F0 from F0 that they ascribe to an intonational accent. They do not ignore intrinsic F0 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