Fowler, C.A. (2004). Speech as a supramodal or amodal phenomenon. In: Gemma A. Calvert, Charles Spence, and Barry E. Stein (eds.), *The Handbook of Multisensory Processes*. MIT Press, pp. 189-201.

12 Speech as a Supramodal or Amodal Phenomenon

CAROL A. FOWLER

Introduction

Speech perceivers are informational omnivores. Although the acoustic medium provides necessary and sometimes sufficient information for phonetic perception, listeners use other sources as well. The surface of the speaking face provides surprising amounts of phonetic information (e.g., Yehia & Vatikiotis-Bateson, 1998), at least some of which observers use when it is available. In noisy settings, perceivers who can see the face of a speaker achieve more accurate percepts than those who cannot (e.g., Sumby & Pollack, 1954). Moreover, given appropriate dubbings of acoustic syllables or words onto visible facial gestures for other syllables or words, perceivers integrate information from the two modalities (the *McGurk effect;* e.g., Brancazio, in press; Massaro, 1998; McGurk & MacDonald, 1976). For example, acoustic *ma* dubbed onto visible *da* is identified predominantly as *na,* an outcome that integrates visible information for place of articulation with acoustic information for manner and voicing. This outcome can be phenomenally striking; people hear one syllable with their eyes open and a different one with their eyes closed.

Another perceptual system that provides useful information about speech is the haptic system. Some individuals who are deaf and blind have learned to talk by placing their hands on the face of (and, in Helen Keller's case, sometimes in the mouth of) a speaker (Lash, 1980; Chomsky, 1986). Moreover, naive normally seeing and hearing individuals show a haptic version of the McGurk effect. With hands in surgical gloves placed over the mouth and jaw of a speaker as the speaker mouths *ga,* perceivers increase identifications as *ga* of syllables along an acoustic continuum ranging from *ba* to *ga* (Fowler & Dekle, 1991).

How should we understand speech perception such that phonetic perception can be achieved in all of these ways?

An analogous question arises when we consider speech production as perceptually guided action. We can first consider speech production guided by the perception of one's own speech. When speakers' own acoustic signals are fed back transformed in some way, their speech is affected. For example, in the Lombard effect, speakers increase their vocal amplitude in the presence of noise (e.g., Lane & Tranel, 1971). When feedback about vowel production is transformed acoustically, speakers change the way they produce vowels as if in compensation (Houde & Jordan, 1998). Hommel, Müsseler, Aschersleben, and Prinz (in press) raise the more general question, how can percepts communicate with action plans? Aren't they coded in different ways, percepts as representations of features of the stimulus input and action plans as some kind of motoric specification? How should we understand speech production and speech perception such that perceived speech can affect produced speech?

Finally, a related question arises when we consider speech as a communicative activity taking place between people. In cooperative conversations, speakers may converge in their dialects, vocal intensity, speaking rate, and rate of pausing (see Giles, Coupland, & Coupland, 1991, for a review). More generally, listeners perceive the phonological message that talkers meant to convey. Speakers talk by producing actions of vocal tract articulators. Ultimately their speech action plan must be to produce those actions. For their part, listeners receive acoustic speech signals, signals that have acoustic, not motoric, features. How can speakers communicate with listeners? How can we understand speaking and listening such that a perceived dialect can affect a produced one, and so, more fundamentally, that a listener can perceive a talker's phonological message? Each of these questions is addressed in this chapter.

Cross-modal speech perception

I consider four approaches to an understanding of cross-modal speech perception. Of these, three invoke the need for a "common currency," that is, a common kind of representation or percept that can be achieved by the different modalities. Effectively, in these approaches, information integrates across the perceptual

modalities, because it is information about a common object or common event. Likewise, perceptual information can affect action because perception and action planning share a common currency. The fourth approach is quite different. It proposes integration by fiat. Information acquired cross-modally integrates because templates for speech percepts in the head associate compatible features acquired by different sensory modalities. Perceived information can communicate with action planning for unspecified reasons.

One approach, derived from Meltzoff and Moore's (1997) AIM (active intermodal mapping) hypothesis, suggests that speech perceivers achieve a supramodal representation of a speaker's utterance—a representation, that is, that somehow transcends the sensory modalities from which it derives. It is the supramodal nature of the representation that allows integration of information derived from different perceptual modalities. A supramodal representation of speech that is compatible with the supramodal representations proposed by Meltzoff and Moore is of vocal tract organs and their actions.

Two other approaches share one idea about speech perception but disagree on another. According to both the motor theory of speech perception (Liberman & Mattingly, 1985, 1989; Liberman & Whalen, 2000) and the direct-realist theory (Best, 1994; Fowler, 1986, 1994), speech perceivers perceive the linguistically significant actions ("gestures") of the vocal tract that occur during speech. Gestural percepts are achieved regardless of the perceptual modality that provides information about them. This proposal is compatible with the extension of AIM to speech, as just proposed.

In the motor theory, these gestural percepts are effectively representations of gestures that are generated in a module of the brain specialized to extract them from perceptual information. These gestural representations, like the supramodal representation of an AIM hypothesis, provide a common currency that allows information to be integrated across, for example, the acoustic and optical modalities (and to be integrated across sensory modalities and motor production systems). I will refer to these representations as amodal, because they are motoric rather than perceptual in nature.

In direct-realist theories of perception (e.g., Gibson, 1979), perception is not mediated by representations. Rather, perceptual systems universally use structure in media that stimulate the sense organs (e.g., patterning in reflected light for vision and patterning in air for hearing) to perceive the world of distal objects and events directly. Information is integrated across the perceptual modalities because it is information about a common object or event. Speech perceivers perceive gestures because that is what information for a speech event, whether it is acoustic, optical, haptic, or proprioceptive, is about. In this case perception is amodal, because it is direct perception of events in the world.

In all of these accounts integration occurs because the representations or the percepts are uniformly of distal properties of the world, not of the proximal stimulation that the sense organs intercept. However, the fourth account of integration discussed in this chapter does not invoke the idea of common currency. Instead, speech percepts achieved auditorily are cued by acoustic features and those achieved optically are cued by visible speech gestures (e.g., Massaro, 1998). The information integrates despite the fact that acoustic features are features of the proximal stimulus, while optical features are properties of the distal event—even in the absence of a common currency—because possible speech percepts are represented mentally associated with both the acoustic cues and the visible gestures that provide information about them. In the following discussion, I elaborate on each of these four theoretical ideas.

SPEECH PERCEPTION AS SUPRAMODAL. Meltzoff and Moore (1997, 1999) proposed their AIM hypothesis to explain the remarkable propensity and ability of newborns to imitate facial gestures. Infants do not necessarily imitate accurately right away. They typically get the organ—tongue or lips—right immediately, but they may not get the gesture right. However, they keep trying and keep refining their own gestures to approximate more closely those of their model over many repeated attempts.

Within hours of being born, infants will imitate a tongue protrusion gesture by an adult (Meltzoff & Moore, 1977). (Imitations are identified as such by coders of videotapes of the infants who are blind to the particular gesture that an infant saw. They use preestablished criteria to classify the infants' facial gestures.) It has been found that tongue gestures by infants are more likely to occur when the adult model is producing or has just produced a tongue gesture rather than a lip gesture. In contrast, lip gestures by infants are more common in response to modeled lip gestures than to tongue gestures. How do the infants do it? As Meltzoff and Moore (1997) point out, the infants can see the model's gesture, but not their own. They can proprioceptively feel their tongue, say, and any gesture that they produce with it, but they cannot feel the model's tongue or its gesture. Accordingly, they must have some way of relating information acquired by different sensory modalities.

Meltzoff and Moore (1997, 1999) propose that perceivers, including young infants, establish "supramodal representations" of bodily organs and their interrelations (e.g., a tongue protruded between teeth). These

representations in some way transcend the sensory modalities, providing a single framework within which information acquired from different perceptual modalities can be compared and contrasted. The representations transcend the modalities by representing the distal world of objects (e.g., tongues) and events (e.g., protrusion gestures) rather than the proximal world of sensory stimulation. That is, given optical proximal stimulation caused by a tongue protrusion gesture, the infants represent not a reflected-light structure but a protruding tongue. Given proprioceptive proximal stimulation caused by their own tongue and its action, they represent not proprioceptive "raw feels" but a tongue in action. Because both modalities yield these representations of distal events, components of the events perceived in different modalities (e.g., tongues) can be compared and identified as equivalent.

Vocal imitation, which infants also show from a young age (by 12 weeks of age in Kuhl & Meltzoff, 1996), Meltzoff and Moore suggest is based essentially on intramodal comparisons. This may appear to pose less of a challenge to understand than the cross-modal imitation of facial gestures. However, vocal imitation and infant speech perception more generally pose some very interesting challenges to theorists. It is likely true that vocal imitation depends most heavily on intramodal comparisons. (However, Kuhl and Meltzoff's model was visible to the imitating infants.) Nevertheless, even in that case the infant solves a rather large problem. There is no way that an infant can come close to reproducing the acoustic signal of an adult model, even setting aside the fact that 12-week-old infants are unskilled speakers. This is because an infant has a tiny vocal tract and an adult has a large one. Accordingly, the fundamental frequencies of infant vocalizations and the formant frequencies of the vowel-like sounds they produce at 12 weeks are much higher in frequency than those of an adult. Infants could not imitate by achieving an acoustic match to adult vowels even if they were skilled enough to produce the same vocalic gestures as an adult. On what basis does the infant detect correspondences between its productions and those of other speakers? Even intramodal comparisons require establishment of a common currency that allows abstract equivalences in events to be detected.

In any case, infant speech perception itself is not unimodal (just as infants' perception generally is cross-modal; e.g., Bahrick, 1987; Bahrick & Watson, 1985; Spelke, 1979). In research by Kuhl and Meltzoff (1982, 1988; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983), infants looked at two films showing a speaker producing the vowel they heard being produced. In the films, which were presented side by side, one speaker mouthed /i/ and the other mouthed /a/. If the acoustic vowel was /a/, infants looked longer at the visible /a/ than at the visible /i/; they looked longer at the visible /i/ if the acoustic vowel was /i/ rather than /a/. Kuhl and Meltzoff (1988) report replicating the finding with the vowels /i/ and /u/.

Compatibly, Rosenblum, Schmuckler, and Johnson (1997) reported finding a McGurk effect in 5-month-old infants. They used a procedure in which discrimination is assessed with a looking-time habituation procedure. As long as infants watched a video display, they heard a sound being produced. When they looked away, the sound stopped. In such circumstances, when the same sound is repeated over trials, once the infants detect the contingency between their looking and the presentation of a sound, they look for a long time initially, but over trials their looking time decreases. For infants in experimental conditions, following habituation to the first sound (indexed by a criterion decrease in looking time), a new sound accompanies the video display. If the infant discriminates the sound, looking time increases again; if he or she does not, looking time remains at a low level or decreases further. Infants in a control condition receive the same sound before and after habituation. Rosenblum et al. used a face mouthing /va/ as the video display and, during habituation, a /va/ acoustic syllable. After habituation, infants in the experimental group received (on alternate trials) video /va/ accompanied by audio /ba/ or video /va/ accompanied by audio /da/. (Adults hear the first pairing as /va/ but the second as /da/.) In the experiment, infants increased their looking time (relative to no-shift control infants) only when the acoustic signal was /da/. A subsequent experiment, in which an unmoving face accompanied the same acoustic syllables as in the first experiment, yielded a different pattern of looking-time changes. In the presence of an unmoving face, adults should hear the prehabituation syllable as /va/ and the posthabituation syllables as /ba/ and /da/. In this experiment, infants' looking times increased relative to the no-shift condition only on /ba/ trials. Accordingly, the experimenters inferred that the result pattern in the first experiment was not due to acoustic /ba/ sounding less different from /va/ than acoustic /da/. Rather, infants experienced cross-modal integration such that video /va/-audio /va/ and video /va/-audio /ba/ sound the same, just as they do for adult listener-viewers.

These findings raise questions very much like those posed for infants who imitate visible facial gestures. In Kuhl's research, how do infants know which facial speech gestures match which acoustic vowel? In the research of Rosenblum et al., how are optical and acoustic speech displays perceived such that the infants

experience cross-modal integration? There has to be a way to compare information acquired in two different perceptual modalities.

Meltzoff and Kuhl (1994; cf. Meltzoff, Kuhl, & Moore, 1991) propose that the basis for infants' cross-modal integrations is a learning process in which infants come to associate vocal gestures and their auditory outcomes during cooing and babbling when they both see and hear speech. These associations permit integration of audiovisual information such as that provided by Kuhl and Meltzoff (1982) and Rosenblum et al. (1997). It may also serve imitation of speech, as observed by Kuhl and Meltzoff (1996). Given a (normalized) acoustic signal produced by a model speaker, infants can use the associations to determine which gestures will generate the equivalent infant vocalization.

This proposal is not analogous to the proposal that Meltzoff and Moore (1997, 1999) developed to explain infants' cross-modal identification of facial gesture equivalences. A proposal analogous to the AIM hypothesis would be that cross-modal speech perception implies establishment of supramodal representations of what it looks like and sounds like to produce a particular sequence of gestures. What would such a representation be like? Meltzoff and Moore propose that supramodal representations supporting facial gesture imitation are representations of real-world—that is, distal—objects (organs) and events (gestures). That is what unites the perceptual input from different sensory modalities. We need an idea like that of Meltzoff and Moore for speech perception. The real-world events of phonetic speech production are, as they are for facial-gesture production and perception, actions of organs of the vocal tract. Facial gestures and speech gestures are alike in this respect. They differ in two ways. Speech actions cause patterning, and not only in the optic array and, for the speaker, proprioceptively; they also structure the air. Moreover, and relatedly, they are linguistically significant vocal tract actions; we refer to them as *phonetic gestures*. If infants (and adults) perceive phonetic gestures from both acoustic information (e.g., Fowler, 1986) and optical information about them, the ability of infants to detect which of two visible faces is mouthing the vowel they hear is understandable.

Supramodally represented speech gestures can also be invoked to account for the findings of Rosenblum et al. (1997). It is understandable that infants as well as adults integrate optical and acoustic information for speech if both modalities yield supramodal, distal percepts—that is, percepts of the speaker's phonetic gestures. In that way, gestures specified by one modality may be compared and, if the information warrants it, identified with those specified by another.

Note that if infants perceive phonetic gestures, we can understand not only their audiovisual speech skills but perhaps their vocal imitation skill as well. Gesture perception provides an abstract equivalence of the sort that infants must achieve in order to compare their own vowels with those of an adult. Infants who perceive acoustic /u/ as a high back gesture of the tongue coupled to a protrusion gesture of the lips when an adult produces it can attempt to generate those gestures in their own vocal tract based on proprioceptive information. Likewise, infants can detect the match between the gestures of /u/ that they hear and the visible lip protrusion of a videotaped speaker producing /u/.

Do infants perceive phonetic gestures (see Lewkowicz and Kraebel, Chap. 41, this volume)? The experimental data are not in yet, but we do know that infants integrate information about speech cross-modally. We also know that they learn to produce speech by perceiving it spoken by others and by themselves. They must get gestural information from what they see and hear if perception guides their learning to talk. I will also suggest next that there is considerable evidence for gesture perception by adults.

EVIDENCE FOR PHONETIC GESTURE PERCEPTION The idea that listeners to speech perceive linguistically significant actions of the vocal tract is central to the account of cross-modal speech perception that this chapter has offered so far. It is justified in the AIM-derived account by the need for a common currency that allows perceptual information to be integrated cross-modally. It is justified in this and other ways by the two accounts to be presented next. Nontheless, the idea is considered radical in the field of speech research.

Here I summarize some of the evidence that gestures are perceived. The review is brief and incomplete; the point here is to justify the idea.

Equivalence of cross-modal information This evidence has already been described. It includes the evidence that infants and adults integrate speech cross-modally and that they imitate speech. As Meltzoff and Moore (1997) propose for facial gestures, there must be a common currency for percepts achieved by different sensory modalities to be compared. The obvious domain is that of distal events—in speech, phonetic gestures.

The rapidity of imitation Luce (1986) points out that there is canonically a 100 to 150 ms difference in latencies to comparable simple and choice response tasks. Simple tasks are those in which participants make speeded responses when they detect an event. The response does not vary with the event. In choice tasks, different responses are made to different events, so the

participant needs to do more than detect the occurrence of an event to choose his or her response. With appropriately chosen speech tasks, however, the difference in latencies can be very much smaller than the canonical difference, and both sets of latencies are in the range of canonical simple response times.

Both Kozhevnikov and colleagues (Kozhevnikov & Chistovich, 1965) and Porter and colleagues (Porter & Castellanos, 1980; Porter & Lubker, 1980) have obtained this outcome. Porter and Lubker had participants "shadow" a model speaker who produced an extended /a/ vowel followed at an unpredictable interval by a consonant-vowel (CV) syllable. In the simple task, participants produced the extended vowel with the model and switched to a designated syllable (/ba/) when the model produced a CV. In the choice task, participants produced the extended vowel with the model, and when the model shifted to a CV, the participant shifted to the same CV. In that experiment, choice task times exceeded simple task times by only 50 ms. In a recent replication of our own (Fowler, Brown, Sabadini, & Weihing, 2003), the difference was 26 ms. In both experiments, responses were very fast and more characteristic of canonical simple than choice latencies. Kozhevnikov et al., Porter et al., and Fowler et al. all interpreted the findings as demonstrating gesture perception. If speech perceivers perceive the speaker's gestures, then, in the choice condition, the percept provides instructions for the required response.

The disposition to imitate Infants (e.g., Meltzoff & Moore, 1977) and adults (e.g., McHugo, Lanzetta, Sullivan, Masters, & Englis, 1985) are disposed to imitate others' facial gestures, and infants (Kuhl & Meltzoff, 1996) and adults (Goldinger, 1998) imitate others' speech. In infancy, imitation may have the function of helping children to acquire culturally appropriate behavior. But the tendency persists into adulthood. The persistence of imitation may occur because perception generally provides instructions for action. In speech, instructions for producing gestures are perceived gestures.

Conflicting and cooperating cues Generally, if two stimuli differ in two ways acoustically, they are easier to discriminate than if they differ in just one of two ways. This is not always the case in speech, however. Fitch, Halwes, Erickson, and Liberman (1980) synthesized stimuli ranging from *slit* to *split* by varying two acoustic cues. A silent interval between the offset of the initial /s/ and the onset of /l/ was varied in duration. In addition, labial formant transitions either followed the silence or were absent. Both a silent interval and labial transitions provide information for /p/. Fitch et al. first found a trading relation between the cues. That is, silence of a long enough duration was sufficient, in the absence of transitions, to yield a *split* percept, but less silence was needed if transitions were present. Some amount of silence could be traded for transitions. Next, Fitch et al. had participants discriminate pairs of the syllables in which pair members differed in either of three ways. In the first way, the syllables differed only in the silence duration; transitions were either present or absent in both members of a pair. In the two other conditions, pair members differed in both cues. In the "cooperating cues" condition, one member of a pair had a longer silent interval than the other, and it had transitions, whereas the other did not. That is, either both cues signaled /p/ or both signaled its absence. In the "conflicting cues" condition, pair members differed in the same two ways as in the cooperating condition, but one cue was swapped between pair members. Now the pair member with the longer silence lacked transitions whereas the other pair member had them. The striking finding was that listeners did discriminate cooperating pairs better than one-cue-different pairs, but they discriminated the conflicting pairs *worse* than one-cue-different pairs. Even though conflicting pairs differed in the same way as one-cue pairs and differed in one more way as well, discrimination performance was worse than in the one-cue condition. This finding was very surprising on acoustic grounds but is expected in gesture theory. In the cooperating condition, one pair member was a clear *split* and the other a clear *slit*. In the conflicting condition, often both members of a pair sounded like *split* or both like *slit*. It is not the acoustic quality of the cues that matters; rather, it is what they inform the listener about the presence or absence of a labial speech gesture. This finding has been replicated by Best, Morrongiello, and Robson (1981) using a *say-stay* continuum. Using a /sa/-/spa/ continuum, Nittrouer and Crowther (2001) have replicated the findings only in part. Five-year-olds discriminated one-cue-different pairs better than two-cue conflicting pairs; 7-year-olds showed no difference, and adults showed the opposite pattern. However, all three age groups discriminated cooperating cues pairs better than conflicting cues pairs (with the findings for 7-year-olds just marginally significant, $P = 0.10$).

Parsing the speech signal Speech is coarticulated so that information about different consonants and vowels is conveyed by the same acoustic structure. For example, formant transitions provide information about both a consonantal and a vocalic gesture. In this way, gestures can have converging effects on the same acoustic dimensions. A good example is provided by fundamental frequency (f_0). Generally, the f_0 contour provides

information about intonation (speech melody). However, it also shows declination, an early to late decrease in f_0 along a stretch of speech. In addition, high vowels are associated with higher f_0 values than low vowels, and voiceless obstruents cause a high falling tone on a following vowel. Research has shown that listeners "parse" effects of distinct gestures from f_0. That is, for /i/ and /a/ vowels to have the same pitch, /i/ has to be higher in f_0 than /a/ (e.g., Fowler & Brown, 1997). Listeners use the f_0 that a high vocalic gesture should have caused as information for vowel height, not vowel pitch or intonation (Reinholt Peterson, 1986). Declination is also parsed from intontation (Pierrehumbert, 1979; Silverman, 1987), as are the effects of stop consonant devoicing (e.g., Pardo & Fowler, 1997). There are similar findings with regard to segmental information for coarticulated consonants and vowels (e.g., Fowler & Smith, 1986). That is, speech perceivers parse the acoustic signal along gestural lines.

The two remaining accounts of cross-modal perception share with the earlier AIM-derived account a claim that listeners to speech perceive gestures.

Speech Perception as Amodal: 1. The Motor Theory

The motor theory of speech perception (e.g., Liberman, 1982; Liberman & Mattingly, 1985) offers an account of cross-modal speech perception that is similar to the account I derived above as an extension of AIM. The motor theory was developed to account for evidence suggesting a closer correspondence between speech percepts and articulation than between speech perception and the mediating acoustic speech signal. Two findings spurred development of the theory in the 1950s. Liberman and colleagues found that intelligible stop consonants (e.g., /b/, /d/, /g/, /p/, /t/, and /k/ in English) in CV or VC syllables could be synthesized in either of two ways, each way providing one of two prominent acoustic signatures of a stop consonant. One way, in CVs, was to provide the formant transitions into the following vowel appropriate to the consonant in the context of the vowel. The transitions are produced after release of the consonant as the vocal tract opens up for the vowel. The other way was to place the stop burst that occurs at release of the consonantal constriction just before a set of steady-state formants for the vowel.

One critical finding (Liberman, Delattre, Cooper, & Gerstman, 1954) was that the second formant transitions of synthetic syllables are acoustically quite different. As shown in Figure 12.1*A*, that for /di/ is a high rise in frequency, while that for /du/ is a low fall.

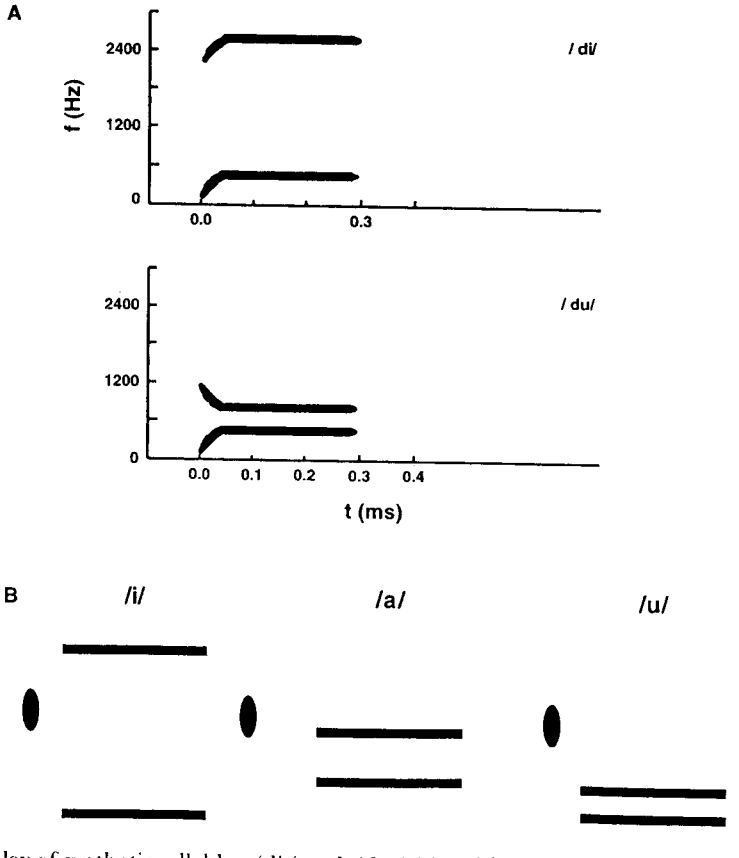

Figure 12.1 Schematic display of synthetic syllables /di/ and /du/ (*A*), with consonants specified by their formant transitions, and of /pi/, /ka/, and /pu/ (*B*), with consonants specified by stop bursts.

Isolated from the vowel, the transitions are audibly quite different. However, they are produced by the same gesture, a constriction of the tongue tip against the alveolar ridge of the palate. The acoustic differences arise from coarticulation, that is, from speakers overlapping temporally successive consonants and vowels. While the tongue tip constriction is being produced, the body of the tongue is conforming itself for the following vowel, as are the lips if the vowel is /u/. Therefore, after release of the tongue tip constriction, the vocal tract opens up into different vowels in the two syllables, and the formant transitions as well as the steady-state values of the formants are different. In this finding, two markedly different acoustic cues for /d/ caused by the same gesture (coarticulating with different ones) sound the same. The percept tracks articulation.

The second finding (Liberman, Delattre, & Cooper, 1952) was complementary. Here, the same stop burst placed before /i/ or /u/ was identified as /p/, but when placed before /a/, it was identified as /k/ (Fig. 12.1*B*). Now the same bit of acoustic signal that, because of coarticulation with the following vowel had to be produced by different constriction gestures before /a/ than before /i/ or /u/, sounded different to listeners. Again, listeners tracked gestures.

The logic by which these findings led Liberman and colleagues to develop the motor theory of speech perception is shown in Figures 12.2*A* and *B*. In the case of synthetic /di/ and /du/ the same oral constriction gesture gives rise, because of coarticulation, to different critical formant transitions; however, the consonants perceived, like the gestures, are invariant. In the case of synthetic /pi/, /ka/, and /pu/, different oral constriction gestures give rise, because of coarticulation, to the same acoustic stop burst; the consonants perceived, like the constriction gestures, are different.

In its final or near-final form, the motor theory (e.g., Liberman & Mattingly, 1985) proposed that gesture perception occurs in a module of the brain specialized for both speech production and speech perception. Coarticulation in speech is necessary, in the view of motor theorists, because consonants and vowels have to be sequenced very fast in order not to exceed the memory span of listeners during an utterance; however, if the segments were produced discretely, without coarticulation, the rate at which they occur would exceed the temporal resolving power of the ear. Coarticulation eliminates the discreteness of speech elements, but it creates a new problem. The relation between the acoustic signal and the segments that listeners must recover to identify words is very complex. The phonetic module evolved to deal with that complexity.

Coarticulation and perception of coarticulation are capabilities that had to coevolve, because neither would be useful without the other. According to the motor theory, the phonetic module is a single specialization of the brain that serves both capabilities, and it is this that underlies our perception of gestures rather than acoustic speech signals in the theory. The phonetic module uses its competence to generate coarticulated speech and to

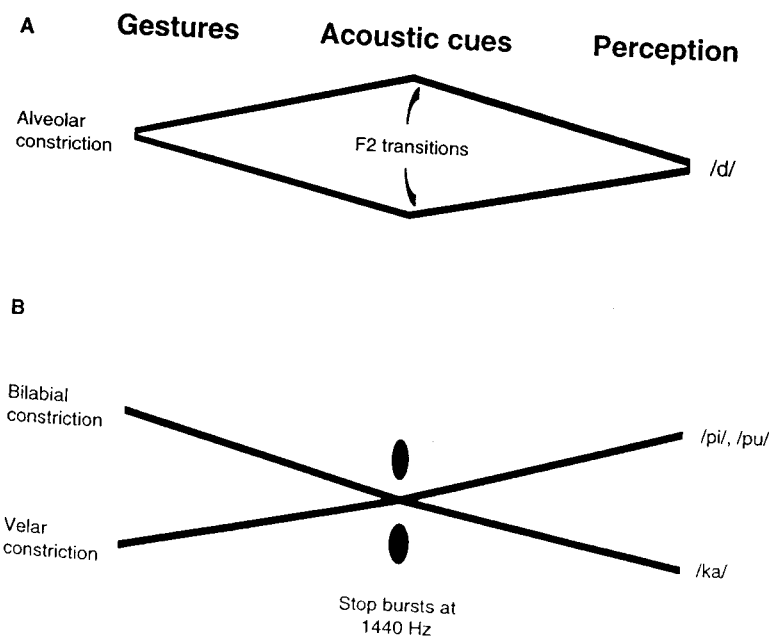

FIGURE 12.2 The logic by which results of perceptual experiments using the stimuli of Figure 12.1 led Liberman to develop his motor theory of speech perception.

recover the speakers' "intended" gestures. (Liberman and Mattingly argued that actual gestures of the vocal tract were not the same as intended gestures, because of coarticulation.)

A gestural percept might be called "amodal" rather than supramodal if the word "modality" refers just to perceptual systems. The percept is motoric in the theory. Otherwise, however, the motor theory's account of cross-modal speech perception is not very different from the account derived as an extension of Meltzoff and Moore's AIM. The gestural percept provides a common currency in which information about gestures derived acoustically, haptically, proprioceptively, and optically can converge.

In Liberman's motor theory, as in the extension of AIM, cross-modal speech perception is achieved by a percept that, in some way, is not tied to any perceptual modality. In the AIM account, the percept achieves that by being about distal properties of an event, not about the proximal stimulation itself. In the motor theory, the percept is not quite about distal properties, because it is of intended gestures rather than actual ones. However, it is similar in not being tied to any perceptual modality. In the motor theory, the percept is a motor percept.

The motor theory makes three central claims. It claims that listeners perceive (intended) speech gestures, that perception is achieved by a module of the brain dedicated to speech production and perception, and that perceptual speech processing recruits the speech motor system and its motor competence. The first claim is common to all three views presented here, and evidence for it has been offered earlier.

Behavioral evidence that has been interpreted as evidence for a dedicated speech processing system is provided by duplex perception. In one version of this finding (e.g., Mann & Liberman, 1983), listeners are presented with synthetic syllables identified as /da/ and /ga/. These syllables can be synthesized so that they are identical except for the third formant transition, which falls for /da/ and rises for /ga/. If the part of the syllable that is the same for /da/ and /ga/ (called the base) is presented to one ear and the distinguishing transition is presented to the other, listeners integrate the information across the ears and hear /da/ or /ga/, depending on the transition. However, at the same time, they also hear the transition as a pitch rise or fall. The finding that part of the signal is heard in two different ways at the same time has been interpreted to suggest that two different perceptual systems are responsible for the two percepts. One, a phonetic processor, integrates the base and transition and yields a phonetic percept, /da/ or /ga/. The other yields a "homomorphic" percept (Liberman & Mattingly, 1989), that is, a percept having the same form as the acoustic signal. Presumably this is an auditory processor.

This interpretation has been challenged on a variety of grounds (e.g., Fowler & Rosenblum, 1990; Pastore, Schmuckler, Rosenblum, & Szczesiul, 1983). I will not review those challenges here. Rather, I note that the motor theoretical interpretation of duplex perception would be strengthened by evidence favoring the third claim of the theory, that there is motor system or motor competence involvement in perceiving speech. This is because theorists do not claim motor involvement in auditory perception generally. If evidence favors motor involvement in speech perception, this would constitute grounds that speech perception is achieved by a different perceptual system than the auditory system, one perhaps dedicated to speech perception.

In fact, evidence for motor involvement in speech perception is weak. However, apparently this is because such evidence has rarely been sought, not because many tests have yielded negative outcomes. I have found three kinds of behavioral data and some suggestive neuropsychological data.

Following a seminal study by Eimas and Corbit (1973), many researchers studied the effects of "selective adaptation" on speech perception. Listeners heard repeated presentations of a syllable at one end of an acoustic continuum, say, /pa/, and then identified members of, say, a /pa/ to /ba/ continuum. The consequence of listeners hearing repeated /pa/ syllables was a reduction in the number of syllables in the ambiguous region of the continuum that listeners identified as /pa/. Eimas and Corbit suggested that phonetic feature detectors (a detector for voicelessness in the example) were being fatigued by the repetitions, so that ambiguous stimuli were more likely to be associated with more activity in the unfatigued voicing feature detector than in the voicelessness detector. This account was challenged (e.g., by Diehl, Kluender, & Parker, 1985). For our purposes, the interpretation is less important than the finding by Cooper (in studies summarized in Cooper, 1979) that repeated presentations of a syllable such as /pi/ had weak but consistent influences on production of the same syllable or another syllable sharing one or more of its features. In the example, voice onset times of produced /pi/s and /ti/s were reduced after adaptation by perceived /pi/s.

Bell-Berti, Raphael, Sawusch, and Pisoni (1978) provided a different kind of evidence for a motor theory. The vowels /i/, /ɪ/ , /e/, and /ɛ/ all occur in English. The differences among the vowels can be described in two ways. On the one hand, they decrease in "height" in the series as listed above. On the other hand, /i/ and

/e/ are described as tense vowels; /ɪ/ and /ɛ/ are their lax counterparts. Within the tense vowel pair and the lax pair, vowels differ in height. Bell-Berti et al. found that speakers differed in how they produced the vowels in the series. In four speakers, activity of the genioglossus muscle (a muscle of the tongue affecting tongue height) gradually decreased in the series of four vowels listed above, suggesting progressively lower tongue heights. In contrast, six of the speakers showed comparable levels of activity for /i/ and /e/ that were much higher than activity levels for the two lax vowels. This suggested use of a tense-lax differentiation of the vowels.

Remarkably, on a test of perception, the ten participants partitioned into the same two groups. Listeners identified vowels along an /i/ to /ɪ/ continuum under two conditions. In one, the vowels along the continuum were equally likely to occur. In the other, anchoring, condition the vowel at the /i/ end of the continuum occurred four times as frequently as the other continuum members. This tends to decrease /i/ identifications. Participants who showed progressively decreasing levels of genioglossus activity in their production of the four vowels showed much larger effects of anchoring than the six speakers who produced /e/ with more genioglossus activity, and presumably a higher tongue, than /ɪ/. The authors speculated that the difference occurred because, for the second group of listeners, /i/ and /ɪ/ are not adjacent vowels, whereas they are for members of the first group. Whatever the appropriate account, it is telling that the participants grouped in the same way as talkers that they did as listeners. This provides some evidence suggesting that speech percepts are mediated by information about production of speech.

A final behavioral finding is provided by Kerzel and Bekkering (2000). These investigators looked for compatibility effects in speech production. They presented a face visibly mouthing /bʌ/ or /dʌ/. At a variable interval after that, they presented either of two symbols that participants had learned to associate with the spoken responses /ba/ and /da/. They found an effect of the irrelevant visible speech gesture on latencies to produce the syllables cued by the symbols such that /ba/ responses were faster when the face mouthed /bʌ/ than when it mouthed /dʌ/. Likewise, /da/ responses were facilitated by visible /dʌ/. Kerzel and Bekkering argued that these effects had to be due to stimulus (visible gesture)-response compatibility, not stimulus-stimulus (that is, visible gesture-visible symbol) compatibility, because the symbols (## and &&) bear an arbitrary relation to the visible gestures, whereas the responses do not. Their interpretation was that the visible gestures activated the speech production system and facilitated compatible speech actions, an account consistent with the motor theory. It has yet to be shown that acoustic speech syllables, rather than visible speech gestures, have the same effect.

There is a little recent neuropsychological evidence providing support for a motor theory of speech perception. Calvert et al. (1997) reported that auditory cortical areas activate when individuals view silent speech or speechlike movements. Moreover, the region of auditory cortex that activated for silent lipreading and for acoustic speech perception were the same. More recently the same group of investigators (MacSweeney et al., 2000) replicated the findings using procedures that ensured that fMRI scanner noise was not the source of the auditory cortical activation.

Using transcranial magnetic stimulation of the motor cortex, Aravind, Sundara, and Chen (2000) recorded motor-evoked potentials in the orbicularis oris muscle of the lips under a variety of audiovisual conditions. Participants listened to acoustic /ba/, they watched visible /ba/, and they watched and listened to audiovisual /ba/. In addition, they watched visible /ta/, as well as audio /ba/ dubbed onto video /ta/. Under the conditions of this dubbing, most people report hearing /da/. In all conditions in which participants perceived a labial consonant, orbicularis oris activity numerically or significantly exceeded that during a baseline condition in which participants saw a motionless face. Lip activity did not exceed baseline in conditions, including the McGurk dubbing condition, in which the percept was a nonlabial consonant.

There is considerable evidence that listeners perceive phonetic gestures and some evidence favoring the motor theory's account that the motor system is involved in perception. This theory explains cross-modal speech perception quite naturally by claiming that all information about gestures is integrated into a motor percept.

SPEECH PERCEPTION AS AMODAL: 2. DIRECT REALISM In both the motor theory and the AIM extension, percepts are achieved in the head. They can be integrated when they are achieved via different sensory modalities because the percepts have distal (or, in the motor theory, almost distal) reference. In speech perception, they are about the gestures of the vocal tract that the speaker used to produce the acoustic speech signal.

In direct realism (e.g., Gibson, 1966, 1979), perceivers do not represent the world, they perceive it. In this account, information extracted about events across different sensory modalities integrates because it is about the same event. It is not that a perceptual representation has distal reference. Rather, perception is of the distal event itself, and distal things are amodal.

In the direct-realist theory of speech perception (Best, 1995; Fowler, 1986), relevant distal events are linguistically significant actions of the vocal tract, that is, phonetic gestures. In contrast to the motor theoretical idea that coarticulation prevents intended gestures from being actual, in the direct-realist view, coarticulation is nondestructive and intended gestures are actual (Fowler & Saltzman, 1993). Despite considerable coarticulatory overlap of /b/ production by surrounding vowels, for example, speakers always get their lips closed, and that is /b/'s constriction gesture (cf. Fowler & Saltzman, 1993).

In direct-realist accounts generally, perception serves a universal function across perceptual modalities. Its very important function is to allow perceiver/actors to know the world. Perception serves its function in just one general way. Structure in informational media (e.g., light for vision, acoustic signals for hearing) is caused by properties of environmental events. Moreover, distinctive properties of events structure the media distinctively and so provide information for their causal source. Perceivers/actors intercept structure in media during their exploratory and performatory actions. The structure, imparted to the sensory modalities, is used perceptually, not as something to be perceived itself but as information for its causal source in the environment.

In this theory, then, as in the previous two, perception is generally of distal properties of the world, not of proximal stimulation. It is this aspect of the theory that underlies its claim, applied to speech, that listeners perceive phonetic gestures. In contrast to AIM, perceived gestures are not mediated by representations in a direct-realist account. In contrast both to AIM and the motor theory, percepts are not representations that are about something distal (or nearly distal in the motor theory), they are distal themselves.

It is also this aspect of the theory that allows it to account for cross-modal speech perception. Speakers cause structure in acoustic signals that constitutes information about gestures. They also cause structure in the light that constitutes information for gestures. If a hand is placed on the face as someone speaks, then the speech also structures the skin on the hand, which provides haptic information for gestures. Because all of the modalities for which speaking generates information use proximal stimulation as information for its causal source in the world, all of them acquire information about gestures, and this is how cross-modal information can integrate.

OTHER ACCOUNTS: NO COMMON CURRENCY The AIM-derived account, the motor theory, and direct realism have much in common in invoking distal speech properties as the common currency that permits cross-modal integration of information. I will summarize just one alternative account of cross-modal speech perception, that offered by Massaro's Fuzzy Logical Model of Perception (FLMP; e.g., Massaro, 1987, 1998). In the theory, speech perceivers have prototypes of CV syllables in their heads. Associated with the prototype are acoustic features and optical gestural descriptions. Audiovisual speech information integrates because of the mutual attachment of acoustic and visible speech information to common prototypes.

When a speech event is both viewed and heard, properties of the event (acoustic cues and visible gesture information) are extracted perceptually. Fuzzy logical values (between 0 and 1) are assigned to each feature of every prototype, with the number reflecting the strength of the evidence that the feature is present in the stimulation. The fuzzy logical numbers for the features of each prototype are multiplied to generate one number per prototype, reflecting its degree of match to the stimulation. The prototype reflecting the best match to the input is chosen as the percept. Because both optical information and acoustic information are associated with prototypes for CVs, both contribute to CV perception.

The FLMP accounts well for findings of audiovisual speech perception. The model can simulate human speech perception data admirably. Although in its current form it does not account for other kinds of cross-modal speech perception that occur, it could by the addition of, say, haptic speech features or proprioceptive ones.

I find the FLMP account unsatisfactory in relation to the other three accounts I have offered, in part because it does not confront the problem of common currency that is necessary to provide not just a simulation of but an explanation of cross-modal speech perception.

In the three gestural accounts that I have offered, proponents identified a need for cross-modal information to be in a common code or currency in order to integrate. This is, for example, a fundamental idea in Meltzoff and Moore's AIM hypothesis that allows it to explain how infants can relate facial gestural information acquired cross-modally. All of the accounts I have offered draw the conclusion that the common currency is gestural. In contrast, in FLMP there is no common currency. The theory proposes that information acquired auditorily is proximal (acoustic cues), whereas information acquired optically is distal (e.g., lip closure for /ba/). Information integrates across the modalities not, therefore, because it is commensurate cross-modally but because optical features for a syllable and

acoustic features for the same syllable are attached to the same prototype. However, the theory is less explanatory than any of the other three described earlier. This is because there is no explanation of how the proximal acoustic cues and distal gestures becomes attached to a prototype or why a set of acoustic and optical features amounts to a particular CV. Why are the auditorily extracted properties proximal and the optical ones distal? Pairings of compatible acoustic and gestural cues (that is, cues for the same CV) in memory might be explained by invoking experiences of both seeing and hearing speech (but see Fowler & Dekle, 1991, for a challenge to that account). However, why does any pairing amount to a particular CV whose "name," say, is /di/ or /du/? Where do the names come from, and why do the two syllables in the example have the same C in their names? The answer cannot be found in the acoustic cues (the critical F2 rises for /di/ but falls for /du/) or in the facial gestures (/d/ in /du/ looks more like /g/ in /gu/ than it looks like /d/ in /di/). All of these questions have answers if there is a common currency underlying the integration of auditorily extracted and visually extracted information about speech. They appear not to in the FLMP account.

Common currency and perceptually guided speaking

The discussion so far has been about cross-modal speech perception and how ideas of supramodal representations or common currencies appear to provide the means by which information can be integrated cross-modally. My topic, however, is not speech perception as a supramodal or amodal phenomenon, but speech itself as a supramodal or amodal phenomenon. It is not commonplace, and perhaps is not good usage, to refer to the motor system as a mode, and indeed, I excluded it as such in referring to the motor theory's representations as amodal. However, it is instructive to look at some issues that arise in accounting for perceptually guided action; they appear strictly analogous to those that arise in accounting for cross-modal perception.

Hommel et al. (2001) ask how percepts can communicate with action plans. If there are perceptual codes that are sets of perceptual features and action plans that are motor commands, the two codes cannot communicate, and actions cannot be perceptually guided. But, of course, they are. Hommel et al. propose that there are codes, and that their common currency is featural. The features are those of the distal environment in which the perceptually guided action occurs. A common code is quite analogous to a supramodal representation. An alternative idea is that the common currency is the distal event itself, not the represented features of that event. It is the event that the talker generates and that the listener perceives cross-modally.

One example of perceptually guided action in speech is the imitation of speech that both infants (Kuhl & Meltzoff, 1996) and adults (Goldinger, 1998) exhibit. How can the speech of a talker guide an imitative response? It can if there is a common currency shared by percepts and preparations for action. Again, we invoke speech gestures. If listeners perceive gestures and plan to produce gestures, an imitative response should be the most straightforward way in which perception can guide action. The instances of convergences in dialect, speaking rate, and vocal intensity cited in the introduction to this chapter are other examples of perceptually guided imitation.

There are other examples of perceptually guided action as well. Houde and Jordan (1998) transformed the formant frequencies of vowels produced by participants so that participants heard fed-back vowels different from the ones they had produced. Just as in cases of adaptation to prisms, this led to compensatory changes in articulation to make the acoustic feedback more appropriate to the vowel that the speaker had intended to say. The fed-back vowel, perceived as the gestures that would have produced it, affected the gestures that the speakers went on to produce.

Common currency and communication

As Liberman understood well there is another critical need for a common currency in speech that a conclusion that listeners perceive phonetic gestures addresses. Liberman and colleagues (e.g., Liberman & Whalen, 2000) referred to a "parity" requirement in speech communication. This, in Liberman and Whalen's article, had three core aspects, the first two of which derived from the fact that speech is a between-person activity. Listeners and talkers have to "agree" on what public actions of a speaker count as producing language forms. (As Liberman and Whalen put it, /ba/ counts, a sniff does not.) In addition, for utterances to do their intended work, listeners have to perceive the language forms that speakers say. In these two ideas of parity, listeners and talkers require a common currency, both in the sense of sharing what the currency is and in the more local sense of listeners recovering the specific forms constituting a speaker's utterance. The third aspect of parity, for Liberman and Whalen, is the necessary coevolution of coarticulation and the ability to perceive coarticulated speech.

The idea of common currency (*parity*, in the theorizing of Liberman) is absent from nongestural theories of

speech perception (e.g., Diehl & Kluender, 1989; Kluender, 1994; Nearey, 1997) but is needed for multiple reasons, not only to explain information integration across the perceptual modalities or perceptually guided action. Liberman and Whalen point out that it is required to explain how speech can communicate. What counts as something phonetic is the same for talkers and listeners. More locally, for communication to take place, generally, listeners have to perceive with sufficient accuracy the language forms produced by the talker. There has to be parity or common currency now between talker and listener.

Conclusion

Speakers and listeners receive multimodal information about language forms. They use the information they receive in all of the modalities that provide information. Their doing so requires a common currency across the sensory modalities. The common currency might be the supramodal representation of Meltzoff and Moore (1997, 1999) or it might, more directly, be the distal object of perception. This same common currency is required to understand how perceptually guided action can occur and how communication can occur between talkers and listeners.

REFERENCES

Aravind, N. K., Sundara, M., & Chen, R. (2000). Effects of auditory and visual stimuli on motor facilitation of speech muscles [Abstract]. *Journal of the Acoustical Society of America, 107,* 2887.

Bahrick, L. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development, 10,* 387–416.

Bahrick, L., & Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology, 21,* 963–973.

Bell-Berti, F., Raphael, L. R., Sawusch, J. R., & Pisoni, D. B. (1978). Some relationships between speech production and perception. *Phonetica, 36,* 373–383.

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

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange & J. J. Jenkins (Eds.), *Cross-language speech perception* (pp. 171–204). Timonium, MD: York Press.

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

Brancazio, L. (2001). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance.*

Calvert, G., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276,* 593–596.

Chomsky, C. (1986). Analytic study of the Tadoma method: Language abilities of three deaf-blind subjects. *Journal of Speech and Hearing Research, 29,* 332–347.

Cooper, W. (1979). *Speech perception and production: Studies in selective adaptation.* Norwood, NJ: Ablex.

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

Diehl, R., Kluender, K., & Parker, E. (1985). Are selective adaptation effects and contrast effects really distinct? *Journal of Experimental Psychology: Human Perception and Performance, 11,* 209–220.

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

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

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

Fowler, C. (1994). Speech perception: Direct realist theory. In *Encyclopedia of language and linguistics* (Vol. 8, pp. 4199–4203). Oxford, England: Pergamon Press.

Fowler, C., & Brown, J. (1997). Intrinsic f0 differences in spoken and sung vowels and their perception by listeners. *Perception & Psychophysics, 59,* 729–738.

Fowler, C., Brown, J., Sabadin, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49,* 396–413.

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,* 816–828.

Fowler, C., & Smith, M. (1986). Speech perception as vector analysis: An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123–136). Hillsdale, NJ: Erlbaum.

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,* 742–754.

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36,* 171–195.

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

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

Giles, H., Coupland N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1–68). Cambridge, England: Cambridge University Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding TEC: A framework for perception and action planning. *Behavioral and Brain Sciences, 24,* 849–937.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 227,* 1213–1216.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus-response compatibility. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 634–647.

Kluender, K. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 173–217). San Diego, CA: Academic Press.

Kozhevnikov, V., & Chistovich, L. (1965). *Speech: Articulation and perception.* Washington, DC: Joint Publications Research Service.

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

Kuhl, P., & Meltzoff, A. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy. The Minnesota Symposia on Child Psychology, 20,* 235–266. Hillsdale, NJ: Erlbaum.

Kuhl, P., & Meltzoff, A. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America, 100,* 2425–2438.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research, 14,* 677–709.

Lash, J. (1980). *Helen and teacher.* Reading, MA: Addison-Wesley.

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

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

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

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4,* 187–196.

Luce, R. D. (1986). *Response times.* New York: Oxford University Press.

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

MacSweeney, M., Amaro, E., Calvert, G., Campbell, R., David, A. S., McGuire, P. et al. (2000). Silent speech reading in the absence of scanner noise: An event-related fMRI study. *NeuroReport, 11,* 1729–1733.

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

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

Massaro, D. (1998). *Perceiving talking faces.* Cambridge, MA: MIT Press.

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

McHugo, G., Lanzetta, J., Sullivan, D., Masters, R., & Englis, B. (1985). Emotional reactions to a political leader's expressive displays. *Journal of Personality and Social Psychology, 49,* 1513–1529.

Meltzoff, A., & Kuhl, P. (1994). Faces and speech: Intermodal processing of biologically relevant signals in infants and adults. In D. Lewkowicz & R. Licklikr (Eds). The development of inter sensory perception: Comparative perspective (pp 335–369). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Meltzoff, A., Kuhl, P., & Moore, M. (1991). Perception, representation, and the control of action in newborns and young infants. In M. J. S. Weiss & R. Zelazo (Eds.), *Newborn attention: Biological constraints and the influence of experience* (pp. 377–411). Norwood, NJ: Ablex.

Meltzoff, A., & Moore, M. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75–78.

Meltzoff, A., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6,* 179–192.

Meltzoff, A., & Moore, K. (1999). Persons and representation: Why infant imitation is important for theories of human development. In J. Nadel & G. Butterworth (Eds.), *Imitation in infancy* (pp. 9–35). Cambridge: Cambridge University Press.

Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101,* 3241–3254.

Nittrouer, S., & Crowther, C. (2001). Coherence in children's speech perception. *Journal of the Acoustical Society of America, 110,* 2129–2140.

Pardo, J., & Fowler, C. A. (1997). Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch. *Perception & Psychophysics, 59,* 1141–1152.

Pastore, R., Schmuckler, M., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception for musical stimuli CHK. *Perception & Psychophysics, 33,* 469–474.

Pierrehumbert, J. (1979). The perception of fundamental frequency. *Journal of the Acoustical Society of America, 66,* 363–369.

Porter, R., & Castellanos, F. X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America, 67,* 1349–1356.

Porter, R., & Lubker, J. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage. *Journal of Speech and Hearing Research, 23,* 593–602.

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

Rosenblum, L. D., Schmuckler, M., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59,* 347–357.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* Unpublished doctoral dissertation, Cambridge University.

Spelke, E. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology, 15,* 626–636.

Sumby, W. H., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Yehia, H., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal tract and facial behavior. *Speech Communication, 26,* 23–44.