We have titled our presentation "The perception of phonetic gestures" as if phonetic gestures are perceived. By phonetic gestures we refer to organized movements of one or more vocal-tract structures that realize phonetic dimensions of an utterance (cf. Browman & Goldstein, 1986; in press a). An example of a gesture is bilabial closure for a stop, which includes contributions by the jaw and the upper and lower lips. Gestures are organized into larger segmental and suprasegmental groupings, and we do not intend to imply that these larger organizations are not perceived as well. We focus on gestures to emphasize a claim that, in speech, perceptual objects are fundamentally articulatory as well as linguistic.

That is, in speech perception, articulatory events have a status quite different from that of their acoustic products. The former are perceived, whereas the latter are the means (or one of the means) by which they are perceived.

A claim that phonetic gestures are perceived is not uncontroversial, of course, and there are other points of view (e.g., Massaro, 1987; Stevens & Blumstein, 1981). We do not intend to consider these other views here, however, but instead to focus on agreements and disagreements between two theoretical perspectives from which the claim is made. Accordingly, we begin by summarizing some of the evidence that, in our view, justifies it.

Phonetic gestures are perceived: Three sources of evidence

1. Correspondence failures between acoustic signal and percept: Correspondences between gestures and percept

Perhaps the most compelling evidence that gestures, and not their acoustic products, are perceptual objects is the failure of dimensions of speech percepts to correspond to obvious dimensions of the acoustic signal and their correspondence, instead, to phonetically-organized articulatory behaviors that produce the signal. We offer three examples, all of them implicating articulatory gestures as perceptual objects and the third showing most clearly that the perceived gestures are not surface articulatory movements, but rather, linguistically-organized gestures.

a. Synthetic /di/ and /du/

One example from the early work at Haskins Laboratories (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) is of synthetic /di/ and /du/. Monosyllables, such as those in Figure 1, can be synthesized that consist only of two formants. The information specifying /d/ (rather than /b/ or /g/) in both syllables is the second formant transition. These transitions are very different in the two syllables, and, extracted from their syllables, they sound very different too. Each sounds more-or-less like the frequency glide it resembles in the visible display. Neither sounds like /d/. In the context of their respective syllables, however, they sound alike and they sound like /d/.

The consonantal segments in /di/ and /du/ are produced alike too, by a constriction and release gesture of the tongue tip against the alveolar ridge of the palate. When listeners perceive the
The Perception of Phonetic Gestures

synthetic /di/ and /du/ syllables of Figure 1, their percepts correspond to the implied constriction and release gestures, not, it seems, to the context-sensitive acoustic signal.

![Figure 1. Synthetic syllables /di/ and /du/. The second formant transitions identify the initial consonant as /d/ rather than as /b/ or /g/.](image)

b. Functional equivalence of acoustic "cues"

We expect listeners to be very good at distinguishing an interval of silence from nonsilence—from a set of frequency glides, for example. Too, we expect them to distinguish acoustic signals that differ in two ways more readily than signals that differ in just one of the two ways. Both of these expectations are violated—another example of noncorrespondence—if the silence and glides are joint acoustic products of a common constriction and release gesture for a stop consonant.

Fitch, Halwes, Erickson and Liberman (1980) created synthetic syllables identified as "slit" or "split" by varying the duration of a silent interval following the fricative and manipulating the presence or absence of transitions for a bilabial stop following the silent interval. A relatively long silent interval and the presence of transitions both signal a bilabial stop, the silent interval cuing the closure and the transitions the release. Fitch et al. found that pairs of syllables differing on both cue dimensions, duration of silence and presence/absence of transitions, were either more discriminable than pairs differing in one of these ways or less discriminable depending on how the cues were combined. A syllable with a long silent interval and transitions was nearly indistinguishable from one with a longer interval and no transitions; both were identified as "split." Syllables differing in two ways are indistinguishable just when the acoustic cues that distinguish them are "functionally equivalent"—that is, they cue the same articulatory gesture. A long silent interval does not normally sound like a set of frequency glides, but it does in a context in which each specifies a consonantal constriction.

c. Perception of intonation

The findings just summarized, among others, reveal that listeners perceive gestures. Apparently, listeners do not perceive the acoustic signal per se.

Nor, however, do they perceive "raw" articulatory motions as such. Rather, they perceive linguistically-organized (phonetic) gestures. Research on the various ways in which fundamental frequency (henceforth, f₀) is perceived shows this most clearly.

Perceived intonational peak height will not, in general, correspond to the absolute rate at which the vocal folds open and close during production of the peak. Instead, perception of the peak corresponds to just those influences on the rate of opening and closing that are caused by gestures intended by the talker to affect intonational peak height. (Largely, intonational melody is implemented by contraction and relaxation of muscles of the larynx that tense the vocal folds; see, e.g., Ohala, 1978.) There are other influences on the rate of vocal fold opening and closing that may either decrease or increase f₀. Some of these influences, due to lung deflation during an expiration ("declination," Gelfer, Harris, & Baer, 1987; Gelfer, Harris, Collier, & Baer, 1985) or to segmental perturbations reflecting vowel height (e.g., Lehiste & Peterson, 1961) and obstruent voicing (e.g., Ohde, 1984), are largely or entirely automatic consequences of other things that talkers are doing (producing an utterance on an expiratory airflow, producing a close or open vowel [Honda, 1981], producing a voiced or voiceless obstruent [Ohde, 1984]; Löfqvist, Baer, McGarr, & Seider Story, in press). They do not sound like changes in pitch; rather, they sound like what they are: information for early-to-late serial position in an utterance in the case of declination (Pierrehumbert, 1979; see also Lehiste, 1982), and information for vowel height (Reinholt Peterson, 1986; Silverman, 1987) or consonant voicing (e.g., Silverman, 1986) in the case of segmental perturbations.
As we will suggest below (under "How acoustic structure may serve as information for speech perceivers"), listeners apparently use configurations of changes in different acoustic variables to recover the distinct, organized articulatory systems that implement the various linguistic dimensions of talkers' utterances. By using acoustic information in this way, listeners can recover what Liberman (1982) has called the talker's "phonetic intents."

2. Audio-visual integration of gestural information

A video display of a face mouthing /ga/ synchronized with an acoustic signal of the speaker saying /ba/ is heard most typically as "da" (MacDonald & McGurk, 1978). Subjects' identifications of syllables presented in this type of experiment reflect an integration of information from the optical and acoustic sources. Too, as Liberman (1982) points out, the integration affects what listeners experience hearing to an extent that they cannot tell what contribution to their perceptual experience is made by the acoustic signal and what by the video display. Why does integration occur? One answer is that both sources of information, the optical and the acoustic, provide information apparently about the same event of talking, and they do so by providing information about the talkers' phonetic gestures.

3. Shadowing

Listeners' latency to repeat a syllable they hear is very short—in Porter's research (Porter, 1978; Porter & Lubker, 1980), around 180 ms on average. Even though these latencies are obtained in a choice reaction time procedure (in which the vocal response required is different for different stimuli to respond), latencies approach simple reaction times (in which the same response occurs to any stimulus to respond), and they are much shorter than choice reaction times using a button press.

Why should these particular choice reaction times be so fast? Presumably, the compatibility between stimulus and response explains the fast response times. Indeed, it effectively eliminates the element of choice. If listeners perceive the talker's phonetic gestures, then the only response requiring essentially no choice at all is one that reproduces those gestures.

The motor theory

Throughout most of its history, the motor theory (e.g., Liberman et al., 1963; Liberman et al., 1967; Liberman & Mattingly, 1985; see also, Cooper, Delattre, Liberman, Borst, & Gerstman, 1952) has been the only theory of speech perception to identify the phonetic gesture as an object of perception. Here we describe the motor theory by discussing what, more precisely, the motor theorists have considered to be the object of perception, how they characterize the process of speech perception and why, recently, they have introduced the idea that speech perception is accomplished by a specialized module.

What is perceived for the motor theorist?

Coarticulation is the reason why the acoustic signal appears to correspond so badly to the sequences of phonemes that talkers intend to produce. Due to coarticulation, phonemes are produced in overlapping time frames so that the acoustic signal is everywhere (or nearly everywhere; see, e.g., Stevens & Blumstein, 1981), context-sensitive. This makes the signal a complex "code" on the phonemes of the language, not a cipher, like an alphabet. In "Perception of the speech code" (1967), Liberman and his colleagues speculated that coarticulatory "encoding" is, in part, a necessary consequence of properties of the speech articulators (their sluggishness, for example). However, in their view, coarticulation is also promoted both by the nature of phonemes themselves—that they are realized by sets of subphonemic features—and by the listener's short-term memory, which would be overtaxed by the slow transmission rate of an acoustic cipher.

In producing speech, talkers exploit the fact that the different articulators—the lips, velum, jaw, etc.—can be independently controlled. Subphonemic features, such as lip rounding, velum lowering and alveolar closure each use subsets of the articulators, often just one; therefore, more than one feature can be produced at a time. Speech can be produced at rapid rates by allowing "parallel transmission" of the subphonemic features of different phonemes. This increases the transmission rates for listeners, but it also creates much of the encoding that is considered responsible for the apparent lack of invariance between acoustic and phonetic segments.
features. To explain why “perception mirrors articulation more closely than sound” (p. 453) and (yet) achieves recovery of discrete unencoded phonemes, the motor theorists proposed as a first hypothesis that perceivers somehow access their speech-motor systems in perception and that the percept they achieve corresponds to a stage in production before encoding of the speech segments takes place. In “Perception of the speech code,” the stage was one in which “motor commands” to the muscles were selected to implement subphonemic features. In “The motor theory revised,” (Liberman & Mattingly, 1985), a revision to the theory reflects developments in our understanding of motor control. Evidence suggests that activities of the vocal tract are products of functional couplings among articulators (e.g., Folkins & Abbs, 1975, 1976; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984), which produce gestures as defined earlier, not independent movements of the articulators identified with subphonemic features in “Perception of the speech code.” In “The motor theory revised,” control structures for gestures have replaced motor commands for subphonemic features as invariants of production and as objects of perception for listeners. In recovering abstract gestures, processes of speech perception yield quite different kinds of perceptual objects than general auditory perception. In auditory perception, more generally, according to Liberman and Mattingly, listeners hear the signal as “ordinary sound” (p. 6); that is, they hear the acoustic signal as such. In other publications, Mattingly and Liberman (1988) refer to this apparently more straightforward perceptual object as “homomorphic” in contrast to objects of speech perception which are “heteromorphic.” An example they offer of homomorphic auditory perception is perception of isolated formant transitions which sound like the frequency glides they resemble in a spectrographic display.

How perception takes place in the motor theory

In the motor theory, listeners use “analysis by synthesis” to recover phonetic gestures from the encoded, informationally-impoverished acoustic signal. This aspect of the theory has never been worked out in detail. However, in general, analysis by synthesis consists in analyzing a signal by guessing how the signal might have been produced (e.g., Stevens, 1960; Stevens & Halle, 1964). Liberman and Mattingly refer to an “internal, innately specified vocal-tract synthesizer that incorporates complete information about the anatomical and physiological characteristics of the vocal tract and also about the articulatory and acoustic consequences of linguistically significant gestures” (p. 26). The synthesizer computes candidate gestures and then determines which of those gestures, in combination with others identified as ongoing in the vocal-tract could account for the acoustic signal.

Speech perception as modular

If speech perception does involve accessing the speech-motor system, then it must indeed be special and quite distinct from general auditory perception. It is special in its objects of perception, in the kinds of processes applied to the acoustic signal, and presumably in the neural systems dedicated to those processes as well. Liberman and Mattingly propose that speech perception is achieved by a specialized module.

A module (Fodor, 1983) is a cognitive system that tends to be narrowly specialized (“domain specific”), using computations that are special (“eccentric”) to its domain; it is computationally autonomous (so that different systems do not compete for resources) and prototypically is associated with a distinct neural substrate. In addition, modules tend to be “informationally encapsulated,” bringing to bear on the processing they do only some of the relevant information the perceiver may have; in particular, processing of “input” (perceptual) systems—prime examples of modules—is protected early on from bias by “top-down” information.

The speech perceptual system of the motor theory has all of these characteristics. It is narrowly specialized and its perception-production link is eccentric; moreover, it is associated with a specialized neural substrate (e.g., Kimura, 1961). In addition, as the remarkable phenomenon of duplex perception (e.g., Liberman, Isenberg, & Rakerd, 1981; Rand, 1974) suggests, the speech
perceiving system is autonomous and informationally encapsulated.

In duplex perception as it is typically investigated (e.g., Liberman et al., 1981; Mann & Liberman, 1983; Repp, Milburn, & Ashkenas, 1983), most of an acoustic CV syllable (the “base” at the left of Figure 2) is presented to one ear while the remainder, generally a formant transition (either of the “chirps” on the right side of Figure 2) is presented to the other ear. Heard in isolation, the base is ambiguous between “da” and “ga,” but listeners generally report hearing “da.” (It was identified as “da” 87% of the time in the study by Repp et al., 1983.) In isolation, the chirps sound like the frequency glides they resemble; they do not sound speech-like. Presented dichotically, listeners integrate the chirp and the base, hearing the integrated “da” or “ga” in the ear receiving the base. Remarkably, in addition, they hear the chirp in the other ear. Researchers who have investigated duplex perception describe it as perception of the same part of an acoustic signal in two ways simultaneously. If that characterization is correct, it implies strongly that the percepts are outputs of two distinct and autonomous perceptual systems, one specialized for speech and the other perhaps general to other acoustic signals.

A striking characteristic of speech perceptual systems that integrate syllable fragments presented to different ears is their imperviousness to information in the spatial separation of the fragments that they cannot possibly be part of the same spoken syllable—an instance, perhaps, of information encapsulation.

In recent work, Mattingly and Liberman (in press) have revised, or expanded on, Fodor's view of modules by proposing a distinction between “closed” and “open” modules. Closed modules, including the speech module and a sound-localization module, for example, are narrowly specialized as Fodor has characterized modules more generally. In addition (among other special properties), they yield heteromorphic percepts—that is, percepts whose dimensions are not those of the proximal stimulation. Although Mattingly and Liberman characterize the heteromorphic percept in this way—in terms of what it does not conform to, it appears that the heteromorphic percept can be characterized in a more positive way as well. The dimensions of heteromorphic percepts are those of distal events, not of proximal stimulation. The speech module renders phonetic gestures; the sound-localization module renders location in space. By contrast, open modules are sort of “everything-else” perceptual systems.

![Figure 2](image.png)

Figure 2. Stimuli that yield duplex perception. The base is presented to one ear and the third formants to another. In the ear to which the base is presented, listeners hear the syllable specified jointly by the base and the transitions; in the other ear, they hear the transitions as frequency glides. (Figure adapted from Whalen & Liberman, 1987).
An open auditory-perceptual module is responsible for perception of most sounds in the environment. According to the theory, outputs of open modules are homomorphic.

In the context of this account of auditory perception, the conditions under which duplex perception is studied are seen as somehow tricking the open module into providing a percept of the isolated formant transition even though the transition is also being perceived by the speech module. Accordingly, two percepts are provided for one acoustic fragment; one percept is homomorphic and the other is heteromorphic.

Prospectus

Our brief overview of the motor theory obviously cannot do justice to it. In our view, it is, to date, superior to other theories of speech perception in at least two major respects: in its ability to handle the full range of behavioral findings on speech perception—in particular, of course, the evidence that listeners recover phonetic gestures—and in having developed its account of speech in the context of a more general theory of biological specializations for perception.

Our purpose here, however, is not just to praise the theory, but to challenge it as well, with the further aim of provoking the motor theorists either to buttress their theory where it appears to us vulnerable, or else to revise it further.

We will raise three general questions from the perspective of our own, direct-realist theory (Fowler, 1986a,b; Rosenblum, 1987). First we question the inference from evidence that perceivers recover phonetic gestures that the listener's own speech-motor system plays a role in perception. The nature of the challenge we mount to this inference leads to a second one. We question the idea that, whereas a specialized speech module—and other closed modules—render heteromorphic percepts, other percepts are homomorphic. Finally, we challenge the idea in any case that duplex perception reveals that speech perception is achieved by a closed module.

Standing behind all of these specific questions we raise about claims of the motor theory is a general issue that needs to be confronted by all of us who study speech perception and, for that matter, perception more generally. The issue is one of determining when behavioral data warrant inferences being drawn about perceptual processes taking place inside perceivers and when the data deserve accounting instead in terms of the nature of events taking place publicly when something is perceived.

Does perceptual recovery of phonetic gestures implicate the listener's speech motor system?

In our view, the evidence that perceivers recover phonetic gestures in speech perception is incontrovertible and any theory of speech perception is inadequate unless it can provide a unified account of those findings. However, the motor theorists have drawn an inference from these findings that, we argue, is not warranted by the general observation that listeners recover gestures. The inference is that recovery of gestures implies access by the perceiver to his own speech-motor system. It is notable, perhaps, that, in neither “Perception of the speech code” nor “The motor theory revised,” do Liberman and his colleagues offer any evidence in support of this claim except evidence that listeners recover gestures (and that human left-cerebral hemispheres are specialized for speech and especially for phonetic perception [e.g., Kimura, 1961; Liberman, 1974; Studdert-Kennedy & Shankweiler, 1970]).

There is another way to explain why listeners recover phonetic gestures. It is that phonetic gestures are among the “distal events” that occur when speech is perceived and that perception universally involves recovery of distal events from information in proximal stimulation.

Distal events universally are perceptual objects: Proximal stimuli universally are not.

Consider first visual perception observed from outside the perceiver. Visual perceivers recover properties of objects and events in their environment (“distal events”). They can do so, in part, because the environment supplies information about the objects and events in a form that their perceptual systems can use. Light reflects from objects and events, which structure it lawfully; given a distal event and light from some source, the reflected light must have the structure that it has. To the extent that the structure in the light is also specific to the properties of a distal event that caused it, it can serve as information to a perceiver about its distal source. The reflected light (“proximal stimulation”) has another property that permits it its central role in perception. It can stimulate the visual system of a perceiver and thereby impart its structure to it. From there, the perceiver can use the structure as information for distal-event perception.
The reflected light does not provide information to the visual system by picturing the world. Information in reflected light for "looming" (that is, for an object on a collision course with the perceiver's head), for example, is a certain manner of expansion of the contours of the object's reflection in the light, that progressively covers the contours of optical reflections of immobile parts of the perceiver's environment. When an object looms, it does not grow; it approaches. However, its optical reflection grows, and, confronted with such an optic array, perceivers (from fiddler crabs to kittens to rhesus monkeys to humans [Schiff, 1965; Schiff, Caviness, & Gibson, 1962]) behave as if they perceive an object on a collision course; that is, they try to avoid it.

Two related conclusions from this characterization of visual perception are first that observers see distal events based on information about them in proximal stimulation and second that, in Mattingly and Liberman's terms, visual perception therefore is quite generally heteromorphic. It is not merely heteromorphic in respect to those aspects of stimulation handled by closed modules (for example, one that recovers depth information from binocular disparity); it is generally the case that the dimensions of the percept correspond with dimensions of distal objects and events and not necessarily with those of a distal-event-free description of the proximal stimulation.7

Auditory perception is analogous to visual perception in its general character, viewed, once again from outside the perceiver. Consider any sounding object, a ringing bell, for example. The ringing bell is a "distal event" that structures an acoustic signal. The structuring of the air by the bell is lawful and, to the extent that it also tends to be specific to its distal source, the structure can provide information about the source to a sensitive perceiver. Like reflected light, the acoustic signal (the proximal stimulation) in fact has two critical properties that allow it to play a central role in perception. It is lawfully structured by some distal event and it can stimulate the auditory system of perceivers, thereby imparting its structure to it. The perceiver then can use the structure as information for its source.

As for structure in reflected light, structure in an acoustic signal does not resemble the sound-producing source in any way. Accordingly, if auditory perception works similarly to visual perception—that is, if perceivers use structure in acoustic signals to recover their distal sources, then auditory percepts, like visual percepts will be heteromorphic.

Liberman and Mattingly (1985; Mattingly & Liberman, 1988) suggest, however, that in general, auditory perceptions are homomorphic. We agree that our intuitions are less clear here than they are in the case of visual perception. However, it is an empirical question whether dimensions of listeners' percepts are better explained in terms of dimensions of distal events or of a distal-event free description of proximal stimulation. To date the question is untested, however; for whatever reason, researchers who study auditory perception rarely study perception of natural sound-producing events (see, however, Repp, 1987; VanDerVeer, 1979; Warren & Verbrugge, 1984).

Now consider speech perception. In speech, the distal event—at least the event in the environment that structures the acoustic speech signal—is the moving vocal tract. If, as we propose, the vocal tract produces phonetic gestures, then the distal event is, at the same time, the set of phonetic gestures that compose the talker's spoken message. The proximal stimulus is the acoustic signal, lawfully structured by movement in the vocal tract. To the extent that the structure in the signal also tends to be specific to the events that caused it, it can serve as information about those events to sensitive perceivers. The information that proximal stimulation provides will be about the phonetic gestures of the vocal tract. Accordingly, if speech perception works like visual perception, then recovery of phonetic gestures is not eccentric and does not require eccentric processing by a speech module. It is, instead, yet another instance of recovery of distal events by means of lawfully-generated structure in proximal stimulation.

The general point we hope to make is that, arguably, all perception is heteromorphic, with dimensions of percepts always corresponding to those of distal events, not to distal-event free descriptions of proximal stimuli. Speech is not special in that regard. A more specific point is that even if evidence were to show that speech perceivers do access their speech-motor systems, that perceptual process would not be needed to provide the reason why listeners' percepts are heteromorphic. The reason percepts are heteromorphic is that perceivers universally use proximal stimuli as information about events taking place in the world; they do not use them as perceptual objects per se.
Are phonetic gestures public or private?

Although in "Perception of the speech code" and "The motor theory revised," evidence that listeners recover gestures is the only evidence cited in favor of the view that perceivers access their speech motor systems, that evidence is not the only reason why the motor theorists and other theorists invoke a construct inside the perceiver rather than the proximal stimulation outside to explain why the percep has the character it does. A very important reason why, for the motor theorists, the proximal stimulation is not by itself sufficient to specify phonetic gestures is that, in their view, phonetic gestures are abstract control structures corresponding to the speakers' intentions, but not to the movements actually taking place in the vocal tract. If phonetic gestures aren't "out there" in the vocal tract, then they cannot be analogous to other distal events, because they cannot, themselves, lawfully structure the acoustic signal.

In our view, this characterization of phonetic gestures is mistaken, however. We can identify two considerations that appear to support it, but we find neither convincing. One is that any gesture of the vocal tract is merely a token action. Yet perceivers do not just recognize the token, they recognize it as a member of a larger linguistically-significant category. That seems to localize the thing perceived in the mind of the perceiver, not in the mouth of the talker. More than that, the same collections of token gestures may be identified as tokens of different categories by speakers of different languages. (So, for example, speakers of English may identify a voiceless unaspirated stop in stressed syllable-initial position as a /b/, whereas speakers of languages in which voiceless unaspirated stops can appear stressed-syllable initially may identify it as an instance of a /p/.) Here, it seems, the information for category membership cannot possibly be in the gestures themselves or in the proximal stimulation; it must be in the head of the perceiver. The second consideration is that coarticulation, by most accounts, prevents nondestructive realization of phonetic gestures in the vocal tract. We briefly address both considerations.

Yet another analogy: There are chairs in the world that do not look very much like prototypical chairs. Recognizing them as chairs may require learning how people typically use them (learning their "proper function" in Millikan's terms [1984]). By most accounts, learning involves some enduring change inside the perceiver. Notice, however, that even if it does, what makes the token chair a chair remains its properties and its use in the world prototypically as a chair. Too, whatever perceivers may learn about that chair and about chairs in general is only what they learn; the chair itself and the means by which its type-hood can be identified remain unquestionably out there in the world. Phonetic gestures and phonetic segments are like chairs (in this respect). Token instances of bilabial closure are members of a type because the tokens all are products of a common coupling among jaw and lips realized in the vocal tract of talkers who achieve bilabial closure. Instances of bilabial closure in stressed-syllable-initial position that have a particular timing relation to a glottal opening gesture are tokens of a phonological category, /b/, in some languages and of a different category, /p/, in others because of the different ways that they are deployed by members of the different language communities. That differential deployment is what allowed descriptive linguists to identify members of phonemic categories as such, and presumably it is also what allows language learners to acquire the phonological categories of their native language. By most accounts, when language learners discover the categories of their language, the learning involves enduring changes inside the learner. However, even if it does, it is no more the case that the phonetic gestures or the phonetic segments move inside the mind than it is that chairs move inside when we learn how to recognize them as such. What we have learned is what we know about chairs and phonetic segments; it is not the chairs or the phonetic segments themselves. They remain outside.

Turning to coarticulation, it is described in the motor theory as "encoding," by Ohala (e.g., 1981) as "distortion," by Daniloff and Hammarberg (1973) as "assimilation" and by Hockett (1955) as "smashing" and "rubbing together" of phonetic segments (in the way that raw eggs would be smashed and rubbed together were they sent through a wringer). None of these characterizations is warranted, however. Coarticulation may instead be characterized as gestural layering—a temporally staggered realization of gestures that sometimes do and sometimes do not share one or more articulators.

In fact, this kind of gestural layering occurs commonly in motor behavior. When someone walks, the movement of his or her arm is seen as pendular. However, the surface movement is a complex (layered) vector including not only the
swing of the arm, but also movement of the whole body in the direction of locomotion. This layering is not described as "encoding," "distortion" or even as assimilation of the arm movement to the movement of the body as a whole. And for good reason; that is not what it is. The movement reflects a convergence of forces of movement on a body segment. The forces are separate for the walker, information in proximal stimulation allows their parsing (Johansson, 1973), and perceivers detect their separation.

There is evidence already suggesting that at least some of coarticulation is gestural layering (Carney & Moll, 1971; Öhman, 1966; also see Browman & Goldstein, in press a), not encoding or distortion or assimilation. There is also convincing evidence that perceivers recover separate gestures more-or-less in the way that Johansson suggests they recover separate sources of movement of body segments in perception of locomotion. Listeners use information for a coarticulating segment that is present in the domain of another segment as information for the coarticulating segment itself (e.g., Fowler, 1984; Mann, 1980; Whalen, 1984); they do not hear the coarticulated segment as assimilated or, apparently, as distorted or encoded (Fowler, 1981; 1984; Fowler & Smith, 1986).

Our colleagues Catherine Browman and Louis Goldstein (1985, 1986, in press a,b) have proposed that phonetic primitives of languages are gestural, not abstract featural. Our colleague Elliot Saltzman (1986; Saltzman & Kelso, 1987; see also, Kelso, Saltzman, & Tuller, 1986) is developing a model that implements phonetic gestures as functional couplings among the articulators and that realizes the gestural layering characteristic of coarticulation. To the extent that these approaches both succeed, they will show that phonetic gestures—speakers' intentions—can be realized in the vocal tract nondestructively, and hence can structure acoustic signals directly.

Do listeners need an innate vocal tract synthesizer to recognize acoustic reflections of phonetic gestures? Although it might seem to help, it cannot be necessary, because there is no analogous way to explain how observers recognize most distal events from their optical reflections. Somehow the acoustic and optical reflections of a source must identify the source on their own. In some instances, we begin to understand the means by which acoustic patternings can specify their gestural sources. We consider one such instance next.

How acoustic structure may serve as information for gestures.

We return to the example previously described of listeners' perception of those linguistic dimensions of an utterance that are cued in some way by variation in f0. A variety of linguistic and paralinguistic properties of an utterance have converging effects on f0. Yet listeners pull apart those effects in perception.

What guides the listeners' factoring of converging effects of f0? Presumably, it is the configuration of acoustic products of the several gestures that have effects, among others, on f0. Intonational peaks are local changes in an f0 contour that are effected by means that, to a first approximation, only affect f0; they are produced, largely, by contraction and relaxation of muscles that stretch or shorten the vocal folds (e.g., Ohala, 1978). In contrast, declination is a global change in f0 that, excepting the initial peak in a sentence, tracks the decline in subglottal pressure (Gelfer et al., 1985; Gelfer et al., 1987). Subglottal pressure affects not only f0, but amplitude as well, and several researchers have noticed that amplitude declines in parallel with f0 and resets when f0 resets at major syntactic boundaries (e.g., Breckenridge, 1977; Maeda, 1976). The parallel decline in amplitude and f0 constitutes information that pinpoints the mechanism behind the f0 decline—gradual lung deflation, incompletely offset by expiratory-muscle activity. That mechanism is distinct from the mechanism by which intonational peaks are produced. Evidence that listeners pull apart the two effects on f0 (Pierrehumbert, 1979; Silverman, 1987) suggests that they are sensitive to the distinct gestural sources of these effects on f0.

By the same token, f0 perturbations due to height differences among vowels are not confused by listeners with information for intonational peak height even though f0 differences due to vowel height are local, like intonational peaks, and are similar in magnitude to differences among intonational peaks in a sentence (Silverman, 1987). The mechanisms for the two effects on f0 are different, and, apparently, listeners are sensitive to that. Honda (1981) shows a strong correlation between activity of the genioglossus muscle, active in pulling the root of the tongue forward for high vowels, and intrinsic f0 of vowels. Posterior fibers of the genioglossus muscle insert into the hyoid bone of the larynx. Therefore, contraction of the genioglossus may pull the hyoid
forward, rotating the thyroid cartilage to which the vocal folds attach, and thereby may stretch the vocal folds. Other acoustic consequences of genioglossus contraction, of course, are changes in the resonances of the vocal tract, which reflect movement of the tongue. These changes, along with those in f₀ (and perhaps others as well) pinpoint a phonetic gesture that achieves a vowel-specific change in vocal-tract shape. If listeners can use that configuration of acoustic reflections of tongue-movement (or, more likely, of coordinated tongue and jaw movement) to recover the vocalic gesture, then they can pull effects on f₀ of the vocalic gesture from those for the intonation contour that cooccur with them.

Listeners do just that. In sentence pairs such as “They only feast before fasting,” with intonational peaks on the “fast” syllables, listeners require a higher peak on “feast” in the second sentence than on “fast” in the first sentence in order to hear the first peak of each sentence as higher than the second (Silverman, 1987). Compatibly, among steady-state vowels on the same f₀, more open vowels sound higher in pitch than more closed vowels (Stoll, 1984). Intrinsic f₀ of vowels does not contribute to perception of an intonation contour or to perception of pitch. But it is not thrown away by perceivers either. Rather, along with spectral information for vowel height, it serves as information for vowel height (Reinholt Peterson, 1986).

We will not review the literature on listeners’ use of f₀ perturbations due to obstructive voicing except to say that it reveals the same picture of the perceiver as the literature on listeners’ use of information for vowel height (for a description of the f₀ perturbations: Ohde, 1984; for studies of listeners’ use of the perturbations: Abramson & Lisker, 1985; Fujimura, 1971; Haggard, Ambler, & Callow, 1970; Silverman, 1986; for evidence that listeners can detect the perturbations when they are superimposed on intonation contours: Silverman, 1986). As the motor theory and the theory of direct perception both claim, listeners’ percepts do not correspond to superficial aspects of the acoustic signal. They correspond to gestures, signaled, we propose, by configurations of acoustic reflections of those gestures.

**Does duplex perception reveal a closed speech module?**

We return to the phenomenon of duplex perception and consider whether it does convincingly reveal distinct closed and open modules for speech perception and general auditory perception respectively. As noted earlier, duplex perception is obtained, typically, when most of the acoustic structure of a synthetic syllable is presented to one ear, and the remainder—usually a formant transition—is presented to the other ear (refer to Figure 2). In such instances, listeners hear two things. In the ear that gets most of the signal, they hear a coherent syllable, the identity of which is determined by the transition presented to the other ear. At the same time, they hear a distinct, non-speech ‘chirp’ in the ear receiving the transition. The percept is duplex—the transition is heard as a critical part of a speech syllable, hypothetically as a result of it’s being processed by the speech module, and it is heard simultaneously as a non-speech chirp, hypothetically as a result of its being processed also by an open auditory module (Liberman & Mattingly, 1985). Here we offer a different interpretation of the findings.

Whalen and Liberman (1987) have recently shown that duplex perception can occur with monaural or diotic presentation of the base and transition of a syllable. In this case, duplexity is attained by increasing the intensity of the third formant transition relative to the base until listeners hear both an integrated syllable (/da/ or /ga/ depending on the transition) and a non-speech ‘whistle’ (sinusoids were used for transitions). In the experiment, subjects first were asked to label the isolated sinusoidal transitions as “da” or “ga”. Although they were consistent in their labeling, reliably identifying one whistle as “da” and the other as “ga,” their overall accuracy was not greater than chance. About half the subjects were consistently right and the remainder were consistently wrong. The whistles are distinct, but they do not sound like “da” or “ga.” Next, Whalen and Liberman determined ‘duplexity thresholds’ for listeners. They presented the base and one of the sinusoids simultaneously and gave listeners control over the intensity of the sinusoid. Listeners adjusted its intensity to the point where they just heard a whistle. At threshold, subjects were able to match these duplex sinusoids to sinusoids presented in isolation. Finally, subjects were asked to identify the integrated speech syllables as “da” or “ga” at sinusoid intensities both 6 dB above and 4 dB below the duplexity threshold. Subjects were consistently good at these tasks yielding accuracy scores well above 90%.

In the absence of any transition, listeners hear only the base and identify it as “da” most of the
time. When a sinusoidal transition is present but at intensities below the duplexity threshold, subjects hear only the unambiguous syllable (“da” or “ga” depending on the transition). Finally, when the intensity of the transition reaches and exceeds the duplexity threshold, subjects hear the “da” or “ga” and they hear a whistle at the same time: i.e., the transition is duplexed.

This experiment reveals two new aspects of the duplex phenomenon. One is that getting a duplex percept requires a sufficiently high intensity of the transition. A second is that the transition integrates with the syllable at intensities below the duplexity threshold. Based on this latter finding, Whalen and Liberman conclude that processing of the sinusoid as speech has priority. It is as if a (neurally-encoded) acoustic signal must first pass through the speech module at which point portions of the signal that specify speech events are peeled off. After the speech module takes its part, any residual is passed on to the auditory module where it is perceived homomorphically. Mattingly and Liberman (1988) refer to this priority of speech processing as “preemptiveness,” and Whalen and Liberman (1987) suggest that it reflects the “profound biological significance of speech.”

There is another way to look at these findings, however. They suggest that duplex perception does not, in fact, involve the same acoustic fragment being perceived in two ways simultaneously. Rather part of the transition integrates with the syllable and the remainder is heard as a whistle or chirp. As Whalen and Liberman themselves describe it:

“... the phonetic mode takes precedence in processing the transitions, using them for its special linguistic purposes until, having appropriated its share, it passes the remainder to be perceived by the nonspeech system as auditory whistles.”
(Whalen & Liberman, 1987, p. 171; our italics).

This is important, because in earlier reports of duplex perception, it was the apparent perception of the transition in two different ways at once that was considered strong evidence favoring two distinct perceptual systems, one for speech and one for general auditory perception. In addition, research to date has only looked for preemptiveness using speech syllables. Accordingly, it is premature to conclude that speech especially is preemptive. Possibly acoustic fragments integrate preferentially whenever the integrated signal specifies some coherent sound-producing event.

We have recently looked for duplex perception in perception of nonspeech sounds (Fowler and Rosenblum, in press). We predicted that it would be possible to observe duplex perception and preemptiveness whenever two conditions are met: 1) A pair of acoustic fragments is presented that, integrated, specify a natural distal event; and 2) one of the fragments is unnaturally intense. Under these conditions, the integrated event should be preemptive and the intense fragment should be duplexed regardless of the type of natural sound-producing event that is involved, whether it is speech or non-speech, and whether it is profoundly biologically significant or biologically trivial.

There have been other attempts to get duplex perception for nonspeech sounds. All the ones of which we are aware have used musical stimuli, however (e.g., Collins, 1985; Pastore, Schmuckler, Rosenblum, & Szczesuil, 1983). We chose not to use musical stimuli because it might be argued that there is a music module. (Music is universal among human cultures, and there is evidence for an anatomical specialization of the brain for music perception (e.g., Shapiro, Grossman, & Gardner, 1981). These considerations led us to choose a non-speech event that evolution could not have anticipated. We chose an event involving a recent human artifact: a slamming metal door.

To generate our stimuli, we recorded a heavy metal door (of a sound-attenuating booth) being slammed shut. A spectrogram of this sound can be seen in Figure 3a. To produce our ‘chirp,’ we high-pass filtered the signal above 3000 Hz. To produce a ‘base,’ we low-pass filtered the original signal, also at 3000 Hz (see bottom panels of Figure 3). To us, the high-passed ‘chirp’ sounded like a can of rice being shaken, while the low-pass-filtered base sounded like a wooden door being slammed shut. (That is, the clanging of the metal door was largely absent.)

We asked sixteen listeners to identify the original metal door, the base and the chirp. The modal identifications of the metal door and the base included mention of a door; however, less than half the subjects reported hearing a door slam. Even so, essentially all of the identifications involved hard collisions of some sort (e.g., boots clomping on stairs, shovel banged on sidewalk). In contrast, no subject identified the chirp as a door sound, and no identifications described hard collisions. Most identifications of the chirp referred to an event involving shaking (tambourine, maracas, castanets, keys).
Given our metal-door chirp and a high-pass-filtered wooden door slam, subjects could not identify which was a filtered metal door slam and which a filtered wooden door slam. Subjects were consistent in their labeling judgments, identifying one of the chirps as a metal door and the other as a wooden door. However, overall, more of them were consistently wrong than right. On average, they identified the metal-door chirp as the sound of a metal door 31% of the time and the wooden-door chirp as a metal door sound 79% of the time.

To test for duplex perception and preemptiveness, we first trained subjects to identify the unfiltered door sound as a "metal door," the base as a "wooden door" and the upper frequencies of the door as a shaking sound. Next we tested them on stimuli created from the base and the chirp. We created 15 different diotic stimuli. All included the base, and almost all included the metal-door chirp. The stimuli differed in the intensity of the chirp. The chirp was attenuated or amplified by multiplying its digitized voltages by the following values: 0, .05, .1, .15, .2, .9, .95, 1, 1.05, 1.1, 4, 4.5, 5, 5.5 and 6. That is, there were 15 different intensities falling into three ranges; five were well below the natural intensity relationship of the chirp to the base, five were in the range of the natural intensity relation, and five were well above it. Three tokens of each of these stimuli were presented to subjects diotically in a randomized order. Listeners were told that they might hear one of the stimuli, metal door, wooden door or shaking sound, or sometimes two of them simultaneously, on each trial. They were to indicate what they heard on each trial by writing an identifying letter or pair of letters on their answer sheets.

In our analyses, we have grouped responses to the 15 stimuli into three blocks of five. In Figure 4, we have labeled these Intensity Conditions low, medium, and high. Figure 4 presents the results as percentages of responses in the various response categories across the three intensity conditions. We show only the three most interesting (and most frequent) responses. The figure shows that the most frequent response for the low intensity condition is 'wooden door,' the label we asked subjects to use when they heard the base. The most frequent response for the medium condition is 'metal door,' the label we asked subjects to use when they heard the metal door slam. The preferred response for the high-intensity block of stimuli is overwhelmingly 'metal door + chirp,' the response that indicates a duplex percept. The changes in response frequency over the three intensity conditions for each response type are highly significant.

Figure 3. Display of stimuli used to obtain duplex perception of closing-door sounds.
Our results can be summarized as follows. First, at very low intensities of the upper frequencies of the door, subjects hear the base only. When the 'chirp' is amplified to an intensity at or near its natural intensity relation to the base, subjects report hearing a metal door the majority of the time. Further amplification of the 'chirp' leads to reports of the metal door and a separate shaking sound. The percept is duplex, and the metal door slam is preemptive.

There are several additional tests that we must run to determine whether our door slams, in fact, are perceived analogously to speech syllables in procedures revealing duplex perception. If we can show that they are, then we will conclude that an account of our findings that invokes a closed module is inappropriate. Evolution is unlikely to have anticipated metal door slams, and metal-door slams aren't profoundly biologically significant. We suggest alternatively that preemptiveness occurs when a chirp fills a "hole" in a simultaneously presented acoustic signal so that, together the two parts of the signal specify some sound-producing distal event. If anything is left over after the hole is filled, the remainder is heard as separate.

**Summary and concluding remarks**

We have raised three challenges to the motor theory. We challenge their inference from evidence that phonetic gestures are perceived that speech perception involves access to the talker's own motor system. The basis for our challenge is a claim that dimensions of percepts always conform to those of distal events even in cases where access to an internal synthesizer for the events is unlikely. A second, related, challenge is to the idea that only some percepts are heteromorphic—just those for which we have evolved closed modules. When Liberman and Mattingly write that speech perception is heteromorphic, they mean heteromorphic with respect to structure in proximal stimulation, but they always mean as well that the percept is homomorphic with respect to dimensions of the distal source of the proximal stimulation. We argue that percepts are *generally* heteromorphic with respect to structure in proximal stimulation, but, whether they are or
The Perception of Phonetic Gestures

not, they are always homomorphic with respect to dimensions of distal events. Finally, we challenge the interpretation of duplex perception that ascribes it to simultaneous processing of one part of an acoustic signal by two modules. We suggest, instead, that duplex perception reflects the listener's parsing of acoustic structure into disjoint parts that specify, insofar as the acoustic structure permits, coherent distal events.

Where (in our view) does this leave the motor theory? It is fundamentally right in its claim that listeners perceive phonetic gestures, and also, possibly, in its claim that humans have evolved neural systems specialized for perception and production of phonetic gestures. It is wrong, we believe, specifically in its claims about what those specialized systems do, and generally in the view that closed modules must be invoked to explain why distal events are perceived.

Obviously, we prefer our own, direct-realist, theory, not so much because it handles the data better, but because, in our view, it fits better in a universal theory of perception. But however our theory may be judged in relation to the motor theory, we recognize that we would not have developed it at all in the absence of the important discoveries of the motor theorists that gestures are perceived.

REFERENCES


The Perception of Phonetic Gestures


FOOTNOTES

1 There is a small qualification to the claim that listeners cannot tell what contributions visible and audible information each have to their perceptual experience in the McGurk effect. Massaro (1987) has shown that effects of the video display can be reduced but not eliminated by instructing subjects to look at, but to ignore, the display.

2 Liberman et al. identify a cipher as a system in which each unique unit of the message maps onto a unique symbol. In contrast, in a code, the correspondence between message unit and symbol is not 1:1.

3 Liberman et al. propose to replace the more conventional view of the features of a phone (for example, that of Jakobson, Fant & Halle, 1951) with one of features as “implicit instructions to separate and independent parts of the motor machinery” (p. 446).

4 With one apparent slip on page 2: “The objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands . . . .”

5 One can certainly challenge the idea that listeners recover the very gestures that occurred to produce a speech signal. Obviously, there are no gestures at all responsible for most synthetic speech or for “sine-wave speech” (e.g., Remez, Rubin, Pisoni, & Carrell, 1981) and quite different behaviors underlie a parrot’s or mynah bird’s mimicking of speech. The claim that we argue is incontrovertible is that listeners recover gestures from speech-like signals, even those generated in some other way. (We direct realists [Fowler, 1986a,b] would also argue that “misperceptions” (hearing phonetic gestures where there are none) can only occur in limited varieties of ways—the most notable being signals produced by certain mirage-producing human artifacts, such as speech synthesizers or mirage-producing birds. Another, however, possibly, includes signals produced to mimic those of normal speakers by speakers with pathologies of the vocal tract that prevent normal realization of gestures.)

6 There are two almost orthogonal perspectives from which perception can be studied. On the one hand, investigators can focus on processes inside the perceiver that take place from the time that a sense organ is stimulated until a percept is achieved or a response is made to the input. On the other hand, they can look outside the perceiver and ask what, in the environment, the organism under study perceives, what information in stimulation to the sense organs allows perception of the things perceived, and finally, whether the organisms in fact use the postulated information. Here we focus on this latter perspective, most closely associated with the work of James Gibson (e.g., 1966; 1979; Reed & Jones, 1982).

7 It is easy to find examples in which perception is homomorphic with respect to the proximal stimulation and homomorphic with respect to distal events—looming, for example. We can also think of some examples in which perception appears homomorphic with respect to proximal stimulation, but in the examples we have come up with, they are homomorphic with respect to the distal event as well (perception of a line drawn by a pencil, for example), and so there is no way to decide whether perception is of the proximal stimulation or of the distal event. We challenge the motor theorists to provide an example in which perception is homomorphic with structure in proximal stimulation that is not also homomorphic with distal event structure. These would provide convincing cases of proximal stimulation perception.

8 Bregman (1987) considers duplex perception to disconfirm his “rule of disjoint allocation” in acoustic scene analysis by listeners. According to the rule, each acoustic fragment is assigned in perception to one and only one environmental source. It seems, however, that duplex perception does not disconfirm the rule.

9 Using a more sensitive, AXB, test, however, we have found that listeners can match the metal door chirp, rather than a wooden door chirp, to the metal door slam at performance levels considerably better than chance.