Cognition, 21 (1985) 1–36

The motor theory of speech perception revised*

ALVIN M. LIBERMAN

Haskins Laboratories, University of Connecticut and Yale University

IGNATIUS G. MATTINGLY

Haskins Laboratories and University of Connecticut

Abstract

A motor theory of speech perception, initially proposed to account for results of early experiments with synthetic speech, is now extensively revised to accommodate recent findings, and to relate the assumptions of the theory to those that might be made about other perceptual modes. According to the revised theory, phonetic information is perceived in a biologically distinct system, a 'module' specialized to detect the intended gestures of the speaker that are the basis for phonetic categories. Built into the structure of this module is the unique but lawful relationship between the gestures and the acoustic patterns in which they are variously overlapped. In consequence, the module causes perception of phonetic structure without translation from preliminary auditory impressions. Thus, it is comparable to such other modules as the one that enables an animal to localize sound. Peculiar to the phonetic module are the relation between perception and production it incorporates and the fact that it must compete with other modules for the same stimulus variations.

Together with some of our colleagues, we have long been identified with a view of speech perception that is often referred to as a 'motor theory'. Not *the* motor theory, to be sure, because there are other theories of perception that, like ours, assign an important role to movement or its sources. But the

*The writing of this paper was supported by a grant to Haskins Laboratories (NIH-NICHD HD-01994). We owe a special debt to Harriet Magen for invaluable help with the relevant literature, and to Alice Dadourian for coping with an ever-changing manuscript. For their patient responses to our frequent requests for information and criticism, we thank Franklin Cooper, Jerry Fodor, Carol Fowler, Scott Kelso, Charles Liberman, Robert Remez, Bruno Repp, Arthur Samuel, Michael Studdert-Kennedy, Michael Turvey, and Douglas Whalen. We also acknowledge the insightful comments of an anonymous reviewer.

Reprint requests should be sent to: Alvin Liberman, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511, U.S.A.

0010-0277/85/$11.30

theory we are going to describe is only about speech perception, in contrast to some that deal with other perceptual processes (e.g., Berkeley, 1709; Festinger, Burnham, Ono, & Bamber, 1967) or, indeed, with all of them (e.g., Washburn, 1926; Watson, 1919). Moreover, our theory is motivated by considerations that do not necessarily apply outside the domain of speech. Yet even there we are not alone, for several theories of speech perception, being more or less 'motor', resemble ours to varying degrees (e.g., Chistovich, 1960; Dudley, 1940; Joos, 1948; Ladefoged & McKinney, 1963; Stetson, 1951). However, it is not relevant to our purposes to compare these, so, for convenience, we will refer to *our* motor theory as *the* motor theory.

We were led to the motor theory by an early finding that the acoustic patterns of synthetic speech had to be modified if an invariant phonetic percept was to be produced across different contexts (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Liberman, Delattre, & Cooper, 1952). Thus, it appeared that the objects of speech perception were not to be found at the acoustic surface. They might, however, be sought in the underlying motor processes, if it could be assumed that the acoustic variability required for an invariant percept resulted from the temporal overlap, in different contexts, of correspondingly invariant units of production. In its most general form, this aspect of the early theory survives, but there have been important revisions, including especially the one that makes perception of the motor invariant depend on a specialized phonetic mode (Liberman, 1982; Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Liberman & Studdert-Kennedy, 1978; Mattingly & Liberman, 1969). Our aim in this paper is to present further revisions, and so bring the theory up to date.

The theory

The first claim of the motor theory, as revised, is that the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations. These gestural commands are the physical reality underlying the traditional phonetic notions—for example, 'tongue backing,' 'lip rounding,' and 'jaw raising'—that provide the basis for phonetic categories. They are the elementary events of speech production and perception. Phonetic segments are simply groups of one or more of these elementary events; thus [b] consists of a labial stop gesture and [m] of that same gesture combined with a velum-lowering gesture. Phonologically, of course, the gestures themselves must be viewed as groups of features, such as 'labial,' 'stop,' 'nasal,' but these features are

attributes of the gestural events, not events as such. To perceive an utterance, then, is to perceive a specific pattern of intended gestures.

We have to say 'intended gestures,' because, for a number of reasons (coarticulation being merely the most obvious), the gestures are not directly manifested in the acoustic signal or in the observable articulatory movements. It is thus no simple matter (as we shall see in a later section) to define specific gestures rigorously or to relate them to their observable consequences. Yet, clearly, invariant gestures of some description there must be, for they are required, not merely for our particular theory of speech perception, but for *any* adequate theory of speech production.

The second claim of the theory is a corollary of the first: if speech perception and speech production share the same set of invariants, they must be intimately linked. This link, we argue, is not a learned association, a result of the fact that what people hear when they listen to speech is what they do when they speak. Rather, the link is innately specified, requiring only epigenetic development to bring it into play. On this claim, perception of the gestures occurs in a specialized mode, different in important ways from the auditory mode, responsible also for the production of phonetic structures, and part of the larger specialization for language. The adaptive function of the perceptual side of this mode, the side with which the motor theory is directly concerned, is to make the conversion from acoustic signal to gesture automatically, and so to let listeners perceive phonetic structures without mediation by (or translation from) the auditory appearances that the sounds might, on purely psychoacoustic grounds, be expected to have.

A critic might note that the gestures do produce acoustic signals, after all, and that surely it is these signals, not the gestures, which stimulate the listener's ear. What can it mean, then, to say it is the gestures, not the signals, that are perceived? Our critic might also be concerned that the theory seems at first blush to assign so special a place to speech as to make it hard to think about in normal biological terms. We should, therefore, try to forestall misunderstanding by showing that, wrong though it may be, the theory is neither logically meaningless nor biologically unthinkable.

An issue that any theory of speech perception must meet

The motor theory would be meaningless if there were, as is sometimes supposed, a one-to-one relation between acoustic patterns and gestures, for in that circumstance it would matter little whether the listener was said to perceive the one or the other. Metaphysical considerations aside, the proximal acoustic patterns might as well be the perceived distal objects. But the relationship between gesture and signal is not straightforward. The reason is

that the timing of the articulatory movements—the peripheral realizations of the gestures—is not simply related to the ordering of the gestures that is implied by the strings of symbols in phonetic transcriptions: the movements for gestures implied by a single symbol are typically not simultaneous, and the movements implied by successive symbols often overlap extensively. This coarticulation means that the changing shape of the vocal tract, and hence the resulting signal, is influenced by several gestures at the same time. Thus, the relation between gesture and signal, though certainly systematic, is systematic in a way that is peculiar to speech. In later sections of the paper we will consider how this circumstance bears on the perception of speech and its theoretical interpretation. For now, however, we wish only to justify consideration of the motor theory by identifying it as one of several choices that the complex relation between gesture and signal faces us with. For this purpose, we will describe just one aspect of the relation, that we may then use it as an example.

When coarticulation causes the signal to be influenced simultaneously by several gestures, a particular gesture will necessarily be represented by different sounds in different phonetic contexts. In a consonant–vowel syllable, for example, the acoustic pattern that contains information about the place of constriction of the consonantal gesture will vary depending on the following vowel. Such context-conditioned variation is most apparent, perhaps, in the transitions of the formants as the constriction is released. Thus, place information for a given consonant is carried by a rising transition in one vowel context and a falling transition in another (Liberman, Delattre, Cooper, & Gerstman, 1954). In isolation, these transitions sound like two different glissandi or chirps, which is just what everything we know about auditory perception leads us to expect (Mattingly, Liberman, Syrdal, & Halwes, 1971); they do not sound alike, and, just as important, neither sounds like speech. How is it, then, that, in context, they nevertheless yield the same consonant?

Auditory theories and the accounts they provide

The guiding assumption of one class of theories is that ordinary auditory processes are sufficient to explain the perception of speech; there is no need to invoke a further specialization for language, certainly not one that gives the listener access to gestures. The several members of this class differ in principle, though they are often combined in practice.

One member of the class counts two stages in the perceptual process: a first stage in which, according to principles that apply to the way we hear all sounds, the auditory appearances of the acoustic patterns are registered, followed by a second stage in which, by an act of sorting or matching to prototypes, phonetic labels are affixed (Crowder & Morton, 1969; Fujisaki

& Kawashima, 1970; Oden & Massaro, 1978; Pisoni, 1973). Just why such different acoustic patterns as the rising and falling transitions of our example deserve the same label is not explicitly rationalized, it being accounted, presumably, a characteristic of the language that the processes of sorting or matching are able to manage. Nor does the theory deal with the fact that, in appropriate contexts, these transitions support phonetic percepts but do not also produce such auditory phenomena as chirps. To the contrary, indeed, it is sometimes made explicit that the auditory stage is actually available for use in discrimination. Such availability is not always apparent because the casual (or forgetful) listener is assumed to rely on the categorical labels, which persist in memory, rather than on the context-sensitive auditory impressions, which do not; but training or the use of more sensitive psychophysical methods is said to give better access to the auditory stage and thus to the stimulus variations—including, presumably, the differences in formant transition—that the labels ignore (Carney, Widin, & Viemeister, 1977; Pisoni & Tash, 1974; Samuel, 1977).

Another member of the class of auditory theories avoids the problem of context-conditioned variation by denying its importance. According to this theory, speech perception relies on there being at least a brief period during each speech sound when its short-time spectrum is reliably distinct from those of other speech sounds. For an initial stop in a stressed syllable, for example, this period includes the burst and the first 10 ms. after the onset of voicing (Stevens & Blumstein, 1978). That a listener is nevertheless able to identify speech sounds from which these invariant attributes have been removed is explained by the claim that, in natural speech, they are sometimes missing or distorted, so that the child must learn to make use of secondary, context-conditioned attributes, such as formant transitions, which ordinarily co-occur with the primary, invariant attributes (Cole & Scott, 1974). Thus, presumably, the different-sounding chirps develop in perception to become the same-sounding (nonchirpy) phonetic element with which they have been associated.

The remaining member of this class of theories is the most thoroughly auditory of all. By its terms, the very processes of phonetic classification depend directly on properties of the auditory system, properties so independent of language as to be found, perhaps, in all mammals (Kuhl, 1981; Miller, 1977; Stevens, 1975). As described most commonly in the literature, this version of the auditory theory takes the perceived boundary between one phonetic category and another to correspond to a naturally-occurring discontinuity in perception of the relevant acoustic continuum. There is thus no first stage in which the (often) different auditory appearances are available, nor is there a process of learned equivalence. An example is the claim that the

distinction between voiced and voiceless stops—normally cued by a complex of acoustic differences caused by differences in the phonetic variable known as voice-onset-time—depends on an auditory discontinuity in sensitivity to temporal relations among components of the signal (Kuhl & Miller, 1975; Pisoni, 1977). Another is the suggestion that the boundary between fricative and affricate on a rise-time continuum is the same as the rise-time boundary in the analogous nonspeech case—that is, the boundary that separates the nonspeech percepts 'pluck' and 'bow' (Cutting & Rosner, 1974; but see Rosen & Howell, 1981). To account for the fact that such discontinuities move as a function of phonetic context or rate of articulation, one can add the assumption that the several components of the acoustic signal give rise to interactions of a purely auditory sort (Hillenbrand, 1984; but see Summerfield, 1982). As for the rising and falling formant transitions of our earlier example, some such assumption of auditory interaction (between the transitions and the remainder of the acoustic pattern) would presumably be offered to account for the fact that they sound like two different glissandi in isolation, but as the same (non-glissando-like) consonant in the context of the acoustic syllable. The clear implication of this theory is that, for all phonetic contexts and for every one of the many acoustic cues that are known to be of consequence for each phonetic segment, the motivation for articulatory and coarticulatory maneuvers is to produce just those acoustic patterns that fit the language-independent characteristics of the auditory system. Thus, this last auditory theory is auditory in two ways: speech perception is governed by auditory principles, and so, too, is speech production.

The account provided by the motor theory

The motor theory offers a view radically different from the auditory theories, most obviously in the claim that speech perception is not to be explained by principles that apply to perception of sounds in general, but must rather be seen as a specialization for phonetic gestures. Incorporating a biologically based link between perception and production, this specialization prevents listeners from hearing the signal as an ordinary sound, but enables them to use the systematic, yet special, relation between signal and gesture to perceive the gesture. The relation is systematic because it results from lawful dependencies among gestures, articulator movements, vocal-tract shapes, and signal. It is special because it occurs only in speech.

Applying the motor theory to our example, we suggest what has seemed obvious since the importance of the transitions was discovered: the listener uses the systematically varying transitions as information about the coarticulation of an invariant consonant gesture with various vowels, and so perceives this gesture. Perception requires no arbitrary association of signal with phone-

tic category, and no correspondingly arbitrary progression from an auditory stage (e.g., different sounding glissandi) to a superseding phonetic label. As Studdert-Kennedy (1976) has put it, the phonetic category 'names itself'.

By way of comparison with the last of the auditory theories we described, we note that, just as this theory is in two ways auditory, the motor theory is in two ways motor. First, because it takes the proper object of phonetic perception to be a motor event. And, second, because it assumes that adaptations of the motor system for controlling the organs of the vocal tract took precedence in the evolution of speech. These adaptations made it possible, not only to produce phonetic gestures, but also to coarticulate them so that they could be produced rapidly. A perceiving system, specialized to take account of the complex acoustic consequences, developed concomitantly. Accordingly, the theory is not indifferently perceptual *or* motor, implying simply that the basis of articulation and the object of perception are the same. Rather, the emphasis is quite one-sided; therefore, the theory fully deserves the epithet 'motor'.

How the motor theory makes speech perception like other specialized perceiving systems

The specialized perceiving system that the motor theory assumes is not unique; it is, rather, one of a rather large class of special systems or 'modules'. Accordingly, one can think about it in familiar biological terms. Later, we will consider more specifically how the phonetic module fits the concept of modularity developed recently by Fodor (1983); our concern now is only to compare the phonetic module with others.

The modules we refer to have in common that they are special neural structures, designed to take advantage of a systematic but unique relation between a proximal display at the sense organ and some property of a distal object. A result in all cases is that there is not, first, a cognitive representation of the proximal pattern that is modality-general, followed by translation to a particular distal property; rather, perception of the distal property is immediate, which is to say that the module has done all the hard work. Consider auditory localization as an example. One of several cues is differences in time of arrival of particular frequency components of the signal at the two ears (see Hafter, 1984, for a review). No one would claim that the use of this cue is part of the general auditory ability to perceive, as such, the size of the time interval that separates the onsets of two different signals. Certainly, this kind of general auditory ability does exist, but it is no part of auditory localization, either psychologically or physiologically. Animals perceive the location of sounding objects only by means of neural structures specialized to take advantage of the

systematic but special relation between proximal stimulus and distal location (see, for example, Knudsen, 1984). The relation is systematic for obvious reasons; it is special because it depends on the circumstance that the animal has two ears, and that the ears are set a certain distance apart. In the case of the human, the only species for which the appropriate test can be made, there is no translation from perceived disparity in time because there is no perceived disparity.

Compare this with the voicing distinction (e.g., [ba] vs. [pa]) referred to earlier, which is cued in part by a difference in time of onset of the several formants, and which has therefore been said by some to rest on a general auditory ability to perceive temporal disparity as such (Kuhl & Miller, 1975; Pisoni, 1977). We believe, to the contrary, that the temporal disparity is only the proximal occasion for the unmediated perception of voicing, a distal gesture represented at the level of articulation by the relative timing of vocal-tract opening and start of laryngeal vibration (Lisker & Abramson, 1964). So we should expect perceptual judgments of differences in signal onset-time to have no more relevance to the voicing distinction than to auditory localization. In neither case do general auditory principles and procedures enlighten us. Nor does it help to invoke general principles of auditory interaction. The still more general principle that perception gives access to distal objects tells us only that auditory localization and speech perception work as they are supposed to; it does not tell us how. Surely the 'how' is to be found, not by studying perception, even auditory perception, in general, but only by studying auditory localization and speech perception in particular. Both are special systems; they are, therefore, to be understood only in their own terms.

Examples of such biologically specialized perceiving modules can be multiplied. Visual perception of depth by use of information about binocular disparity is a well-studied example that has the same general characteristics we have attributed to auditory localization and speech (Julesz, 1960, 1971; Poggio, 1984). And there is presumably much to be learned by comparison with such biologically coherent systems as those that underlie echolocation in bats (Suga, 1984) or song in birds (Marler, 1970; Thorpe, 1958). But we will not elaborate, for the point to be made here is only that, from a biological point of view, the assumptions of the motor theory are not bizarre.

How the motor theory makes speech perception different from other specialized perceiving systems

Perceptual modules, by definition, differ from one another in the classes of distal events that form their domains and in the relation between these events and the proximal displays. But the phonetic module differs from others in at least two further respects.

Auditory and phonetic domains

The first difference is in the locale of the distal events. In auditory localization, the distal event is 'out there', and the relation between it and the proximal display at the two ears is completely determined by the principles of physical acoustics. Much the same can be said of those specialized modules that deal with the primitives of auditory quality, however they are to be characterized, and that come into play when people perceive, for example, whistles, horns, breaking glass, and barking dogs. Not so for the perception of phonetic structure. There, the distal object is a phonetic gesture or, more explicitly, an 'upstream' neural command for the gesture from which the peripheral articulatory movements unfold. It follows that the relation between distal object and proximal stimulus will have the special feature that it is determined not just by acoustic principles but also by neuromuscular processes internal to the speaker. Of course, analogues of these processes are also available as part of the biological endowment of the listener. Hence, some kind of link between perception and production would seem to characterize the phonetic module, but not those modules that provide auditory localization or visual perception of depth. In a later section, we will have more to say about this link. Now we will only comment that it may conceivably resemble, in its most general characteristics, those links that have been identified in the communication modules of certain nonhuman creatures (Gerhardt & Rheinlaender, 1982; Hoy, Hahn, & Paul, 1977; Hoy & Paul, 1973; Katz & Gurney, 1981; Margolish, 1983; McCasland & Konishi, 1983; Nottebohm, Stokes, & Leonard, 1976; Williams, 1984).

The motor theory aside, it is plain that speech somehow informs listeners about the phonetic intentions of the talker. The particular claim of the motor theory is that these intentions are represented in a specific form in the talker's brain, and that there is a perceiving module specialized to lead the listener effortlessly to that representation. Indeed, what is true of speech in this respect is true for all of language, except, of course, that the more distal object for language is some representation of linguistic structure, not merely of gesture, and that access to this object requires a module that is not merely phonetic, but phonological and syntactic as well.

Competition between phonetic and auditory modes

A second important difference between the phonetic module and the others has to do with the question: how does the module cooperate or compete with others that use stimuli of the same broadly defined physical form? For auditory localization, the key to the answer is the fact that the module is turned on by a specific and readily specifiable characteristic of the proximal stimulus: a particular range of differences in time of arrival at the two ears.

Obviously, such differences have no other utility for the perceiver but to provide information about the distal property, location; there are no imaginable ecological circumstances in which a person could use this characteristic of the proximal stimuli to specify some other distal property. Thus, the proximal display and the distal property it specifies only complement the other aspects of what a listener hears; they never compete.

In phonetic perception, things are quite different because important acoustic cues are often similar to, even identical with, the stimuli that inform listeners about a variety of nonspeech events. We have already remarked that, in isolation, formant transitions sound like glissandi or chirps. Now surely we don't want to perceive these as glissandi or chirps when we are listening to speech, but we do want to perceive them so when we are listening to music or to birdsong. If this is true for all of the speech cues, as in some sense it presumably is, then it is hard to see how the module can be turned on by acoustic stigmata of any kind—that is, by some set of necessary cues defined in purely acoustic terms. We will consider this matter in some greater detail later. For now, however, the point is only that cues known to be of great importance for phonetic events may be cues for totally unrelated nonphonetic events, too. A consequence is that, in contrast to the generally complementary relation of the several modules that serve the same broadly defined modality (e.g., depth and color in vision), the phonetic and auditory modules are in direct competition. (For a discussion of how this competition might be resolved, see Mattingly & Liberman, 1985.)

Experimental evidence for the theory

Having briefly described one motive for the motor theory—the context-conditioned variation in the acoustic cues for constant phonetic categories—we will now add others. We will limit ourselves to the so-called segmental aspects of phonetic structure, though the theory ought, in principle, to apply in the suprasegmental domain as well (cf. Fowler, 1982).

The two parts of the theory—that gestures are the objects of perception and that perception of these gestures depends on a specialized module—might be taken to be independent, as they were in their historical development, but the relevant data are not. We therefore cannot rationally apportion the data between the parts, but must rather take them as they come.

A result of articulation: The multiplicity, variety, and equivalence of cues for each phonetic percept

When speech synthesis began to be used as a tool to investigate speech per-

ception, it was soon discovered that, in any specific context, a particular local property of the acoustic signal was sufficient for the perception of one phonetic category rather than another and, more generally, that the percept could be shifted along some phonetic dimension by varying the synthetic stimulus along a locally-definable acoustic dimension. For example, if the onset frequency of the transition of the second formant during a stop release is sufficiently low, relative to the frequency of the following steady state, the stop is perceived as labial; otherwise, as apical or dorsal (Liberman et al., 1954). A value along such an acoustic dimension that was optimal for a particular phonetic category, or, more loosely, the dimension itself, was termed an 'acoustic cue'.

Of course, the fact that particular acoustic cues can be isolated must, of itself, tell us something about speech perception, for it might have been otherwise. Thus, it is possible to imagine a speech-perception mechanism, equipped, perhaps, with auditory templates, that would break down if presented with anything other than a wholly natural and phonetically optimal stimulus. Listeners would either give conflicting and unreliable phonetic judgments or else not hear speech at all. Clearly, the actual mechanism is not of this kind, and the concept of cue accords with this fact.

Nevertheless, the emphasis on the cues has, perhaps, been unfortunate, for the term 'cue' might seem to imply a claim about the elemental units of speech perception. But 'cue' was simply a convenient bit of laboratory jargon referring to acoustic variables whose definition depended very much on the design features of the particular synthesizers that were used to study them. The cues, as such, have no role in a theory of speech perception; they only describe some of the facts on which a theory might be based (cf. Bailey & Summerfield, 1980). There are, indeed, several generalizations about the cues—some only hinted at by the data now available, others quite well founded—that are relevant to such a theory.

One such generalization is that every 'potential' cue—that is, each of the many acoustic events peculiar to a linguistically significant gesture—is an *actual* cue. (For example, every one of 18 potential cues to the voicing distinction in medial position has been shown to have some perceptual value; Lisker, 1978.) All possible cues have not been tested, and probably never will be, but no potential cue has yet been found that could not be shown to be an actual one.

A closely related generalization is that, while each cue is, by definition, more or less sufficient, none is truly necessary. The absence of any single cue, no matter how seemingly characteristic of the phonetic category, can be compensated for by others, not without some cost to naturalness or even intelligibility, perhaps, but still to such an extent that the intended category is, in

fact, perceived. Thus, stops can be perceived without silent periods, fricatives without frication, vowels without formants, and tones without pitch (Abramson, 1972; Inoue, 1984; Remez & Rubin, 1984; Repp, 1984; Yeni-Komshian & Soli, 1981).

Yet another generalization is that even when several cues are present, variations in one can, within limits, be compensated for by offsetting variations in another (Dorman, Studdert-Kennedy, & Raphael, 1977; Dorman, Raphael, & Liberman, 1979; Hoffman, 1958; Howell & Rosen, 1983; Lisker, 1957; Summerfield & Haggard, 1977). In the case of the contrast between fricative-vowel and fricative-stop-vowel (as in [sa] vs. [sta]), investigators have found that two important cues, silence and appropriate formant transitions, engage in just such a trading relation. That this bespeaks a true equivalence in perception was shown by experiments in which the effect of variation in one cue could, depending on its 'direction', be made to 'add to' or 'cancel out' the effect of the other (Fitch, Halwes, Erickson, & Liberman, 1980). Significantly, this effect can also be obtained with sine-wave analogues of speech, but only for subjects who perceive these signals as speech, not for those who perceive them as nonspeech tones (Best, Morrongiello, & Robson, 1981).

Putting together all the generalizations about the multiplicity and variety of acoustic cues, we should conclude that there is simply no way to define a phonetic category in purely acoustic terms. A complete list of the cues—surely a cumbersome matter at best—is not feasible, for it would necessarily include all the acoustic effects of phonetically distinctive articulations. But even if it were possible to compile such a list, the result would not repay the effort, because none of the cues on the list could be deemed truly essential. As for those cues that might, for any reason, be finally included, none could be assigned a characteristic setting, since the effect of changing it could be offset by appropriate changes in one or more of the others. This surely tells us something about the design of the phonetic module. For if phonetic categories were acoustic patterns, and if, accordingly, phonetic perception were properly auditory, one should be able to describe quite straightforwardly the acoustic basis for the phonetic category and its associated percept. According to the motor theory, by contrast, one would expect the acoustic signal to serve only as a source of information about the gestures; hence the gestures would properly define the category. As for the perceptual equivalence among diverse cues that is shown by the trading relations, explaining that on auditory grounds requires ad hoc assumptions. But if, as the motor theory would have it, the gesture is the distal object of perception, we should not wonder that the several sources of information about it are perceptually equivalent, for they are products of the same linguistically significant gesture.

A result of coarticulation: I. Segmentation in sound and percept

Traditional phonetic transcription represents utterances as single linear sequences of symbols, each of which stands for a phonetic category. It is an issue among phonologists whether such transcriptions are really theoretically adequate, and various alternative proposals have been made in an effort to provide a better account. This matter need not concern us here, however, since all proposals have in common that phonetic units of some description are ordered from left to right. Some sort of segmentation is thus always implied, and what theory must take into account is that the perceived phonetic object is thus segmented.

Segmentation of the phonetic percept would be no problem for theory if the proximal sound were segmented correspondingly. But it is not, nor can it be, if speech is to be produced and perceived efficiently. To maintain a straightforward relation in segmentation between phonetic unit and signal would require that the sets of phonetic gestures corresponding to phonetic units be produced one at a time, each in its turn. The obvious consequence would be that each unit would become a syllable, in which case talkers could speak only as fast as they could spell. A function of coarticulation is to evade this limitation. There is an important consequence, however, which is that there is now no straightforward correspondence in segmentation between the phonetic and acoustic representations of the information (Fant, 1962; Joos, 1948). Thus, the acoustic information for any particular phonetic unit is typically overlapped, often quite thoroughly, with information for other units. Moreover, the span over which that information extends, the amount of overlap, and the number of units signalled within the overlapped portion all vary according to the phonetic context, the rate of articulation, and the language (Magen, 1984; Manuel & Krakow, 1984; Öhman, 1966; Recasens, 1984; Repp, Liberman, Eccardt, & Pesetsky, 1978; Tuller, Harris, & Kelso, 1982).

There are, perhaps, occasional stretches of the acoustic signal over which there is information about only one phonetic unit—for example, in the middle of the frication in a slowly articulated fricative-vowel syllable and in vowels that are sustained for artificially long times. Such stretches do, of course, offer a relation between acoustic patterns and phonetic units that would be transparent if phonetic perception were merely auditory. But even in these cases, the listener automatically takes account of, not just the transparent part of the signal, but the regions of overlap as well (Mann & Repp, 1980, 1981; Whalen, 1981). Indeed, the general rule may be that the phonetic percept is normally made available to consciousness only after all the relevant acoustic information is in, even when earlier cues might have been sufficient (Martin & Bunnell, 1981, 1982; Repp et al., 1978).

What wants explanation, then, is that the percept is segmented in a way

that the signal is not, or, to put it another way, that the percept does not mirror the overlap of information in the sound (cf. Fowler, 1984). The motor theory does not provide a complete explanation, certainly not in its present state, but it does head the theoretical enterprise in the right direction. At the very least, it turns the theorist away from the search for those unlikely processes that an auditory theory would have him seek: how listeners learn phonetic labels for what they hear and thus re-interpret perceived overlap as sequences of discrete units; or how discrete units emerge in perception from interactions of a purely auditory sort. The first process seems implausible on its face, the second because it presupposes that the function of the many kinds and degrees of coarticulation is to produce just those combinations of sounds that will interact in accordance with language-independent characteristics of the auditory system. In contrast, the motor theory begins with the assumption that coarticulation, and the resulting overlap of phonetic information in the acoustic pattern, is a consequence of the efficient processes by which discrete phonetic gestures are realized in the behavior of more or less independent articulators. The theory suggests, then, that an equally efficient perceptual process might use the resulting acoustic pattern to recover the discrete gestures.

A result of coarticulation: II. Different sounds, different contexts, same percept
That the phonetic percept is invariant even when the relevant acoustic cue is not was the characteristic relation between percept and sound that we took as an example in the first section. There, we observed that variation in the acoustic pattern results from overlapping of putatively invariant gestures, an observation that, as we remarked, points to the gesture, rather than the acoustic pattern itself, as the object of perception. We now add that the articulatory variation due to context is pervasive: in the acoustic representation of every phonetic category yet studied there are context-conditioned portions that contribute to perception and that must, therefore, be taken into account by theory. Thus, for stops, nasals, fricatives, liquids, semivowels, and vowels, the always context-sensitive transitions are cues (Harris, 1958; Jenkins, Strange, & Edman, 1983; Liberman et al., 1954; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957; Strange, Jenkins, & Johnson, 1983). For stops and fricatives, the noises that are produced at the point of constriction are also known to be cues, and, under some circumstances at least, these, too, vary with context (Dorman et al., 1977; Liberman et al., 1952; Whalen, 1981).

An auditory theory that accounts for invariant perception in the face of so much variation in the signal would require a long list of apparently arbitrary assumptions. For a motor theory, on the other hand, systematic stimulus

variation is not an obstacle to be circumvented or overcome in some arbitrary way; it is, rather, a source of information about articulation that provides important guidance to the perceptual process in determining a representation of the distal gesture.

A result of coarticulation: III. Same sound, different contexts, different percepts

When phonetic categories share one feature but differ in another, the relation between acoustic pattern and percept speaks, again, to the motor theory and its alternatives. Consider, once more, the fricative [s] and the stop [t] in the syllables [sa] and [sta]. In synthesis, the second- and third-formant transitions can be the same for these two categories, since they have the same place of articulation; and the first-formant transition, normally a cue to manner, can be made ambiguous between them. For such stimuli, the perception of [sta] rather than [sa] depends on whether there is an interval of silence between the noise for the [s] and the onsets of the transitions.

Data relevant to an interpretation of the role of silence in thus producing different percepts from the same transition come from two kinds of experiments. First are those that demonstrate the effectiveness of the transitions as cues for the place feature of the fricative in fricative-vowel syllables (Harris, 1958). The transitions are not, therefore, masked by the noise of the [s] friction, and thus the function of silence in a stop is not, as it might be in an auditory theory, to protect the transitions from such masking. The second kind of experiment deals with the possibility of a purely auditory interaction—in this case, between silence and the formant transitions. Among the findings that make such auditory interaction seem unlikely is that silence affects perception of the formant transitions differently in and out of speech context and, further, that the effectiveness of silence depends on such factors as continuity of talker and prosody (Dorman et al., 1979; Rakerd, Dechovitz, & Verbrugge, 1982). But perhaps the most direct test for auditory interaction is provided by experiments in which such interaction is ruled out by holding the acoustic context constant. This can be done by exploiting 'duplex perception', a phenomenon to be discussed in greater detail in the next section. Here it is appropriate to say only that duplex perception provides a way of presenting acoustic patterns so that, in a fixed context, listeners hear the same second- or third-formant transitions in two phenomenally different ways simultaneously: as nonspeech chirps and as cues for phonetic categories. The finding is that the presence or absence of silence determines whether formant transitions appropriate for [t] or for [p], for example, are integrated into percepts as different as stops and fricatives; but silence has no effect on the perception of the nonspeech chirps that these same transitions produce (Liberman, Isenberg, &

Rakerd, 1981). Since the latter result eliminates the possibility of auditory interaction, we are left with the account that the motor theory would suggest: that silence acts in the specialized phonetic mode to inform the listener that the talker completely closed his vocal tract to produce a stop consonant, rather than merely constricting it to produce a fricative. It follows, then, that silence will, by its presence or absence, determine whether identical transitions are cues in percepts that belong to the one manner or the other.

An acoustic signal diverges to phonetic and auditory modes

We noted earlier that a formant transition is perceptually very different depending on whether it is perceived in the auditory mode, where it sounds like a chirp, or in the phonetic mode, where it cues a 'nonchirpy' consonant. Of course, the comparison is not entirely fair, since acoustic context is not controlled: the transition is presented in isolation in the one case, but as an element of a larger acoustic pattern in the other. We should, therefore, call attention to the fact that the same perceptual difference is obtained even when, by resort to a special procedure, acoustic context is held constant (Liberman, 1979; Rand, 1974). This procedure, which produces the duplex percept referred to earlier, goes as follows. All of an acoustic syllable except only the formant transition that decides between, for example, [da] and [ga] is presented to one ear. By itself, this pattern, called the 'base', sounds like a stop-vowel syllable, ambiguous between [da] and [ga]. To the other ear is presented one or the other of the transitions appropriate for [d] or [g]. In isolation, these sound like different chirps. Yet, when base and transition are presented dichotically, and in the appropriate temporal relationship, they give rise to a duplex percept: [da] or [ga], depending on the transition, and, simultaneously, the appropriate chirp. (The fused syllable appears to be in the ear to which the base had been presented, the chirp in the other.)

Two related characteristics of duplex perception must be emphasized. One is that it is obtained only when the stimulus presented to one ear is, like the 'chirpy' transition, of short duration and extremely unspeechlike in quality. If that condition is not met, as, for example, when the first two formants are presented to one ear and the entire third formant to the other, perception is not duplex. It is, on the contrary, simplex; one hears a coherent syllable in which the separate components cannot be apprehended. (A very different result is obtained when two components of a musical chord are presented to one ear, a third component to the other. In that case, listeners can respond to the third component by itself and also to that component combined with the first two (Pastore, Schmuckler, Rosenblum, & Szczesiul, 1983).

The other, closely related characteristic of duplex perception is that it is precisely duplex, not triplex. That is, listeners perceive the nonspeech chirp

and the fused syllable, but they do not also perceive the base—that is, the syllable, minus one of the formant transitions—that was presented to one ear (Repp, Milburn, & Ashkenas, 1983). (In the experiment with musical chords by Pastore et al., 1983, referred to just above, there was no test for duplex, as distinguished from triplex, perception.)

The point is that duplex perception does not simply reflect the ability of the auditory system to fuse dichotically presented stimuli and also, as in the experiment with the chords, to keep them apart. Rather, the duplex percepts of speech comprise the only two ways in which the transition, for example, can be heard: as a cue for a phonetic gesture and as a nonspeech sound. These percepts are strikingly different, and, as we have already seen, they change in different, sometimes contrasting ways in response to variations in the acoustic signals—variations that must have been available to all structures in the brain that can process auditory information. A reasonable conclusion is that there must be two modules that can somehow use the same input to produce simultaneous representations of two distal objects. (For speculation about the mechanism that normally prevents perception of this ecologically impossible situation, and about the reason why that highly adaptive mechanism might be defeated by the procedures used to produce duplex perception, see Mattingly & Liberman, 1985.)

Acoustic and optical signals converge on the phonetic mode

In duplex perception, a single acoustic stimulus is processed simultaneously by the phonetic and auditory modules to produce perception of two distal objects: a phonetic gesture and a sound. In the phenomenon to which we turn now, something like the opposite occurs: two different stimuli—one acoustic, the other optical—are combined by the phonetic module to produce coherent perception of a single distal event. This phenomenon, discovered by McGurk and McDonald (1976), can be illustrated by this variant on their original demonstration. Subjects are presented acoustically with the syllables [ba], [ba], [ba] and optically with a face that, in approximate synchrony, silently articulates [bɛ], [vɛ], [ðɛ]. The resulting and compelling percept is [ba], [va], [ða], with no awareness that it is in any sense bimodal—that is, part auditory and part visual. According to the motor theory, this is so because the perceived event is neither; it is, rather, a gesture. The proximal acoustic signal and the proximal optical signal have in common, then, that they convey information about the same distal object. (Perhaps a similar convergence is implied by the finding that units in the optic tectum of the barn owl are bimodally sensitive to acoustic and optical cues for the same distal property, location in space; Knudsen, 1982).

Even prelinguistic infants seem to have some appreciation of the relation

between the acoustic and optical consequences of phonetic articulation. This is to be inferred from an experiment in which it was found that infants at four to five months of age preferred to look at a face that articulated the vowel they were hearing rather than at the same face articulating a different vowel (Kuhl & Meltzoff, 1982). Significantly, this result was not obtained when the sounds were pure tones matched in amplitude and duration to the vowels. In a related study it was found that infants of a similar age looked longer at a face repeating the disyllable they were hearing than at the same face repeating, another disyllable, though both disyllables were carefully synchronized with the visible articulation (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). Like the results obtained with adults in the McGurk–MacDonald kind of experiment, these findings with infants imply a perception–production link and, accordingly, a common mode of perception for all proper information about the gesture.

The general characteristics that cause acoustic signals to be perceived as speech
The point was made in an earlier section that acoustic definitions of phonetic contrasts are, in the end, unsatisfactory. Now we would suggest that acoustic definitions also fail for the purpose of distinguishing in general between acoustic patterns that convey phonetic structures and those that do not. Thus, speech cannot be distinguished from nonspeech by appeal to surface properties of the sound. Surely, natural speech does have certain characteristics of a general and superficial sort—for example, formants with characteristic bandwidths and relative intensities, stretches of waveform periodicities that typically mark the voiced portion of syllables, peaks of intensity corresponding approximately to syllabic rhythm, etc.—and these can be used by machines to detect speech. But research with synthesizers has shown that speech is perceived even when such general characteristics are absent. This was certainly true in the case of many of the acoustic patterns that were used in work with the Pattern Playback synthesizer, and more recently it has been shown to be true in the most extreme case of patterns consisting only of sine waves that follow natural formant trajectories (Remez, Rubin, Pisoni, & Carrell, 1981). Significantly, the converse effect is also obtained. When reasonably normal formants are made to deviate into acoustically continuous but abnormal trajectories, the percept breaks into two categorically distinct parts: speech and a background of chirps, glissandi, and assorted noises (Liberman & Studdert-Kennedy, 1978). Of course, the trajectories of the formants are determined by the movements of the articulators. Evidently, those trajectories that conform to possible articulations engage the phonetic module; all others fail.

We conclude that acoustic patterns are identified as speech by reference to deep properties of a linguistic sort: if a sound can be 'interpreted' by the specialized phonetic module as the result of linguistically significant gestures, then it is speech; otherwise, not. (In much the same way, grammatical sentences can be distinguished from ungrammatical ones, not by lists of surface properties, but only by determining whether or not a grammatical derivation can be given.) Of course, the kind of mechanism such an 'interpretation' requires is the kind of mechanism the motor theory presumes.

Phonetic and auditory responses to the cues

Obviously, a module that acts on acoustic signals cannot respond beyond the physiological limits of those parts of the auditory system that transmit the signal to the module. Within those limits, however, different modules can be sensitive to the signals in different ways. Thus, the auditory-localization module enables listeners to perceive differences in the position of sounding objects given temporal disparity cues smaller by several orders of magnitude than those required to make the listener aware of temporal disparity as such (Brown & Deffenbacher, 1979, chap. 7; Hirsh, 1959). If there is, as the motor theory implies, a distinct phonetic module, then in like manner its sensitivities should not, except by accident, be the same as those that characterize the module that deals with the sounds of non-speech events.

In this connection, we noted in the first section of the paper that one form of auditory theory of speech perception points to auditory discontinuities in differential sensitivity (or in absolute identification), taking these to be the natural bases for the perceptual discontinuities that characterize the boundaries of phonetic categories. But several kinds of experiments strongly imply that this is not so.

One kind of experiment has provided evidence that the perceptual discontinuities at the boundaries of phonetic categories are not fixed; rather, they move in accordance with the acoustic consequences of articulatory adjustments associated with phonetic context, dialect, and rate of speech. (For a review, see Repp & Liberman, in press.) To account for such articulation-correlated changes in perceptual sensitivities by appeal to auditory processes requires, yet again, an ultimately countless set of ad hoc assumptions about auditory interactions, as well as the implausible assumption that the articulators are always able to behave so as to produce just those sounds that conform to the manifold and complex requirements that the auditory interactions impose. It seems hardly more plausible that, as has been suggested, the discontinuities in phonetic perception are really auditory discontinuities that were caused to move about in phylogenetic or ontogenetic development as a result of experience with speech (Aslin & Pisoni, 1980). The difficulty with this as-

sumption is that it presupposes the very canonical form of the cues that does not exist (see above) and, also, that it implies a contradiction in assuming, as it must, that the auditory sensitivities underwent changes in the development of speech, yet somehow also remained unchanged and nonetheless manifest in the adult's perception of nonspeech sounds.

Perhaps this is the place to remark about categorical perception that the issue is not, as is often supposed, whether nonspeech continua are categorically perceived, for surely some do show tendencies in that direction. The issue is whether, given the same (or similar) acoustic continua, the auditory and phonetic boundaries are in the same place. If there are, indeed, auditory boundaries, and if, further, these boundaries are replaced in phonetic perception by boundaries at different locations (as the experiments referred to above do indicate), then the separateness of phonetic and auditory perception is even more strongly argued for than if the phonetic boundaries had appeared on continua where auditory boundaries did not also exist.

Also relevant to comparison of sensitivity in phonetic and auditory modes are experiments on perception of acoustic variations when, in the one case, they are cues for phonetic distinctions, and when, in some other, they are perceived as nonspeech. One of the earliest of the experiments to provide data about the nonspeech side of this comparison dealt with perception of frequency-modulated tones—or 'ramps' as they were called—that bear a close resemblance to the formant transitions. The finding was that listeners are considerably better at perceiving the pitch at the end of the ramp than at the beginning (Brady, House, & Stevens, 1961). Yet, in the case of stop consonants that are cued by formant transitions, perception is better syllable-initially than syllable-finally, though in the former case it requires information about the beginning of the ramp, while in the latter it needs to know about the end. Thus, if one were predicting sensitivity to speech from sensitivity to the analogous nonspeech sounds, one would make exactly the wrong predictions. More recent studies have made more direct comparisons and found differences in discrimination functions when, in speech context, formant transitions cued place distinctions among stops and liquids, and when, in isolation, the same transitions were perceived as nonspeech sounds (Mattingly et al., 1971; Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975).

More impressive, perhaps, is evidence that has come from experiments in which listeners are induced to perceive a constant stimulus in different ways. Here belong experiments in which sinewave analogues of speech, referred to earlier, are presented under conditions that cause some listeners to perceive them as speech and others not. The perceived discontinuities lie at different places (on the acoustic continuum) for the two groups (Best et al., 1981; Best & Studdert-Kennedy, 1983; Studdert-Kennedy & Williams, 1984; Williams,

Verbrugge, & Studdert-Kennedy, 1983). Here, too, belongs an experiment in which the formant-transitions appropriate to a place contrast between stop consonants are presented with the remainder of a syllable in such a way as to produce the duplex percept referred to earlier: the transitions cue a stop consonant and, simultaneously, nonspeech chirps. The result is that listeners yield quite different discrimination functions for exactly the same formant transitions in exactly the same acoustic context, depending on whether they are responding to the speech or nonspeech sides of the duplex percept; only on the speech side of the percept is there a peak in the discrimination function to mark a perceptual discontinuity at the phonetic boundary (Mann & Liberman, 1983).

Finally, we note that, apart from differences in differential sensitivity to the transitions, there is also a difference in absolute-threshold sensitivity when, in the one case, these transitions support a phonetic percept, and when, in the other, they are perceived as nonspeech chirps. Exploiting, again, the phenomenon of duplex perception, investigators found that the transitions were effective (on the speech side of the percept) in cueing the contrast between stops at a level of intensity 18 db lower than that required for comparable discrimination of the chirps (Bentin & Mann, 1983). At that level, indeed, listeners could not even hear the chirps, let alone discriminate them; yet they could still use the transitions to identify the several stops.

The several aspects of the theory

For the purpose of evaluating the motor theory, it is important to separate it into its more or less independent parts. First, and fundamentally, there is the claim that phonetic perception is perception of gesture. As we have seen, this claim is based on evidence that the invariant source of the phonetic percept is somewhere in the processes by which the sounds of speech are produced. In the first part of this section we will consider where in those processes the invariant might be found.

The motor theory also implies a tight link between perception and production. In the second part of this section we will ask how that link came to be.

Where is the invariant phonetic gesture?

A phonetic gesture, as we have construed it, is a class of movements by one or more articulators that results in a particular, linguistically significant deformation, over time, of the vocal-tract configuration. The linguistic function of the gesture is clear enough: phonetic contrasts, which are of course the basis

of phonological categories, depend on the choice of one particular gesture rather than another. What is not so clear is how the gesture relates to the actual physical movements of articulators and to the resulting vocal-tract configurations, observed, for example, in X-ray films.

In the early days of the motor theory, we made a simplifying assumption about this relation: that a gesture was effected by a single key articulator. On this assumption, the actual movement trajectory of the articulator might vary, but only because of aerodynamic factors and the physical linkage of this articulator with others, so the neural commands in the final common paths (observable with electromyographic techniques) would nevertheless be invariant across different contexts. This assumption was appropriate as an initial working hypothesis, if only because it was directly testable. In the event, there proved to be a considerable amount of variability which the hypothesis could not account for.

In formulating this initial hypothesis, we had overlooked several serious complications. One is that a particular gesture typically involves not just one articulator, but two or more; thus 'lip rounding', for example, is a collaboration of lower lip, upper lip, and jaw. Another is that a single articulator may participate in the execution of two different gestures at the same time; thus, the lips may be simultaneously rounding and closing in the production of a labial stop followed by a rounded vowel, for example, [bu]. Prosody makes additional complicating demands, as when a greater displacement of some or all of the active articulators is required in producing a stressed syllable rather than an unstressed one; and linguistically irrelevant factors, notably speaking rate, affect the trajectory and phasing of the component movements.

These complications might suggest that there is little hope of providing a rigorous physical definition of a particular gesture, and that the gestures are hardly more satisfactory as perceptual primitives than are the acoustic cues. It might, indeed, be argued that there is an infinite number of possible articulatory movements, and that the basis for categorizing one group of such movements as 'lip rounding' and another as 'lip closure' is entirely a priori.

But the case for the gesture is by no means as weak as this. Though we have a great deal to learn before we can account for the variation in instances of the same gesture, it is nonetheless clear that, despite such variation, the gestures have a virtue that the acoustic cues lack: instances of a particular gesture always have certain topological properties not shared by any other gesture. That is, for any particular gesture, the same sort of distinctive deformation is imposed on the current vocal-tract configuration, whatever this 'underlying' configuration happens to be. Thus, in lip rounding, the lips are always slowly protruded and approximated to some appreciable extent, so that the anterior end of the vocal tract is extended and narrowed, though the

relative contributions of the tongue and lips, the actual degrees of protrusion and approximation, and the speed of articulatory movement vary according to context. Perhaps this example seems obvious because lip rounding involves a local deformation of the vocal-tract configuration, but the generalization also applies to more global gestures. Consider, for example, the gesture required to produce an 'open' vowel. In this gesture, tongue, lips, jaw, and hyoid all participate to contextually varying degrees, and the actual distance between the two lips, as well as that between the tongue blade and body and the upper surfaces of the vocal tract, are variable; but the goal is always to give the tract a more open, horn-shaped configuration than it would otherwise have had.

We have pointed out repeatedly that, as a consequence of gestural overlapping, the invariant properties of a particular gesture are not manifest in the spectrum of the speech signal. We would now caution that a further consequence of this overlapping is that, because of their essentially topological character, the gestural invariants are usually not obvious from inspection of a single static vocal-tract configuration, either. They emerge only from consideration of the configuration as it changes over time, and from comparison with other configurations in which the same gesture occurs in different contexts, or different gestures in the same context.

We would argue, then, that the gestures do have characteristic invariant properties, as the motor theory requires, though these must be seen, not as peripheral movements, but as the more remote structures that control the movements. These structures correspond to the speaker's intentions. What is far from being understood is the nature of the system that computes the topologically appropriate version of a gesture in a particular context. But this problem is not peculiar to the motor theory; it is familiar to many who study the control and coordination of movement, for they, like us, must consider whether, given context-conditioned variability at the surface, motor acts are nevertheless governed by invariants of some sort (Browman & Goldstein, 1985; Fowler, Rubin, Remez, & Turvey, 1980; Tuller & Kelso, 1984; Turvey, 1977).

The origin of the perception–production link

In the earliest accounts of the motor theory, we put considerable emphasis on the fact that listeners not only perceive the speech signal but also produce it. This, together with doctrinal behaviorist considerations, led us to assume that the connection between perception and production was formed as a wholly learned association, and that perceiving the gesture was a matter of picking up the sensory consequences of covert mimicry. On this view of the

genesis of the perception–production link, the distinguishing characteristic of speech is only that it provides the opportunity for the link to be established. Otherwise, ordinary principles of associative learning are adequate to the task; no specialization for language is required.

But then such phenomena as have been described in this paper were discovered, and it became apparent that they differed from anything that association learning could reasonably be expected to produce. Nor were these the only relevant considerations. Thus, we learned that people who have been pathologically incapable from birth of controlling their articulators are nonetheless able to perceive speech (MacNeilage, Rootes, & Chase, 1967). From the research pioneered by Eimas, Siqueland, Jusczyk and Vigorito (1971), we also learned that prelinguistic infants apparently categorize phonetic distinctions much as adults do. More recently, we have seen that even when the distinction is not functional in the native language of the subjects, and when, accordingly, adults have trouble perceiving it, infants nevertheless do quite well up to about one year of age, at which time they begin to perform as poorly as adults (Werker & Tees, 1984). Perhaps, then, the sensitivity of infants to the acoustic consequences of linguistic gestures includes all those gestures that could be phonetically significant in any language, acquisition of one's native language being a process of losing sensitivity to gestures it does not use. Taking such further considerations as these into account, we have become even more strongly persuaded that the phonetic mode, and the perception–production link it incorporates, are innately specified.

Seen, then, as a view about the biology of language, rather than a comment on the coincidence of speaking and listening, the motor theory bears at several points on our thinking about the development of speech perception in the child. Consider, first, a linguistic ability that, though seldom noted (but see Mattingly, 1976), must be taken as an important prerequisite to acquiring the phonology of a language. This is the ability to sort acoustic patterns into two classes: those that contain (candidate) phonetic structures and those that do not. (For evidence, however indirect, that infants do so sort, see Alegria & Noirot, 1982; Best, Hoffman, & Glanville, 1982; Entus, 1977; Molfese, Freeman, & Palermo, 1975; Segalowitz & Chapman, 1980; Witelson, 1977; but see Vargha-Khadem & Corballis, 1979). To appreciate the bearing of the motor theory on this matter, recall our claim, made in an earlier section, that phonetic objects cannot be perceived as a class by reference to acoustic stigmata, but only by a recognition that the sounds might have been produced by a vocal tract as it made linguistically significant gestures. If so, the perception–production link is a necessary condition for recognizing speech as speech. It would thus be a blow to the motor theory if it could be shown that infants must develop empirical criteria for this purpose. Fortunately for the

theory, such criteria appear to be unnecessary.

Consider, too, how the child comes to know, not only that phonetic structures are present, but, more specifically, just what those phonetic structures are. In this connection, recall that information about the string of phonetic segments is overlapped in the sound, and that there are, accordingly, no acoustic boundaries. Until and unless the child (tacitly) appreciates the gestural source of the sounds, he can hardly be expected to perceive, or ever learn to perceive, a phonetic structure. Recall, too, that the acoustic cues for a phonetic category vary with phonetic factors such as context and with extra-phonetic factors such as rate and vocal-tract size. This is to say, once again, that there is no canonical cue. What, then, is the child to learn? Association of some particular cue (or set of cues) with a phonetic category will work only for a particular circumstance. When circumstances change, the child's identification of the category will be wrong, sometimes grossly, and it is hard to see how he could readily make the appropriate correction. Perception of the phonetic categories can properly be generalized only if the acoustic patterns are taken for what they really are: information about the underlying gestures. No matter that the child sometimes mistakes the phonological significance of the gesture, so long as that which he perceives captures the systematic nature of its relation to the sound; the phonology will come in due course. To appreciate this relation is, once again, to make use of the link between perception and production.

How 'direct' is speech perception?

Since we have been arguing that speech perception is accomplished without cognitive translation from a first-stage auditory register, our position might appear similar to the one Gibson (1966) has taken to regard to 'direct perception'. The similarity to Gibson's views may seem all the greater because, like him, we believe that the object of perception is motoric. But there are important differences, the bases for which are to be seen in the following passage (Gibson, 1966, p. 94):

> An articulated utterance is a source of a vibratory field in the air. The source is biologically 'physical' and the vibration is acoustically 'physical'. The vibration is a potential stimulus, becoming effective when a listener is within range of the vibratory field. The listener then *perceives* the articulation because the invariants of vibration correspond to those of articulation. In this theory of speech perception, the units and parts of speech are present both in the mouth of the speaker and in the air between the speaker and listener. Phonemes are in the air. They

> can be considered physically real if the higher-order invariants of sound waves are admitted to the realm of physics.

The first difference between Gibson's view and ours relates to the nature of the perceived events. For Gibson, these are actual movements of the articulators, while for us, they are the more remote gestures that the speaker intended. The distinction would be trivial if an articulator were affected by only one gesture at a time, but, as we have several times remarked, an articulatory movement is usually the result of two or more overlapping gestures. The gestures are thus control structures for the observable movements.

The second difference is that, unlike Gibson, we do not think articulatory movements (let alone phonetic structures) are given directly (that is, without computation) by 'higher-order invariants' that would be plain if only we had a biologically appropriate science of physical acoustics. We would certainly welcome any demonstration that such invariants did exist, since, even though articulatory movement is not equivalent to phonetic structure, such a demonstration would permit a simpler account of how the phonetic module works. But no higher-order invariants have thus far been proposed, and we doubt that any will be forthcoming. We would be more optimistic on this score if it could be shown, at least, that articulatory movements can be recovered from the signal by computations that are purely analytic, if nevertheless complex. One might then hope to reformulate the relationship between movements and signal in a way that would make it possible to appeal to higher-order invariants and thus obviate the need for computation. But, given the many-to-one relation between vocal-tract configurations and acoustic signal, a purely analytic solution to the problem of recovering movements from the signal seems to be impossible unless one makes unrealistic assumptions about excitation, damping, and other physical variables (Sondhi, 1979). We therefore remain skeptical about higher-order invariants.

The alternative to an analytic account of speech perception is, of course, a synthetic one, in which case the module compares some parametric description of the input signal with candidate signal descriptions. As with any form of 'analysis-by-synthesis' (cf. Stevens & Halle, 1967), such an account is plausible only if the number of candidates the module has to test can be kept within reasonable bounds. This requirement is met, however, if, as we suppose, the candidate signal descriptions are computed by an analogue of the production process—an internal, innately specified vocal-tract synthesizer, as it were (Liberman, Mattingly, & Turvey, 1972; Mattingly & Liberman, 1969)—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. Further con-

straints become available as experience with the phonology of a particular language reduces the inventory of possible gestures and provides information about the phonotactic and temporal restrictions on their occurrence. The module has then merely to determine which (if any) of the small number of gestures that might have been initiated at a particular instant could, in combination with gestures already in progress, account for the signal.

Thus, we would claim that the processes of speech perception are, like other linguistic processes, inherently computational and quite indirect. If perception seems nonetheless immediate, it is not because the process is in fact straightforward, but because the module is so well-adapted to its complex task.

The motor theory and modularity

In attributing speech perception to a 'module,' we have in mind the notion of modularity proposed by Fodor (1983). A module, for Fodor, is a piece of neural architecture that performs the special computations required to provide central cognitive processes with representations of objects or events belonging to a natural class that is ecologically significant for the organism. This class, the 'domain' of the module, is apt also to be 'eccentric,' for the domain would be otherwise merely a province of some more general domain, for which another module must be postulated anyway. Besides domain-specificity and specialized neural architecture, a module has other characteristic properties. Because the perceptual process it controls is not cognitive, there is little or no possibility of awareness of whatever computations are carried on within the module ('limited central access'). Because the module is specialized, it has a 'shallow' output, consisting only of rigidly definable, domain-relevant representations; accordingly, it processes only the domain-relevant information in the input stimulus. Its computations are thus much faster than those of the less specialized processes of central cognition. Because of the ecological importance of its domain for the organism, the operation of the module is not a matter of choice, but 'mandatory'; for the same reason, its computations are 'informationally encapsulated', that is, protected from cognitive bias.

Most psychologists would agree that auditory localization, to return to an example we have mentioned several times, is controlled by specialized processes of some noncognitive kind. They might also agree that its properties are those that Fodor assigns to modules. At all events, they would set auditory localization apart from such obviously cognitive activities as playing chess, proving theorems, and recognizing a particular chair as a token of the type called 'chair'. As for perception of language, the consensus is that it qualifies as a cognitive process par excellence, modular only in that it is supported by

the mechanisms of the auditory modality. But in this, we and Fodor would argue, the consensus is doubly mistaken; the perception of language is neither cognitive nor auditory. The events that constitute the domain of linguistic perception, however they may be defined, must certainly be an ecologically significant natural class, and it has been recognized since Broca that linguistic perception is associated with specialized neural architecture. Evidently, linguistic perception is fast and mandatory; arguably, it is informationally encapsulated—that is, its phonetic, morphological and syntactic analyses are not biased by knowledge of the world—and its output is shallow—that is, it produces a linguistic description of the utterance, and only this. These and other considerations suggest that, like auditory localization, perception of language rests on a specialization of the kind that Fodor calls a module.

The data that have led us in the past to claim that 'speech is special' and to postulate a 'speech mode' of perception can now be seen to be consistent with Fodor's claims about modularity, and especially about the modularity of language. (What we have been calling a phonetic module is then more properly called a linguistic module.) Thus, as we have noted, speech perception uses all the information in the stimulus that is relevant to phonetic structures: every potential cue proves to be an actual cue. This holds true even across modalities: relevant optical information combines with relevant acoustic information to produce a coherent phonetic percept in which, as in the example described earlier, the bimodal nature of the stimulation is not detectable. In contrast, irrelevant information in the stimulus is *not* used: the acoustic properties that might cause the transitions to be heard as chirps are ignored—or perhaps we should say that the auditory consequences of those properties are suppressed—when the transitions are in context and the linguistic module is engaged. The exclusion of the irrelevant extends, of course, to stimulus information about voice quality, which helps to identify the speaker (perhaps by virtue of some other module) but has no phonetic importance, and even to that extraphonetic information which might have been supposed to help the listener distinguish sounds that contain phonetic structures from those that do not. As we have seen, even when synthetic speech lacks the acoustic properties that would make it sound natural, it will be treated as speech if it contains sufficiently coherent phonetic information. Moreover, it makes no difference that the listener knows, or can determine on auditory grounds, that the stimulus was not humanly produced; because linguistic perception is informationally encapsulated and mandatory, he will hear synthetic speech as speech.

As might be expected, the linguistic module is also very good at excluding from consideration the acoustic effects of unrelated objects and events in the environment; the resistance of speech perception to noise and distortion is well known. These other objects and events are still perceived, because they are dealt with by other modules, but they do not, within surprisingly wide

limits, interfere with speech perception (cf. Darwin, 1984). On the other hand, the module is not necessarily prepared for nonecological conditions, as the phenomenon of duplex perception illustrates. Under the conditions of duplex perception the module makes a mistake it would never normally make: it treats the same acoustic information both as speech and as nonspeech. And, being an informationally encapsulated and mandatorily operating mechanism, it keeps on making the same mistake, whatever the knowledge or preference of the listener.

Our claim that the invariants of speech perception are phonetic gestures is much easier to reconcile with a modular account of linguistic perception than with a cognitive account. On the latter view, the gestures would have to be inferred from an auditory representation of the signal by some cognitive process, and this does not seem to be a task that would be particularly congenial to cognition. Parsing a sentence may seem to bear some distant resemblance to the proving of theorems, but disentangling the mutually confounding auditory effects of overlapping articulations surely does not. It is thus quite reasonable for proponents of a cognitive account to reject the possibility that the invariants are motoric and to insist that they are to be found at or near the auditory surface, heuristic matching of auditory tokens to auditory prototypes being perfectly plausible as a cognitive process.

Such difficulties do not arise for our claim on the modular account. If the invariants of speech are phonetic gestures, it merely makes the domain of linguistic perception more suitably eccentric; if the invariants were auditory, the case for a separate linguistic module would be the less compelling. Moreover, computing these invariants from the acoustic signal is a task for which there is no obvious parallel among cognitive processes. What is required for this task is not a heuristic process that draws on some general cognitive ability or on knowledge of the world, but a special-purpose computational device that relates gestural properties to the acoustic patterns.

It remains, then, to say how the set of possible gestures is specified for the perceiver. Does it depend on tacit knowledge of a kind similar, perhaps, to that which is postulated by Chomsky to explain the universal constraints on syntactic and phonological form? We think not, because knowledge of the acoustic-phonetic properties of the vocal tract, unlike other forms of tacit knowledge, seems to be totally inaccessible: no matter how hard they try, even post-perceptually, listeners cannot recover aspects of the process—for example, the acoustically different transitions—by which they might have arrived at the distal object. But, surely, this is just what one would expect if the specification of possible vocal-tract gestures is not tacit knowledge at all, but rather a direct consequence of the eccentric properties of the module itself. As already indicated, we have in earlier papers suggested that speech perception is accomplished by virtue of a model of the vocal tract that embodies the

relation between gestural properties and acoustic information. Now we would add that this model must be part of the very structure of the language module. In that case, there would be, by Fodor's account, an analogy with all other linguistic universals.

Perception and production: One module or two?

For want of a better word, we have spoken of the relation between speech perception and speech production as a 'link', perhaps implying thereby that these two processes, though tightly bonded, are nevertheless distinct. Much the same implication is carried, more generally, by Fodor's account of modularity, if only because his attention is almost wholly on perception. We take pains, therefore, to disown the implication of distinctness that our own remarks may have conveyed, and to put explicitly in its place the claim that, for language, perception and production are only different sides of the same coin.

To make our intention clear, we should consider how language differs from those other modular arrangements in which, as with language, perception and action both figure in some functional unity: simple reflexes, for example; or the system that automatically adjusts the posture of a diving gannet in accordance with optical information that specifies the time of contact with the surface of the water (Lee & Reddish, 1981). The point about such systems is that the stimuli do not resemble the responses, however intimate the connection between them. Hence, the detection of the stimulus and the initiation of the response must be managed by separate components of the module. Indeed, it would make no great difference if these cases were viewed as an input module hardwired to an output module.

Language is different: the neural representation of the utterance that determines the speaker's production is the distal object that the listener perceives; accordingly, speaking and listening are both regulated by the same structural constraints and the same grammar. If we were to assume two modules, one for speaking and one for listening, we should then have to explain how the same structures evolved for both, and how the representation of the grammar acquired by the listening module became available to the speaking module.

So, if it is reasonable to assume that there is such a thing as a language module, then it is even more reasonable to assume that there is only one. And if, within that module, there are subcomponents that correspond to the several levels of linguistic performance, then each of these subcomponents must deal both with perception and production. Thus, if sentence planning is the function of a particular subcomponent, then sentence parsing is a function of the same subcomponent, and similarly, *mutatis mutandis*, for speech production and speech perception. And, finally, if all this is true, then the

corresponding input and output functions must themselves be as computationally similar as the inherent asymmetry between production and perception permits, just as they are in man-made communication devices.

These speculations do not, of course, reveal the nature of the computations that the language module carries out, but they do suggest a powerful constraint on our hypotheses about them, a constraint for which there is no parallel in the case of other module systems. Thus, they caution that, among all plausible accounts of language input, we should take seriously only those that are equally plausible as accounts of language output; if a hypothesis about parsing cannot be readily restated as a hypothesis about sentence-planning, for example, we should suppose that something is wrong with it.

Whatever the weaknesses of the motor theory, it clearly does conform to this constraint, since, by its terms, speech production and speech perception are both inherently motoric. On the one side of the module, the motor gestures are not the means to sounds designed to be congenial to the ear; rather, they are, in themselves, the essential phonetic units. On the other side, the sounds are not the true objects of perception, made available for linguistic purposes in some common auditory register; rather, they only supply the information for immediate perception of the gestures.

References

Abramson, A.S. (1972) Tonal experiments with whispered Thai. In A. Valdman (Ed.), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, 31–44. The Hague: Mouton.

Alegria, J., & Noirot, E. (1982) Oriented mouthing activity in neonates: Early development of differences related to feeding experiences. In J. Mehler, S. Franck, E.C.T. Walker and M. Garrett (Eds.), *Perspectives on Mental Representation*. Hillsdale, NJ: Erlbaum.

Aslin, R.N., & Pisoni, D.B. (1980) Some developmental processes in speech perception. In G.H. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child Phonology*. New York: Academic Press.

Bailey, P.J., & Summerfield, Q. (1980) Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 536–563.

Bentin, S. & Mann, V.A. (1983) Selective effects of masking on speech and nonspeech in the duplex perception paradigm. *Haskins Laboratories Status Report on Speech Research, SR-76*, 65–85.

Berkeley, G. (1709) *An essay towards a new theory of vision*. Dublin: Printed by Aaron Rhames for Jeremy Pepyal.

Best, C.T., Hoffman, H., & Glanville, B.B. (1982) Development of infant ear asymmetries for speech and music. *Perception and Psychophysics, 31*, 75–85.

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

Best, C.T. & Studdert-Kennedy, M. (1983) Discovering phonetic coherence in acoustic patterns. In A. Cohen & M.P.R. van den Broecke (Eds.), *Abstracts of the Tenth International Congress of Phonetic Sciences*. Dordrecht, The Netherlands: Foris Publications.

Brady, P.T., House, A.S., & Stevens, K.N. (1961) Perception of sounds characterized by a rapidly changing resonant frequency. *Journal of the Acoustical Society of America, 33*, 1357–1362.

Browman, C.P. & Goldstein, L.M. (1985) Dynamic modeling of phonetic structure. In V. Fromkin (Ed.),

Phonetic Linguistics. New York: Academic Press.

Brown E.L. & Deffenbacher, K. (1979) *Perception and the Senses*. New York: Oxford University Press.

Carney, A.E., Widin, G.P., & Viemeister, N.F. (1977) Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America, 62*, 961–970.

Chistovich, L.A. (1960) Classification of rapidly repeated speech sounds. *Akustichneskii Zhurnal, 6*, 392–398. Trans. in *Soviet Physics-Acoustics, 6*, 393–398 (1961).

Cole, R.A. & Scott, B. (1974) Toward a theory of speech perception. *Psychological Review, 81*, 348–374.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., & Gerstman, L.J. (1952) Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America, 24*, 597–606.

Crowder, R.G. & Morton, J. (1969) Pre-categorical acoustic storage (PAS). *Perception and Psychophysics, 5*, 365–373.

Cutting, J.E. & Rosner, B.S. (1974) Categories and boundaries in speech and music. *Perception and Psychophysics, 16*, 564–570.

Darwin, C.J. (1984) Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America, 76*, 1636–1647.

Dorman, M.F., Raphael, L.J., & Liberman, A.M. (1979) Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America, 65*, 1518–1532.

Dorman, M.F., Studdert-Kennedy, M., & Raphael, L.J. (1977) Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics, 22*, 109–122.

Dudley, H. (1940) The carrier nature of speech. *Bell Systems Technical Journal, 19*, 495–515.

Eimas, P., Siqueland, E.R., Jusczyk, P., & Vigorito, J. (1971) Speech perception in early infancy. *Science, 171*, 304–306.

Entus, A.K. (1977) Hemispheric asymmetry in processing dichotically presented speech and nonspeech stimuli by infants. In S.J. Segalowitz and F.A. Greber (Eds.), *Language Development and Neurological Theory*. New York: Academic Press.

Fant, C.G.M. (1962) Descriptive analysis of the acoustic aspects of speech. *Logos, 5*, 3–17.

Festinger, L., Burnham, C.A., Ono, H., & Bamber, D. (1967) Efference and the conscious experience of perception. *Journal of Experimental Psychology Monograph, 74*, (4, Pt. 2).

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

Fodor, J. (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fowler, C.A. (1982) Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General, 112*, 386–412.

Fowler, C.A. (1984) Segmentation of coarticulated speech in perception. *Perception and Psychophysics, 36*, 359–368.

Fowler, C.A., Rubin, P., Remez, R.E., & Turvey, M.T. (1980) Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language Production*. New York: Academic Press.

Fujisaki, M. & Kawashima, T. (1970) Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute* (Faculty of Engineering, University of Tokyo), *29*, 207–214.

Gerhardt, H.C. & Rheinlaender, J. (1982) Localization of an elevated sound source by the green tree frog. *Science. 217*, 663–664.

Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Hafter, E.R. (1984) Spatial hearing and the duplex theory: How viable is the model? In G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.), *Dynamic Aspects of Neocortical Function*. New York: Wiley.

Harris, K.S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech, 1*, 1–7.

Hillenbrand, J. (1984) Perception of sine-wave analogs of voice onset time stimuli. *Journal of the Acoustical Society of America, 75*, 231–240.

Hirsh, I.J. (1959) Auditory perception of temporal order. *Journal of the Acoustical Society of America, 31*, 759–767.

Hoffman, H.S. (1958) Study of some cues in the perception of the voiced stop consonants. *Journal of the Acoustical Society of America, 30*, 1035–1041.

Howell, P. & Rosen, S. (1983) Closure and frication measurements and perceptual integration of temporal cues for the voiceless affricate/fricative contrast. *Speech Hearing and Language Work in Progress*. University College London, Department of Phonetics and Linguistics.

Hoy, R., Hahn, J., & Paul, R.C. (1977) Hybrid cricket auditory behavior: Evidence for genetic coupling in animal communication. *Science, 195*, 82–83.

Hoy, R. & Paul, R.C. (1973) Genetic control of song specificity in crickets. *Science, 180*, 82–83.

Inoue, A. (1984) A perceptual study of Japanese voiceless vowels and its implications for the phonological analysis of voiceless consonants. Unpublished manuscript.

Jenkins, J.J., Strange, W., & Edman, T.R. (1983) Identification of vowels in 'voiceless' syllables. *Perception and Psychophysics, 34*, 441–450.

Joos, M. (1948) Acoustic phonetics. *Language Monograph 23*, Supplement to *Language, 24*.

Julesz, B. (1960) Binocular depth perception of computer-generated patterns. *Bell System Technical Journal 39*, 1125–1162.

Julesz, B. (1971) *Foundations of Cyclopean Perception*. Chicago: University of Chicago Press.

Katz, L.C. & Gurney, M.E. (1981) Auditory responses in the zebra finch's motor system for song. *Brain Research, 221*, 192–197.

Knudsen, E.I. (1982) Auditory and visual maps of space in the optic tectum of the owl. *Journal of Neuroscience, 2*, 1117–1194.

Knudsen, E.I. (1984) Synthesis of a neural map of auditory space in the owl. In G.M. Edelman, W.E. Gall, & W.M. Cowan, *Dynamic Aspects of Neocortical Function*. New York: Wiley.

Kuhl, P.K. (1981) Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America, 70*, 340–349.

Kuhl, P.K. & Meltzoff, A.N. (1982) The bimodal perception of speech in infancy. *Science, 218*, 1138–1144.

Kuhl, P.K. & Miller, J.D. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*, 69.

Ladefoged, P. & McKinney, N. (1963) Loudness, sound pressure, and subglottal pressure in speech. *Journal of the Acoustical Society of America, 35*, 454–460.

Lee, D.N. & Reddish, P.E. (1981) Plummeting gannets: A paradigm of ecological optics. *Nature, 293*, 293–294.

Liberman, A.M. (1979) Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jorgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the IXth International Congress of Phonetic Sciences*. Copenhagen: University of Copenhagen.

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

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review, 74*, 431–461.

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

Liberman, A.M., Isenberg, D., & Rakerd, B. (1981) Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics, 30*, 133–143.

Liberman, A.M., Mattingly, I.G., & Turvey, M. (1972) Language codes and memory codes. In A.W. Melton and E. Martin (Eds.), *Coding Processes and Human Memory*. Washington, DC: Winston.

Liberman, A.M., & Studdert-Kennedy, M. (1978) Phonetic perception. In R. Held, H.W. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of Sensory Physiology, Vol. VIII: Perception*. New York: Springer-Verlag.

Lisker, L. (1957) Closure duration, first-formant transitions, and the voiced-voiceless contrast of intervocalic stops. *Haskins Laboratories Quarterly Progress Report, 23*, Appendix 1.

Lisker, L. (1978) Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, SR-54*, 127–132.

Lisker, L. & Abramson, A. (1964) A cross-language study of voicing in initial stops: Acoustical measurement. *Word, 20*, 384–422.

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

MacNeilage, P.F., Rootes, T.P., & Chase, R.A. (1967) Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. *Journal of Speech and Hearing Research, 10*, 449–468.

Magen, H. (1984) Vowel-to-vowel coarticulation in English and Japanese. *Journal of the Acoustical Society of America, 75*, S41.

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

Mann, V.A. & Repp, B.H. (1980) Influence of vocalic context on the perception of [ʃ]-[s] distinction: I. Temporal factors. *Perception and Psychophysics, 28*, 213–228.

Mann, V.A. & Repp, B.H. (1981) Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69*, 548–558.

Manuel, S.Y. & Krakow, R.A. (1984) Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Report on Speech Research, SR-77/78*, 69–78.

Margolish, D. (1983) Acoustic parameters underlying the responses of song specific neurons in the white-crowned sparrow. *Journal of Neuroscience, 3*, 1039–1057.

Marler, P. (1970) Birdsong and speech development: Could there be parallels? *American Scientist, 58*, 669–673.

Martin, J.G. & Bunnell, H.T. (1981) Perception of anticipatory coarticulation effects in /stri, stru/ sequences. *Journal of the Acoustical Society of America, 69*, S92.

Martin, J.G. & Bunnell, H.T. (1982) Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 473–488.

Mattingly, I.G. (1976) Phonetic prerequisites for first-language acquisition. In W. Von Raffler-Engel, & Y. Lebrun (Eds.), *Baby Talk and Infant Speech*. Lisse, The Netherlands: Swets & Zeitlinger.

Mattingly, I.G. & Liberman, A.M. (1969) The speech code and the physiology of language. In K.N. Leibovic (Ed.), *Information Processing in the Nervous System*. New York: Springer-Verlag.

Mattingly, I.G. & Liberman, A.M. (1985) Verticality unparalleled. *The Behavioral and Brain Sciences, 8*, 24–26.

Mattingly, I.G., Liberman, A.M., Syrdal, A.M., & Halwes, T. (1971) Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2*, 131–157.

McCasland, J.S. & Konishi, M. (1983) Interaction between auditory and motor activities in an avian song control nucleus. *Proceedings of the National Academy of Sciences, 78*, 7815–7819.

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

Miller, J.D. (1977) Perception of speech sounds in animals: Evidence for speech processing by mammalian auditory mechanisms. In T.H. Bullock (Ed.), *Recognition of Complex Acoustic Signals* (Life Sciences Research Report 5), p. 49. Berlin: Dahlem Konferenzen.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A.M., Jenkins, J.J., & Fujimara, O. (1975) An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics, 18*, 331–340.

Molfese, D.L., Freeman, R.B., & Palermo, D.S. (1975) The ontogeny of brain lateralization for speech and nonspeech stimuli. *Brain and Language, 2*, 356–368.

Nottebohm, F., Stokes, T.M., & Leonard, C.M. (1976) Central control of song in the canary, Serinus canarius. *Journal of Comparative Neurology, 165*, 457–486.

O'Connor, J.D., Gerstman, L.J., Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1957) Acoustic cues for the perception of initial /w, r, l/ in English. *Word, 13*, 25–43.

Oden, G.C. & Massaro, D.W. (1978) Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

Ohman, S.E.G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151–168.

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

Pisoni, D.B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*, 253–260.

Pisoni, D.B. (1977) Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. *Journal of the Acoustical Society of America*, *61*, 1352–1361.

Pisoni, D.B. & Tash, J. (1974) Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, *15*, 285–290.

Poggio, G.F. (1984) Processing of stereoscopic information in primate visual cortex. In G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.), *Dynamic Aspects of Neocortical Function*. New York: Wiley.

Rakerd, B., Dechovitz, D.R., & Verbrugge, R.R. (1982) An effect of sentence finality on the phonetic significance of silence. *Language and Speech*, *25*, 267–282.

Rand, T.C. (1974) Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, *55*, 678–680.

Recasens, D. (1984) Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America*, *76*, 1624–1635.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981) Speech perception without traditional speech cues. *Science*, *212*, 947–950.

Remez, R.E. & Rubin, P.E. (1984) On the perception of intonation from sinusoidal signals: Tone height and contour. *Journal of the Acoustical Society of America*, *75*, S39.

Repp, B.H. (1984) The role of release bursts in the perception of [s]-stop clusters. *Journal of the Acoustical Society of America*, *75*, 1219–1230.

Repp, B.H. & Liberman, A.M. (in press) Phonetic categories are flexible. In S. Harnad (Ed.), *Categorical Perception*. Cambridge: Cambridge University Press.

Repp, B.H., Liberman, A.M., Eccardt, T., & Pesetzky, D. (1978) Perceptual integration of acoustic cues for stop, fricative and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 621–637.

Repp, B.H., Milburn, C., & Ashkenas, J. (1983) Duplex perception: Confirmation of fusion. *Perception and Psychophysics*, *33*, 333–337.

Rosen, S.M. & Howell, P. (1981) Plucks and bows are not categorically perceived. *Perception and Psychophysics*, *30*, 156–168.

Sondhi, M.M. (1979) Estimation of vocal-tract areas: the need for acoustical measurements. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *ASSP-27*, 268–273.

Samuel, A.G. (1977) The effect of discrimination training on speech perception: Noncategorical perception. *Perception and Psychophysics*, *22*, 321–330.

Segalowitz, S.J. & Chapman, J.S. (1980) Cerebral asymmetry for speech in neonates: A behavioral measure. *Brain and Language*, *9*, 281–288.

Stetson, R.H. (1951) *Motor Phonetics: A Study of Speech Movements in Action*. Amsterdam: North-Holland.

Stevens, K.N. (1975) The potential role of property detectors in the perception of consonants. In G. Fant & M.A. Tatham (Eds.) *Auditory Analysis and Perception of Speech*. New York: Academic Press.

Stevens, K.N. & Halle, M. (1967) Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.). *Models for the Perception of Speech and Visual Form*. Cambridge, MA: MIT Press.

Stevens, K.N. & Blumstein, S.E. (1978) Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *64*, 1358–1368.

Strange, W., Jenkins, J.J., & Johnson, T.L. (1983) Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, *74*, 695–705.

Studdert-Kennedy, M. (1976) Speech perception. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics*. New York: Academic Press.

Studdert-Kennedy, M. & Williams, D.R. (1984) Range effects for speech and nonspeech judgments of sine wave stimuli. *Journal of the Acoustical Society of America, 75*, S64.

Suga, N. (1984) The extent to which bisonar information is represented in the auditory cortex. In G.M. Edelman, W.E. Gall, & W.M. Cowan (Eds.), *Dynamic Aspects of Neocortical Function*. New York: Wiley.

Summerfield, Q. (1982) Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. *Journal of the Acoustical Society of America, 72*, 51–61.

Summerfield, Q. & Haggard, M. (1977) On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America, 62*, 436–448.

Thorpe, W.H. (1958) The learning of song patterns by birds, with especial reference to the song of the chaffinch, Fringilla coelebs. *Ibis, 100*, 535–570.

Tuller, B., Harris, K., & Kelso, J.A.S. (1982) Stress and rate: Differential transformations of articulation. *Journal of the Acoustical Society of America, 71*, 1534–1543.

Tuller, B. & Kelso, J.A.S. (1984) The relative timing of articulatory gestures: Evidence for relational invariants. *Journal of the Acoustical Society of America, 76*, 1030–1036.

Turvey, M. (1977) Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Physiology*. Hillsdale, NJ: Erlbaum.

Vargha-Khadem, F. & Corballis, M. (1979) Cerebral asymmetry in infants. *Brain and Language, 8*, 1–9.

Washburn, M.F. (1926) Gestalt Psychology and Motor Psychology. *American Journal of Psychology, 37*, 516–520.

Watson, J.B. (1919) *Psychology from the Standpoint of a Behaviorist*. Philadelphia: J.B. Lippincott Co.

Werker, J.F. & Tees, R.C. (1984) Cross-language speech perception: Evidence for perceptual organization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Whalen, D.H. (1981) Effects of vocalic formant transition and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America, 69*, 275–282.

Williams, H. (1984) *A motor theory of bird song perception*. Unpublished doctoral dissertation. The Rockefeller University.

Williams, D.R., Verbrugge, R.R., & Studdert-Kennedy, M. (1983) Judging sine wave stimuli as speech and nonspeech. *Journal of the Acoustical Society of America, 74*, S66.

Witelson, S. (1977) Early hemispher specialization and interhemispheric plasticity: An empirical and theoretical review. In S.J. Segalowitz & F.A. Gruber (Eds.), *Language Development and Neurological Theory*. New York: Academic Press.

Yeni-Komshian, G.H. & Soli, S.D. (1981) Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 70*, 966–975.

Résumé

Une théorie motrice de la perception proposée initialement pour rendre compte des résultats des premières expériences avec de la parole synthétique a été largement révisée afin d'interpréter les données récentes et de relier les propositions de cette théorie à celles que l'on peut faire pour d'autres modalités de perception. La révision de cette théorie stipule que l'information phonétique est fournie par un système biologique distinct, un 'module' spécialisé pour détecter les gestes que le locuteur a eu l'intention de faire: ces gestes fondent les catégories phonétiques. La relation entre les gestes et les patterns acoustiques dans lesquels ceux-ci sont imbriqués de facon variée est unique mais régulée. Cette relation est construite dans la structure du module. En conséquence le module provoque la perception de la structure phonétique sans traduction à partir d'impressions auditives préliminaires. Ce module est ainsi comparable à d'autres modules tels que celui qui permet à l'animal de localiser les sons. La particularité de ce module tient à la relation entre perception et production qu'il incorpore et au fait qu'il doit rivaliser avec d'autres modules pour de mêmes variations de stimulus.