Analytic Approaches to Human Cognition
J. Alegria, D. Holender, J. Junça de Morais and M. Radeau (eds.)

813

CHAPTER 2

Plausibility, Parsimony, and Theories of Speech

Alvin M. Liberman
Haskins Laboratories
New Haven

ABSTRACT

According to a somewhat unconventional view, speech is managed by a specialization for language—a phonetic module—at the level of action and perception. There, the processes and primitives are specifically phonetic, not, as in more commonly assumed, generally motor and auditory. The less conventional view is nevertheless the more plausible because it (1) better illuminates the biological nature of the difference between spoken and written forms of language, and (2) provides the better account of how speech meets the specific requirement of phonological communication that the elements be commutable, as well as the general requirement of all communication systems that there be parity between sender and receiver. Also relevant to the argument of plausibility is the fact that, while the phonetic module is unique to language, it is not without biological precedent, since it has important properties in common with such older (and better understood) specializations as stereopsis and sound localization.

It is, for me, a happy privilege to be part of an occasion that honors Paul Bertelson, dear friend and valued colleague. As my contribution to the occasion, I offer a few reflections on a question I have often discussed with Paul: Is there a specialization for language at the precognitive level? Is there, in other words, a specifically linguistic mode of action and perception? Put in one form or another, this question goes to the heart of claims about the modular nature of linguistic processes. It arises wherever in language one happens to look, but it assumes what I take to be its most pointed manifestation at the level of phonetic structure. There lie two or three dozen consonants and vowels, familiar objects of a seemingly simple sort. Yet they are the elements of which all languages are made. Moreover, their proper use is a distinguishing mark of the human species and a principal component of its linguistic faculty. Accordingly, the question I raise about their management is a question about the biology of language.

Together with some of my colleagues, including especially Ignatius Mattingly, I believe the answer to the question is yes—the biology of

language does, indeed, incorporate a precognitive specialization for the production and perception of consonants and vowels, a specialization we have chosen to call a phonetic module. We take this module to be an integral part of the larger specialization for language, adopting what Fodor (1983) would characterize as a *vertical* view in which the relevant structures and processes are seen as specific to the linguistic function they serve. The opposite view, which is more widely held, is that speech is to be accounted for by the most general principles of motor activity and auditory perception; accordingly, this view is appropriately referred to as *horizontal*.

My aim in this paper is to promote the less conventional vertical view, not by reference to the results of particular and putatively critical experiments, but rather by taking account, in very general form, of a few commonly neglected considerations that are relevant to its plausibility and parsimony. A fuller description of the vertical view, together with an account of the nature of its empirical support, is to be found elsewhere (Liberman & Mattingly, 1985; Liberman & Mattingly, 1989; Mattingly & Liberman, 1988). As will be seen there, this view comprehends both the production and perception of speech; indeed, it assumes an organic relation between the two. It happens, however, that the considerations I mean to offer in this paper are concerned primarily with perception, so I will bias the emphasis in that direction, as I do in the following brief account of the difference between the vertical view and its horizontal opposite.

The horizontal view varies in its particulars from one theorist to another, but the basic assumptions are much the same. Thus, the several proponents are in agreement that perception of speech is no different from perception of other sounds (Ades, 1977; Bregman, 1991; Cole & Scott, 1974; Crowder & Morton, 1969; Diehl & Kluender, 1989a; Fujisaki & Kawashima, 1970; Howell & Rosen, 1984; Kuhl, 1981; Miller, 1977; Lane, 1965; Lindblom, 1991; Oden & Massaro, 1978; Stevens, 1981). All such perception is supposed to depend on the same general processes of hearing, processes that occupy a common domain and evoke in a common sensory register a common set of auditory primitives, including, for example, pitch, loudness, and timbre. Of course, the perceptual representations of a stop consonant and, say, a squeaking door must be different, but the difference is supposed to be only in the relative values that are assigned to the primitives they have in common; there are no specifically phonetic primitives. Thus, the primary perceptual representations of speech are taken to be generally auditory, not specifically phonetic. That being so, proponents of the horizontal view are required to explain how, being independent of language, the auditory representations gain access to a system in which they are specifically marked for linguistic significance and used for a specifically linguistic purpose.

Some proponents explicitly meet this requirement by supposing that, given the auditory percepts, the listener elevates them to linguistic status by attaching phonetic labels, fitting them to phonetic prototypes, or associating

them with such cognitive units as *distinctive features* (Ades, 1977; Crowder & Morton, 1969; Fujisaki & Kawashima, 1970; Pisoni, 1973; Rosen & Howell, 1987; Stevens, 1975, 1989). Since these labels, prototypes, and features are neither acts nor percepts, they deserve to be called ideas. But whatever they are called, they are the end products of a cognitive translation that converts auditory percepts into a form appropriate to language. Getting from speech signal to the primary level of language is, therefore, a two-stage process: evocation of an auditory percept in the first stage, followed by conversion to a phonetic representation in the second. In this important respect, the horizontal view implausibly makes perceiving speech no different in principle from perceiving Morse code or, for that matter, the letters of the alphabet; in all cases, the perceiver must attribute linguistic significance to percepts that are not inherently linguistic (see Liberman, in press, for further discussion).

There are at least two other assumptions of the horizontal view, but these are commonly left unsaid, though they are, to the vertical theorist, of great importance. One, which seems to be tacitly accepted, not as an assumption but as background fact, is that phonetic elements are sounds. The other, which is commonly unspoken because it must appear on this view to be irrelevant, is that the gestures and motor control processes of speech production are, like the processes of speech perception, independent of language. Presumably, language simply appropriated movements and motor mechanisms that are part of a general faculty for action, just as it appropriated for its own special purposes the general mechanisms of audition. It is, therefore, necessary for the speaker, just as it is for the listener, to make a cognitive translation between two very different kinds of representations, one linguistic, the other not. According to the horizontal view, then, it should not matter in this regard whether one produces language by speaking it, by operating a Morse-code key, or by wielding a pen. Putting this observation about production together with the earlier one about perception, we see that the horizontal view must fail in both domains to provide a plausible basis for distinguishing the biologically primary processes of speech from their obviously secondary extensions.

The vertical view is different at all points. Seen vertically, apprehending phonetic structures is managed by a distinct, language-specific system that has its own phonetic domain, its own phonetic mode of signal processing, and its own phonetic primitives. Perception of phonetic structure is therefore precognitive, which is to say immediate; there is no translation from a nonphonetic (auditory) representation because there is no such representation. It is, of course, in precisely this respect that perception of speech differs, plausibly, from perception of Morse code or of scripts.

There are two other assumptions of the vertical view that contrast starkly with its conventional counterpart. One is that the elements of phonetic structure are gestures, not the sounds those gestures produce. These acts are, then, the ultimate constituents of language, the primitives that must be exchanged between speaker and listener if communication by language is to

occur. The second assumption is that these gestures, as well as the processes that control them, are specifically phonetic, having evolved for phonological communication and for nothing else. Unlike a Morse code operator or a writer, a speaker is directly using motor representations that are inherently linguistic. There is no need to connect a nonlinguistic act (pressing a key or writing an alphabetic character) to some linguistic unit of a cognitive sort. Nor is such a unit required, more generally, to serve as a common referent through which a nonlinguistic act and a correspondingly nonlinguistic (auditory) percept can be connected to each other. On the vertical view, the specifically phonetic gestures that are managed by the module in production are recovered by the module as the specifically phonetic primitives of perception, thereby completing the communicative link without cognitive intervention, while also making speech an integral part of language, not, as on the horizontal view, an artifactual adjunct.

ARE THERE ACOUSTIC SUBSTITUTES FOR SPEECH?

According to the horizontal view, speech percepts are supposed to be auditory in the same way that the percepts evoked by the letters of the alphabet are known to be visual. In the visual case, the only limit to the number and variety of optical shapes that can be made to serve as alphabetic characters is in the constraints imposed by the visual system, and they are few. Given the conventional view of speech, one would suppose that a similar situation would exist there. Of course, the auditory channel is neither so wide nor so deep as the visual, but, still, the number of sounds that can be identified is very great, so one should expect that it would be possible, even easy, to find alternative acoustic vehicles.

The foregoing implication of the horizontal view is exactly what my colleagues and I tacitly accepted when, in 1945, we were enlisted in an attempt to build a device that would convert print into intelligible sound and so serve as a reading machine for the blind. We should, of course, have wanted a machine that would make the print speak English, but there were at the time no such things as optical character readers, and, even if there had been, we should not have known how to synthesize speech from their outputs. However, we considered this to be of no great consequence, for we could quite easily make the print control the parameters of various nonspeech sounds, and so produce an acoustic cipher differing only in detail from the speech to which the blind users were accustomed. Given our tacit assumptions about the nature of speech, we supposed that they would learn to connect these sounds to phonetic units, much as they had earlier done with the sounds of speech.

A detailed account of our unsuccessful attempts to substitute nonspeech sounds for the sounds of speech would not be enlightening here, for it would only make the point that, try as we might, we did not come anywhere near to

succeeding. Of course, we could not then, and cannot now, expect to test all possible sounds, nor could we readily arrange for people to have with nonspeech the amount of experience they must have had with speech. Still, we were then, as we are now, convinced that nonspeech sounds simply will not do, not just because they failed the tests we put them to, but because they failed in ways that made it plain why we should never have expected them to succeed. The difficulty was not primarily that the sounds were indiscriminable or unidentifiable, but rather that every arrangement we tried was defeated in one way or another by the variable of rate. Thus, we found that, as the rate of scan approached the lower bound of what would be even marginally acceptable in speech or in reading, performance (as measured by ability to learn a selected set of words) decreased appreciably. Worse yet, listeners lost the ability to identify the individual letter sounds and to apprehend their order, responding instead to some overall auditory pattern characteristic of the word. Thus, to the extent that the words could be learned at all, they had to be treated logophonically, as it were, with attention directed to the way the sound differed holistically from the sound for any other word. The tremendous advantage of the combinatorial principle that phonology exploits was therefore lost, and, given that a purely logographic system cannot really work very well even in reading (De Francis, 1989; Mattingly, in press), one can imagine how vastly more unsuited it would be as a basis for speech perception.

The final blow was dealt by our observation that when we ourselves undertook to master one of these nonspeech systems, we found little transfer of training across rates. Letters and words learned at one rate could not be recognized at other rates that were still within the range of what was reasonable if the machine was to have any utility. Words tended not only to become hard-to-analyze wholes, but the phenomenal nature of the whole changed quite drastically from one rate to another. A user would have been required, therefore, to learn a different set of associations for every significantly different rate.

In hindsight, it is apparent that if we had ever bothered to think about the requirements of phonological communication, and then measured these against the known properties of the ear, we should have realized, without any research at all, that an acoustic-auditory strategy of the kind suggested by the horizontal view was bound to fail. The point is that phonological communication requires commutable, hence discrete and invariant, representations. But if such invariance is to exist in the auditory domain, as it must on the view that we had unthinkingly adopted, then rates of transmission that are normal in speech would seriously strain and sometimes overreach the temporal resolving power of the ear and also its ability to perceive the order of the segments (Liberman, Cooper, & Studdert-Kennedy, 1968). (Speech production would be equally problematic, since invariant and discrete auditory percepts would require correspondingly invariant and discrete gestures, with the result that people could not really speak, they could only spell).

But we had to learn the hard way, as it were, that nonspeech sounds—that is, sounds that do not approximate the results of linguistically significant gestures—cannot be efficient vehicles for language. It was, indeed, this painfully-arrived-at conclusion that initially motivated Frank Cooper and me to begin our speech research. Our aim, very simply, was to find out why the sounds of speech, but no others, can meet the commutability and rate requirements of phonological communication. The answer our research brought us to seems to me now so plausible, not to say obvious, that I wonder we did not arrive at it earlier, simply by thinking about the matter. For what it comes to is that evolution did not ever confront the problems of commutability and rate, simply because it avoided the acoustic-auditory strategy (of the horizontal view) that would have given rise to them. What evolved was a brilliantly successful strategy that defined the invariant elements of phonetic structure not as sounds, but as gestures. The critically important advantage of this strategy was that, given gestures that can somehow be characterized as remote structures of motor control, and given a mode of action specifically adapted to matching these to the needs of phonology, it was possible by overlapping and merging (that is, coarticulation) of the peripheral movements to achieve the high rates of production that characterize speech communication.

As for perception, which was initially our single-minded concern, the advantage is that coarticulation effects parallel transmission of information about successive phonetic segments, and so relaxes the constraints on rate of perception that underlay the failure of our nonspeech reading machines. But this gain has an obvious cost, for coarticulation creates a complex relation between signal and message, a specifically phonetic code that is opaque except as the scientist or perceiving device can take account of the phonetically specific processes that produced it. Once research on speech had convinced us that this was so, we felt challenged to explain, if only in the most general terms, how listeners manage. We rejected the possibility that they *break* the code by some deliberate, cognitive process, preferring, instead, to suppose that they rely on a biologically coherent module specifically adapted to providing the articulatory key. But whatever the plausibility of this proposed solution, it was never plausible to suppose that perception of linguistic structure is so much controlled by general auditory processes that it can be achieved as well with sounds other than speech. That we nevertheless thought it was is testimony to the unquestioning faith we had in what was then, and is now, the received view.

WHENCE COMES THE FIT OF PERCEPTUAL FORM TO PHONOLOGICAL FUNCTION?

Given that the function of phonology is to use the combinatorial principle to generate a large number of words, the units must, as already noted, be discrete and invariant, which is to say categorical, as they are seen from a

linguistic point of view. It is adaptive therefore that the units be correspondingly categorical in immediate perception. Listeners would only be disconcerted by the sense, if it should be their sense, that a particular phonetic token, X, lay half way between X and Y, or that it really sounded like Z, except as it was reinterpreted so as to take account of the fact that it was followed by A. Fortunately, listeners do not have either sense: The much-investigated peaks of discriminability at the acoustic boundaries of the phonetic unit reflect category-producing discontinuities in perception, and it is characteristic of phonetic perception that these categories remain stable across all context-conditioned variation in the stimulus.

What, then, is the source of these stable perceptual categories? On the horizontal view, it must, of course, be in the properties of the auditory system. Accordingly, theorists of this persuasion take comfort in the experiments that find categories in the responses of nonhuman animals to speech and in the responses of human listeners to acoustic nonspeech analogues (Diehl & Walsh, 1989; Kluender, 1991; Kluender, Diehl, & Killeen, 1987; Kluender, Diehl, & Wright, 1988; Kuhl & Miller, 1975; Massaro, 1987; Parker, 1988; Parker, Diehl, & Kluender, 1986; Pastore, 1987; Pisoni, 1973; Pisoni, Carrell, & Gans, 1983). The opposite result is also found, much to the satisfaction of the vertical theorists, who must believe that this kind of categorical perception is specifically phonetic (Best, Morrongiello, & Robson, 1981; Best, Studdert-Kennedy, Manuel, Rubin-Spitz, 1989; Mann & Liberman, 1983; Liberman, Isenberg, & Rakerd, 1981; Mattingly, Liberman, Syrdal, & Halwes, 1971; Sinnott, 1976; Waters & Wilson, 1976). However, I do not mean here to offer a critical evaluation of the experimental evidence pro and con the one assumption or the other, but, rather, in keeping with the spirit of this paper, to argue that the horizontal (auditory) interpretation is simply implausible on its face.

It is relevant, first, to take into account how very great is the variation in stimulus for any given perceptual category (Repp & Liberman, 1987). For all phones, there is variation as a function of phonetic context, position in the syllable, and vocal-tract size. In some cases, there are changes depending on articulatory rate and stress. And, of course, there are the differences that exist across languages. Indeed, so gross is this stimulus variation, and so numerous its sources, that it is impossible to estimate how very many alternative category boundaries the auditory system would need if the percepts were to be held constant, and implausible to suppose that these boundaries could exist in such numbers. Surely, they could not have been selected in the evolution of the auditory system just against the possibility that phonology would one day come along and find them useful. Yet, as properties of the auditory system, they serve no other imaginable purpose. Indeed, from an auditory standpoint, they would be dysfunctional, since they would necessarily distort the perception of nonspeech sounds.

Even if one assumes, against all reason, that this numerous variety of boundaries does exist in the auditory system, is it plausible to suppose that coarticulatory manoeuvres vary as they do with phonetic context and with rate just in order to produce sounds that match the way categories of the auditory system happen, independently of coarticulation, to adjust to variation in the acoustic stimulus?

Moving, now, from implausibility to impossibility, I remark the fact that, as is well known, the articulation of every phonetic unit has multiple acoustic consequences, and that listeners are more or less sensitive to all of them. So, if speakers had somehow managed to produce a second-formant transition to fit some auditory category, what then would they do about the third-formant transition and the burst? The answer has got to be nothing, since it is not possible to control these acoustic consequences independently.

It is also true of these multiple sources of information that, no matter how numerous and acoustically various they may be, they nevertheless evoke a unitary, categorical percept. This equivalence of the acoustically very different components of the speech signal is reflected in, and measured by, the trading relations, so-called, that speech researchers report (Diehl & Kluender, 1989b; Fitch, Halwes, Erickson, & Liberman, 1980; Repp, 1982). But one hardly needs experiments like those to make the point. For, surely, there is no doubt that there are multiple and acoustically very different sources of acoustic information for every phone, and it is common experience that the result is a unitary perceptual category, not a collage in which the several fragments represent the disparate auditory consequences of the different acoustic cues. Is it even conceivable that speakers produce these heterogeneous combinations of sounds by design, and that they do so because they once discovered that the auditory system just happens to cause them to evoke the same percept. It would, again, be dysfunctional if the auditory system did that, for it would effectively prevent the discrimination (or identification) of most ordinary acoustic events; indeed, it would tend to make all of them sound like speech.

Nor can one reasonably suppose that such categories as the auditory system apparently does have might somehow have served as starting points for the development of phonetic perception (Kuhl, 1981). Which contexts, rates, vocal-tract sizes, and languages might have been taken as the linguistic canon? And even if these auditory categories are appropriate in some phonetic circumstances, would they not be inappropriate, hence dysfunctional, in all others? Indeed, auditory categories, to the extent that they exist, should make us the more convinced of the validity of the vertical view, since they require of the phonetic system that it be so independent as to ignore their potentially interfering representations.

Is it not far more plausible to suppose about all these cases that the variable and multiple sources of information in the speech signal are simply the inevitable consequences of acts that are specifically adapted to a phonological

function, and that perception is managed by a corresponding adaptation to those same acts and that same function?

WHAT IS THE PLACE OF SPEECH IN THE BIOLOGICAL SCHEME OF THINGS?

If, as the horizontal view would have it, there is no specialization for language at the level of action and perception, then, as I have already implied, language must begin one step up, where, by a purely cognitive process, a select set of nonlinguistic representations is given a phonetic cast and so made appropriate for whatever specialized language processing the theorist wishes to assume. The same conclusion follows if the theorist should, by a seemingly logical extension, embrace the more broadly horizontal assumption that there is no specifically linguistic process at any level, that just as speech is merely one among many expressions of the general faculties of action and perception, so does syntax fall out of a general faculty of cognition. On either version, however, it will be hard to provide a parsimonious answer to a fundamental question about the biology of speech: How are the acts and percepts of speech marked in evolution for linguistic significance, and so set apart from all others?

Perhaps the most explicit attempt to answer this question from a horizontal point of view has been made by Lindblom (1991) who says that "languages make their selection of phonetic gesture inventories under the strong influence of motor and perceptual constraints that are language independent and in no way special to speech (the functional adaptation of phonetic gestures).". Then, referring to the unconventional assumption that there are specializations at the level of perception and action, he says, "If so, why do inventories of vowels and consonants show evidence of being optimized with respect to motor and perceptual limitations that must be regarded as biologically general and not at all special to speaking and listening?".

As a criticism of the vertical view, which is how it was intended, Lindblom's argument can be dismissed as irrelevant to the question that this view is designed to answer. That question is not whether language somehow evolved out of what was already there, for it could hardly have done otherwise, but, rather, what it was that evolved. Lindblom's answer is that there was, at the precognitive level, no evolution of anything, only a selection from among the possibilities offered by general faculties that were, and presumably still are, independent of language. Of course, that must have been exactly what happened in the development of, say, a cursive writing system, for surely the selection of its characters must have been strongly influenced by "motor and perceptual constraints that are language independent". But such an observation, true though it is, enlightens us not at all about the evolution of language, for what developed in the case of cursive writing were artifacts, not the biologically primary units of the language that those artifacts are taken to

represent. Obviously, the artifacts can have been marked for linguistic significance only by agreement, not by the processes of biological evolution. It is up to each user, then, to honor the agreement by mastering, at a cognitive level, the wholly arbitrary connection between the selected characters and the primary units of the language. On Lindblom's account, the same must be said of speech and the speaker-listener. For if speech production and perception are not distinctly linguistic, the primary units of language must, as earlier noted, be in the nature of ideas (i.e., the labels, prototypes, distinctive features, etc.) to which the nonlinguistic representations of speech become connected. Such ideas might have been a result of the inventiveness that large brains and cognitive power make possible, in which case, the biology of speech would be the biology of large brains and cognitive power. Or, alternatively, they might have become part of the genetic inheritance of human beings, in which case the biology of speech would be the biology of innate ideas. In neither case would there be a place for speech in the biology of language.

According to the vertical view, the biology of speech embraces specifically phonetic structures and processes that are adapted to specific linguistic functions. What evolved, on this view, was a special mode of communication (the phonological mode), that serves a distinctly linguistic function (the generation of a large vocabulary by use of the combinatorial principle), and imposes phonology-specific requirements (among which are the rapid production and perception of commutable elements). The primitives of this mode are correspondingly special, being specifically linguistic and so appropriate for their role in the larger specialization for language, including, for example, the syntactic component. On that basis, it seems plausible to suppose that the elements and processes of the phonological mode were selected according to their ability to meet its special requirements. On the side of action, I should think that an important factor was not ease of production as such, but rather the extent to which the gestures lent themselves to the coarticulatory manoeuvres that effectively circumvent the constraints on rate that would have been imposed had discrete gestures been produced seriatim. On the perceptual side, a decisive factor must have been the immense advantage conferred by a complex kind of parallel transmission that extends the limit on rate set by the temporal resolving power of the ear. It would appear then that, so far from being driven to exploit the strengths of the general motor and auditory systems, as Lindblom's comments imply, the evolution of speech must have been guided, rather, by the need to find ways around what must be seen, from a phonological point of view, as their weaknesses. It must also have been guided, even more generally, by the need to meet the requirement of parity by establishing an identity between the communicative acts of the speaker and the communicative percepts of the listener. This it did by incorporating in the precognitive biology of speech the special mechanisms that allow articulatory gestures—the constituents of language that must be common to speaker and listener—to survive the rigors of the communicative exchange.

It is also relevant to the plausibility of a theory of speech to expose, among its biological implications, the relation of speech to other forms of natural communication. On any theory, the gulf between speech and other systems must, of course, be seen to be very wide, though one would surely be inclined to look with favor on a theory that nevertheless managed some kind of bridge. It therefore counts against the horizontal view that it fails to do that. For if there is no precognitive specialization for speech, then, as has been noted several times already, speech must be matched to phonetic ideas. The horizontal theorists apparently find that consequence acceptable as it applies to human beings and their language. But would they not hesitate to extend it to the nonhuman case? Presumably, they would, given the abundant evidence that nonhuman communication is underlain by specializations for producing and perceiving specifically communicative signals of one sort or another. Are we to suppose, then, that unlike the nonhuman animals, which communicate as they do because of the nature of their precognitive specializations, we humans speak because, having risen above that mean level, we take advantage of innate ideas and intelligence? The vertical view, on the other hand, permits us to see that we and the other creatures are all precognitively specialized for communication; the important difference is that our specialization comprises a phonology and a syntax, while theirs does not.

There remains the biologically relevant question: What more general phenomena are exemplified by the processes of speech? Here, the horizontal view might appear to have the advantage, since it takes speech production and perception to be not different from other forms of action and perception. Accordingly, speech processes are as general as those that manage all of auditory perception and all of motor activity. The vertical view, on the other hand, abjures this kind of generality, holding that speech processes are specific to the linguistic function they serve. Indeed, it is precisely on this score that the unconventional view has been criticized as unparsimonious. As I have already tried to show, however, it is just because of the assumption about special processes that the unconventional view is the more parsimonious, since assuming another precognitive specialization is presumably less in need of Occam's razor than assuming a set of innate phonetic ideas.

At all events, assuming a specialization for speech is no more unparsimonious than making the corresponding assumption for other systems that are biologically adapted to stimulus events and properties that are of great ecological significance to the species. Consider, for example, echolocation in the bat, sound localization in the barn owl, song in the bird, or, indeed, stereopsis in the human. Like the speech specialization as characterized by the vertical view, each of these is to be understood only by reference to the special mechanisms by which it serves its special function. While each system is therefore different from every other, they have in common the properties that Fodor has identified as characteristic of the modules that he takes as the functional elements of the precognitive mind. Moreover, the specializations

named above have in common with each other and with speech that they all belong to a class of modules called *closed* by Mattingly and me, and claimed by us to share the following properties (Liberman & Mattingly, 1989; Mattingly & Liberman, 1988).

1. The representations are heteromorphic. That is, the dimensions of the percept are incommensurate with the dimensions of the stimulus. Thus, in stereoscopic vision, the viewer perceives heteromorphic depth, not homomorphic disparity (doubling of images). In speech, the listener perceives, heteromorphically, a string of discrete consonants and vowels, not the continuously varying timbres (chirps, whistles, bleats, etc.) that constitute the homomorphic representations of the continuously changing formant tracks.

2. The modules preempt the stimulus information that is of interest to them, using it to form the heteromorphic percept, while leaving none for the homomorphic counterpart (Bentin & Mann, 1990; Liberman & Mattingly, 1989; Whalen & Liberman, 1987). Thus, over a range of binocular disparities, the viewer perceives depth; disparity is not also seen. In a similar way, listeners perceive phonetic structures, not phonetic structures and also the homomorphic chirps and whistles that the components of the acoustic signal would otherwise represent.

3. The modules are highly plastic, which allows them to be calibrated and recalibrated by relevant environmental conditions that accumulate over time, or that change, whether naturally or by design of an experimenter (Knudsen, 1988). Thus, stereopsis adjusts at the precognitive level to the changes in binocular disparity that occur as the child's head grows bigger. The phonetic module is similarly calibrated over time according to the phonetic environment to which it is exposed. At all events, the plasticity of these modules is so great that they accommodate stimulus patterns that fall some distance beyond what is possible ecologically. Thus, viewers perceive depth with disparities far greater than could ever be provided by the distance between the eyes. Phonetic perception is possible with a wide variety of departures from the normal acoustic structure of speech, including even sine-wave analogs of the formant tracks.

4. When the limit of plasticity is exceeded, preemptiveness fails, with the result that heteromorphic and homomorphic representations are evoked simultaneously. Thus, in stereopsis, as the disparity is progressively increased, a point is reached at which the viewer sees heteromorphic depth but also homomorphic disparity. In speech, as the experimenter introduces a discordance or discontinuity between two parts of the signal, a point is reached at which the listener perceives the heteromorphic structure but also the chirps, whistles, or bleats that constitute the homomorphic representation. As it occurs in speech, this phenomenon has come to be known as *duplex perception* (Bentin & Mann, 1990; Liberman et al., 1981;

Mann & Liberman, 1983; Rand, 1974; Whalen & Liberman, 1987). The point to be made here is simply that duplex perception is not a freak phenomenon, limited to speech, but is, rather, what happens to a closed module when, as a consequence of limits on its plasticity, it can no longer preempt the stimulus information.

5. In the case of stereopsis, it has been shown that, as the disparity is increased over the range of duplex perception, the heteromorphic percept progressively diminishes while the homomorphic percept grows until, finally, only the homomorphic percept is represented (Richards, 1971). (It is as if there were a conservation of stimulus information: some, or all, of the information goes to form the one percept, the remainder goes to the other, and vice versa. Is there, perhaps, some imaginable sense in which the perceptual *sum* can be said to remain constant?) Mattingly, Yi Xu, and I are currently testing the hypothesis that duplex perception in speech follows a course similar to that found in the duplex range of stereopsis. But whatever the outcome of this test, there is already considerable evidence for the conclusion that the properties of the phonetic module are similar to those that characterize other biological specializations for perception.

In the domain of speech, there are, then, two quite different kinds of biological generality, one for each theory. The horizontal theory claims generality by associating speech with processes that cut across a variety of perceptual, motor, and cognitive functions. The vertical view finds it in the integral relation of speech to language and in the resemblance of speech to other specializations at the precognitive level. The question, then, is not which theory relates speech more generally to other aspects of biology but rather which kind of generality corresponds more closely to the true state of affairs.

The vertical view of speech—that the constituents are gestures, not sounds, and that these constituents are managed by a phonetic specialization—is apparently rejected by most students of speech as implausible and unparsimonious: Implausible, because it flies in the face of the common-sense observation that speech consists of sounds that fall on the ear and therefore excite the auditory system; unparsimonious, because it requires the assumption of a distinct and hitherto unacknowledged mode of action and perception. My aim in this paper has been to show that the shoe is on the other foot. The general form of the argument is that the horizontal view is implausible because the nonlinguistic modalities of action and perception it relies on are manifestly ill suited to the special requirements of phonological communication; it is unparsimonious because it requires cognitive processes of one sort or another if the general auditory and motor units of speech are to be connected to language. The vertical view is designed to avoid these flaws.

REFERENCES

Ades, A. E. (1977). Vowels, consonants, speech and nonspeech. *Psychological Review*, **84**, 524-530.

Bentin, S., & Mann, V. A. (1990). Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes. *Journal of the Acoustical Society of America*, **88(1)**, 64-74.

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

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in auditory patterns. *Perception and Psychophysics*, **45**, 237-250.

Bregman, A. (1991). The compositional process in cognition with applications to speech perception. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

De Francis, J. (1989). *Visible Speech*. University of Hawaii Press: Hawaii.

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

Diehl, R. L., & Kluender, K. R. (1989b). On the categorization of speech sounds. In S. Harnad (Ed.), *Categorical Perception*. Cambridge: England.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basic for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**, 2154-2164.

Fitch, H., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception and Psychophysics*, **27(4)**, 343-350.

Fodor, J. A. (1983) *The Modularity of Mind*, MIT Press, Cambridge, Mass.

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.

Howell, P. & Rosen, S. (1984). Natural auditory sensitivities as universal determiners of phonemic contrasts. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations for Language Universals*. The Hague: Mouton.

Kluender, K. R. (1991). Effects of first formant onset properties on voicing judgments result from processes not specific to humans. *Journal of the Acoustical Society of America*, **90**, 83-96.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science*, **237**, 1195-1197.

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*, **16**, 153-169.

Knudsen, E. I. (1988). Experience shapes sound localization and auditory unit properties during development in the barn owl. In G. M. Edelman, W. E. Gall, & M. W. Cowan (Eds.), *Auditory Function: Neurobiological Bases of Hearing*, pp. 137-149, Wiley: New York.

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

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, **72**, 275-309.

Liberman, A. M. (in press). The relation of speech to reading and writing. Proceedings of the Bellagio Conference on Speech and Reading.

Liberman, A. M., Cooper, F. S., Studdert-Kennedy, M. (1968). Why are spectrograms hard to read? *American Annals of the Deaf*, **113**, 127-133.

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

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

Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics*, **30(2)**, 133-143.

Lindblom, B. (1991). The status of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Lawrence Erlbaum.

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

Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization. In S. Harnad (Ed.), *Categorical Perception*. Cambridge: England.

Mattingly, I. G. (in press). Linguistic awareness and orthographic form. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning*. Amsterdam: North Holland.

Mattingly, I. G. & Liberman, A. M. (1988). Specialized perceiving systems for speech and other biologically significant sounds. In G. M. Edelman, W. E. Gall, and W. M. Cowan (Eds.), *Functions of the Auditory System*. (pp. 775-793). New York: Wiley.

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

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.

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

Parker, E. M. (1988). Auditory constraints on the perception of voice-onset-time: The influence of lower tone frequency on judgments of tone-onset simultaneity. *Journal of the Acoustical Society of America*, **83**, 1597-1607.

Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. *Perception and Psychophysics*, **39**, 129-142.

Pastore, R. E. (1987). Categorical perception: Some psychophysical models. In S. Harnad (Ed.), *Categorical Perception*. Cambridge: England.

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., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, **34**, 314-322.

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

Repp, B. H. (1982). Phonetic trading relations and context effects: New evidence for a phonetic mode of perception. *Psychological Bulletin*, **92**, 81-110.

Repp, B. H., & Liberman, A. M. (1987). Phonetic categories are flexible. In S. Harnad (Ed.) *Categorical Perception*, (pp. 89-112). Cambridge University Press.

Richards, W. (1971). Anomalous stereoscopic depth perception. *Journal of the Optical Society of America*, **61**, 410-414.

Rosen, S., & Howell, P. (1987). Auditory, articulatory, and learning explanations of categorical perception in speech. In S. Harnad (Ed.), *Categorical Perception*. Cambridge: England.

Sinnott, J. M. (1976) Speech sound discrimination by monkeys and humans. *Journal of the Acoustical Society of America*, **60**, 687-695.

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*, Academic Press, New York.

Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Evidence from phonology, acoustics, and psychoacoustics. In T. Myers, J. Laver, & J. Anderson (Eds.), *Advances in Psychology: The Cognitive Representation of Speech*. Amsterdam: North Holland.

Stevens, K. N. (1989). On the Quantal Nature of Speech. *Journal of Phonetics*, **17**, 3-45.

Waters, R. S. & Wilson, W. A. Jr. (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. *Perception and Psychophysics*, **19**, 285-289.

Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, Vol. 23, 169-171.