728

In G.M. Edelman, W.E. Gall and W.M. Cowan (Eds.), Signal and Sense: Local and Global Order in Perceptual Maps. New York: Wiley. 1990.

Chapter 19

Speech and Other Auditory Modules

IGNATIUS G. MATTINGLY
ALVIN M. LIBERMAN

ABSTRACT

Speech perception must be counted a distinct module of the auditory system, because, as experiments show, it does not depend on the outputs of the other modules: Under appropriate conditions, a coherent speech percept is evoked by information in two signals that are perceived as separate sources; moreover, one of the signals is simultaneously also perceived as nonspeech.

The several auditory modules fall into two classes, open and closed, according to the kind of perceptual representations they produce (for example, pitch and loudness versus localized sound sources and speech) and in the way they adapt to environmental influences. Given the nature of these classes, they can hardly have equal access to the flow of stimulus information, as they would if they were in parallel. A more suitable architecture would allow the closed modules to preempt just the information that is of interest to them, and so prevent it from reaching the open modules at all. In the case of speech, there is evidence that such preemption does occur.

It is our view that perception of speech is accomplished not by ordinary auditory and cognitive mechanisms operating in ordinary ways, or even by ordinary auditory and cognitive mechanisms operating in special ways, but by a mechanism that is specific to speech and quite distinct from the others. In earlier papers, we offered this view in general terms (Liberman et al., 1967; Mattingly and Liberman, 1969; Liberman et al., 1972). More recently, we made it more specific, claiming that speech perception is the business of a phonetic (really, a linguistic) "module," as Fodor (1983) defined it (Liberman and Mattingly, 1985; Mattingly and Liberman, 1985; Mattingly and Liberman, 1988). Here we propose not merely to reaffirm our view, but also to offer some further reasons for believing it. These are found in demonstrable relations between speech perception and other auditory modules.

SPEECH AND THE MODULAR BASIS OF AUDITORY PERCEPTION

We assume that auditory perception relies on certain precognitive processes. One such process, "scene analysis," parses the acoustic input at the two ears, reconstructing an auditory scene in which several sound sources, at fixed or changing

locations, may be perceived, and segregating acoustic information into separate streams according to source (Bregman, 1978). Other processes attribute certain primitive properties—for example, pitch, timbre, loudness—to the stream of sound from each source. There is, as we shall argue, ample ground for distinguishing these precognitive processes from later, presumably cognitive, processes that associate patterns of pitch, timbre, and loudness at localized sources with particular events in the environment.

Each of the precognitive processes is presumably controlled by a distinct module, specialized to deal with acoustic information in a particular way and to provide the cognitive processes with a representation of some particular perceptual domain. We believe that speech perception (indeed, linguistic processing in general) is controlled by just such a module. Speech perception, therefore, is thoroughly precognitive and part of the larger specialization for language. This claim is controversial, for, in the more conventional view, speech perception is accomplished by auditory processes of the most general sort (e.g., Stevens, 1975; Miller, 1977; Kuhl, 1981) and in some versions requires a cognitive stage in which the outputs of the standard auditory modules are matched to phonetic prototypes or otherwise assigned to phonetic categories (e.g., Crowder and Morton, 1969; Fujisaki and Kawashima, 1970; Pisoni, 1973; Oden and Massaro, 1978). Thus, certain patterns of temporal variation in the pitch and timbre of certain sources are heard as phonetic events ([ba], [ga]) for exactly the same reasons that other such patterns are heard as nonphonetic events (the sawing of wood or the breaking of glass). Seen this way, speech perception depends on precognitive processes that some may regard as modular, but it does not have a module of its own.

We have taken the opposite view, arguing that speech perception satisfies the various criteria that Fodor (1983) proposes for modulehood, including, most importantly, "domain specificity." This is to suppose that speech perception has its own perceptual primitives. These primitives, we have said, are the intended articulatory gestures that constitute phonetic events (Liberman and Mattingly, 1985). Speech perception is thus independent of the modules for pitch and timbre. A simple example will illustrate what we mean. If a brief synthesized resonance of changing center frequency, like one of the two shown in Figure 1a, is presented in isolation, listeners hear, not unexpectedly, a nonspeech chirp. Its pitch depends on the fundamental frequency at which the resonance is excited and its changing timbre on the slope of the center frequency. But if such a resonance is the initial portion (the "transition") of the third-formant trajectory in a speechlike, three-formant pattern, like one of the two shown in Figure 1b, listeners hear a stop-vowel syllable, [da] or [ga]. Which of these syllables is heard depends, again, on the slope of the transition, the two patterns being otherwise identical. The pattern with the falling transition is heard as [da], the one with the rising transition, as [ga].[1] Thus the resonance can be perceived in two different ways, as nonspeech or as speech. Our view is that when the resonance is presented in isolation, the auditory module for timbre interprets its center-frequency slope as a chirp, but when this same resonance is in the appropriate context, the phonetic module interprets its slope, together with other parts of the pattern, as a phonetic event.

[1] In natural speech, there are many acoustic cues that distinguish [da] and [ga], but for our experimental purposes, all but the third-formant transitions have been either omitted or effectively neutralized.

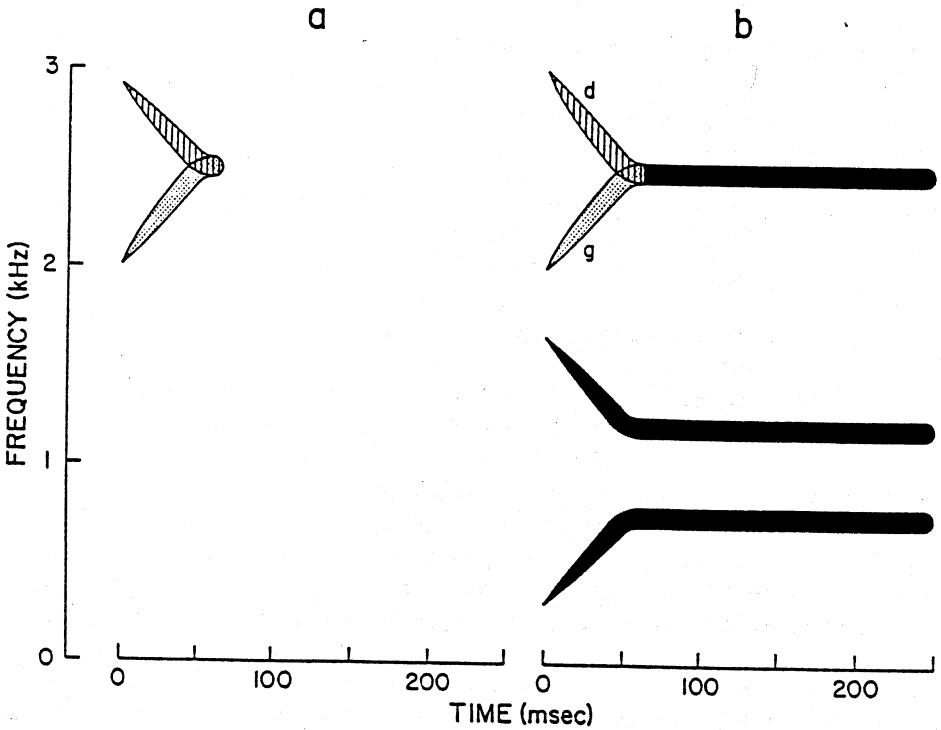

Figure 1. *Patterns that illustrate the "domain specificity" of speech perception. a:* Isolated resonances. *b:* The same resonances producing [d] and [g] in a speechlike context.

Our aim now is to develop this phenomenon further as evidence for a distinct phonetic module and also to consider more broadly how modules might be classified and what kind of architecture they might form.

EXPERIMENTAL EVIDENCE OF THE INDEPENDENCE OF THE PHONETIC MODULE

Our claim is that the different percepts evoked by the sloping resonances are produced by different modules, but other interpretations, more in accord with the conventional view, are, of course, conceivable. Thus it might be suggested that the chirp and the stop consonant are merely different cognitive interpretations of the same basic auditory representation. Or it might be suggested that though the chirp and the consonant are indeed distinct auditory representations, they are nevertheless of the same kind, both being representations of pitch and timbre. In that case, they differ either because the sloping resonance, like certain optical figures, happens to be an inherently ambiguous stimulus, permitting two alternative percepts, or because of some auditory interaction between the resonance and its context. The experiments we discuss below appear to rule out such explanations and to confirm the view that the two percepts are of genuinely different kinds, having been arrived at by different modular processes.

Consider first the possibility that the speech percept is only a cognitive reinterpretation of its nonspeech counterpart, as if people had simply learned that percepts common to speech and nonspeech are, in some contexts, to be given phonetic labels. If that were the correct interpretation, listeners would be expected to perceive the nonspeech chirps in the speech, but, in fact, they cannot. When presented with patterns like those shown in Figure 1 and asked to match the sound of the isolated transitions to the full syllables, listeners performed at chance rates (Repp et al., 1983). The same result was obtained in a later experiment (see below), in which the effective transition cue was not a resonance (as in the patterns in Figure 1), but a sinusoid that tracked the center frequency of the resonance (Whalen and Liberman, 1987). (In isolation this sinusoidal transition sounds like a whistle, not a chirp, and, though discordant in harmonic structure with respect to the remainder of the syllable, it nevertheless produces a normally perceived difference between [da] and [ga].)

Also relevant to the possibility of cognitive reinterpretation are experiments in which listeners are asked not to attach phonetic labels but simply to discriminate among the transitions—isolated in the one condition, incorporated into the syllables in the other—on the basis of any difference they can hear. If the speech percept were a cognitive reinterpretation, listeners should be able to access the underlying percepts that would, on that assumption, be common to speech and nonspeech. In that case, the discrimination functions in the two conditions should be the same. In fact, however, they are different, having very different shapes. The difference was shown first in an experiment in which the synthetic speech patterns were like those in Figure 1, except that the transitions were of the second formant and they were more numerous, being separated by small, equal-onset intervals and covering the full range of [ba], [da], and [ga] (Mattingly et al., 1971). When these transitions were presented in isolation, pairs that differed sufficiently in onset frequency were discriminated at above-chance rates, just as psychoacoustic considerations would predict. However, discrimination among the same transitions in the full syllables did not show these psychoacoustic effects at all, but rather reflected the phonetic categorization: Discrimination was relatively poor when the members of the stimulus pair belonged to the same phonetic category and relatively good when they belonged to different categories. A later study, using three-formant patterns like those in Figure 1, took advantage of a procedure that causes the formant transitions to be perceived as chirps and speech simultaneously (see below). With this procedure, numerous transitions of the third formant (separated by equal-onset intervals within the range shown in Figure 1) were discriminated in each of their simultaneously available speech and nonspeech manifestations. The resulting functions were grossly different in shape, just as they had been in the earlier experiment (Mann and Liberman, 1983).

We now turn to experiments that not only provide further evidence against the possibility that the chirp and the consonant are different cognitive interpretations of the same auditory percept but also refute the possibility that, though they may be different percepts, they are made of the same kind of primitives. These experiments exploit the phenomenon known as "duplex perception," first reported by Rand (1974), in which a part of the acoustic stimulus is apparently heard as speech and nonspeech simultaneously. In earlier papers (Liberman and Mattingly, 1985; Mattingly and Liberman, 1985, 1988), we dealt with only one form of duplex percep-

tion, treating it as if it were a freak phenomenon that just happened to provide controls appropriate to the conclusion that perceiving the resonances as speech and nonspeech reflects the outcome of different modules. Here we propose to view the phenomenon more broadly and so to show, perhaps more compellingly, that the earlier conclusion is justified. For the broader view makes it plain, we think, that phonetic perception does not depend on the output representations of the modules for pitch, loudness, and timbre, but is, rather, formed of its own distinctly phonetic primitives. It also allows us to see, as we had not before, that the phonetic module is equally independent of the output representations of the module for scene analysis.

To produce duplex perception in the most general case, two simultaneously presented parts of a stimulus are made in some way acoustically inconsistent with one another, so that two separate sound sources will be heard; but the information required for the perception of a particular speech sound is divided between these two parts (Bregman, 1987; Mattingly, 1987). The result is that one source is perceived as a nonspeech sound that, not surprisingly, depends on information in one of the two separated parts of the stimulus, but the other is perceived as the particular speech sound that depends on information in both parts. Thus, one of the two parts is being perceived as speech and nonspeech at the same time.

The different forms of duplexity can be conveniently exemplified by variations on the three-formant synthetic speech patterns in Figure 1b. Assume that each of these patterns is divided into two parts, the third formant and the combination of the two lower formants, as in Figure 2, and that these two parts are synthesized separately, but with identical fundamental frequency contours. If the third formant is presented in isolation, it is perceived as a buzz. The timbre of this buzz depends on the center-frequency trajectory of the formant; the pitch, on the fundamental frequency at which the formant is excited. Like the chirp described earlier, the buzz is categorically nonphonetic: Listeners cannot match the buzz having the falling transition to the three-formant [da] or the buzz having the rising transition to the three-formant [ga] (D. Whalen, personal communication). Clearly, then, listeners do not perceive the buzz as a constituent of the syllable. If the combination of the two lower formants is presented in isolation, it is perceived as a syllable, because these two formants are normally sufficient to engage phonetic perception; but the syllable is ambiguous between [da] and [ga], because the distinguishing cue in the third formant is missing.

If the two parts are electrically summed and presented diotically—that is, both parts to both ears—listeners hear, of course, exactly what they would have heard had the stimulus never been divided—[da] or [ga], and nothing else—and the lateralization of the single percept toward one side of the head or the other can be controlled by manipulating the relative intensity of the signals at the ears. But even if the two parts are presented dichotically—one part at one ear and one at the other—listeners still perceive only the single [da] or [ga] (cf. Broadbent, 1955). Moreover, as we, together with Lawrence Rosenblum and Carol Fowler, have recently found, manipulating the relative intensity of each of the dichotically presented parts affects the location of the percept but does not produce two percepts. Though the signals at the two ears have completely different spectra, scene analysis is not led to treat them as separate sources.

But now suppose that the two parts of the stimulus pattern in Figure 2 are synthesized with considerably different fundamental frequency contours (cf. Dar-

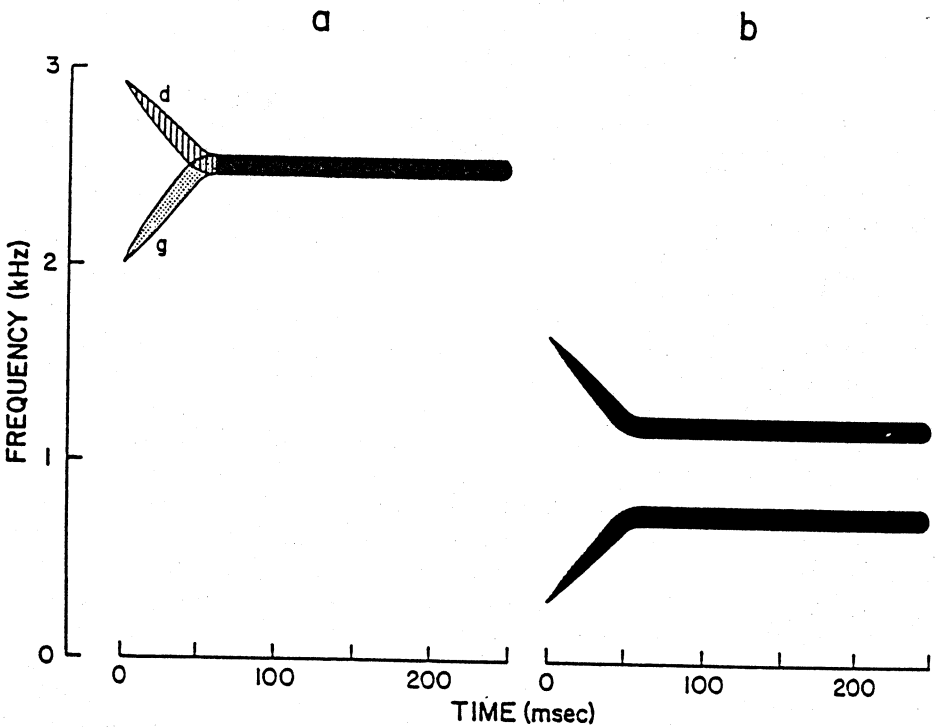

Figure 2. *One way of partitioning the stimulus pattern so as to produce duplex perception. a:* The third format with its variable [d]-producing and [g]-producing resonances. *b:* The fixed remainder of the pattern.

win, 1981). The phonetic information in each part remains essentially the same, but the two parts have inconsistent harmonic structures. In a diotic presentation of these discordant stimuli, perception is duplex. Because the two parts have different harmonic structures, scene analysis treats them as separate sources. Thus, two sounds are heard, with two different pitch contours, though at the same lateralization. One is the same categorically nonphonetic buzz that is heard when the third formant is presented in isolation. This is just what would be expected, assuming that scene analysis treats the third formant as a separate source because of its distinctive harmonic structure and that the modules for pitch and timbre respect this analysis. What is remarkable is the nature of the other percept, for it is not at all what would have been expected had phonetic perception accepted the division of information by source provided by scene analysis. If that had occurred, the other percept would have been the same ambiguous syllable heard when the two lower formants are presented in isolation. But this ambiguous syllable is not heard. The other percept is, instead, the unambiguous [da] or the unambiguous [ga], depending on the third formant, just as when all three formants are synthesized with the same fundamental. Evidently the third formant is being used simultaneously to form two percepts, one phonetic, the other not; and the phonetic percept could have been formed only by combining information from two parts of the stimulus, even though these two parts are treated as separate sources.

If the two discordant parts of the stimulus in Figure 2 are dichotically presented (cf. Cutting, 1976; Darwin, 1981), the results are very similar to those in the diotic

presentation just described except for the lateralization of the two percepts. The nonspeech buzz is now heard at the ear receiving the third formant, and the unambiguous [da] or [ga] at the ear receiving the other two formants. What the dichotic presentation makes clear is that the two parts of the stimulus are in fact separately lateralized in both presentations; they are heard at the same place in the head in the diotic case only because the relative interaural intensities of the two parts are not independently varied. In either presentation, it is the difference in harmonic structure between the two parts, and not their spectral differences, that leads the scene analysis module to define two sources. And in either presentation, one of the two sources is a nonspeech percept evoked by one of the two parts, while the other is a speech percept that can only have been evoked by both.

Duplex perception of another kind can be produced by a different partitioning of the two basic syllable patterns (Figure 3). Now one part of the stimulus is just the 50-msec transition of the third formant. In isolation, as we have already seen, this transition is heard as a chirp, its timbre depending on the slope of the transition. The other part is the remainder of the pattern—the rest of the third formant and all of the first and second formants. In isolation, this part is perceived as a syllable that is once again ambiguous between [da] and [ga]. To produce duplex perception in a diotic presentation, the intensity of the initial third-formant transition is made sufficiently greater than that of the rest of the third formant. In this case, scene analysis defines two sources, because the abrupt change in intensity between the transition and the rest of the third formant is not paralleled by simultaneous changes in the

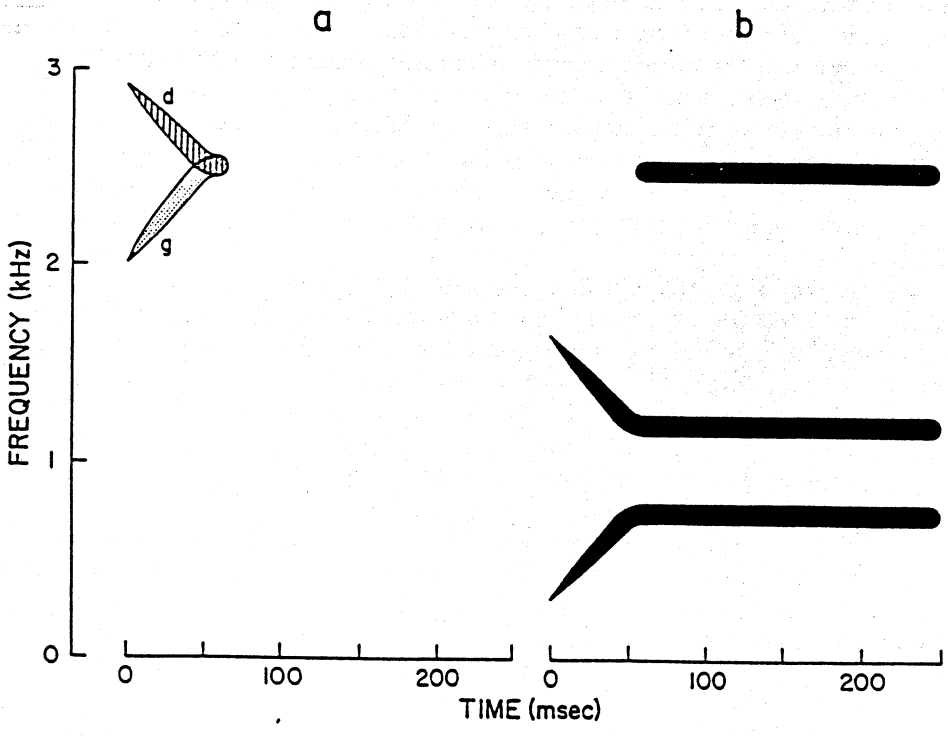

Figure 3. *A second way of partitioning the stimulus pattern so as to produce duplex perception. a:* The variable [d]-producing and [g]-producing resonances. *b:* The fixed remainder of the pattern.

intensities of the other two formants. The listener now hears the nonspeech chirp and the unambiguous [ga] or [da], with a common lateralization (D. H. Whalen, personal communication; cf. Whalen and Liberman, 1987). Again, acoustic information assigned by scene analysis to different sources is used to form one phonetic percept.

If these same two parts of the pattern—the third-formant transition and the remainder—are presented dichotically, duplex perception results without any adjustment of the intensity of the third-formant transition. In this case, the scene analysis module defines two sources because the relative interaural intensity does not change consistently: The third formant shifts abruptly from one ear to the other after only 50 msec, while the other two formants remain at the same ear throughout the stimulation (Bregman, 1987; Mattingly, 1987). As in the first kind of dichotic duplexity, the nonspeech chirp is heard at the ear receiving the transition, the unambiguous [da] or [ga] at the ear receiving the other part of the stimulus (Mann and Liberman, 1983; cf. Rand, 1974; Liberman, 1979; Liberman et al., 1981; Repp et al., 1983; Repp and Bentin, 1984).

To show what we meant by saying at the outset that phonetic perception has its own primitives and is, in that important sense, a module in its own right, we offer as an example the fact that a formant transition is perceived in isolation as a chirp but in appropriate acoustic context as [da] or [ga]. Alternative interpretations are ruled out, we think, by the observations reported in this section. Thus they demonstrate that the speech percept and the chirp depend on different representations, not merely on different cognitive interpretations of the same representations. As we pointed out earlier, this is implied by the inability of listeners to match the one to the other and by the different patterns of discriminability, but it is shown even more forcefully by the fact of duplex perception. For if the speech percept were a cognitive reinterpretation, we should expect in the duplex experiments that perception would be triplex, not duplex. That is, listeners would hear the nonspeech sound and the unambiguous syllable, just as they do; but they would also hear the ambiguous syllable, as if the unambiguous syllable were simply the sum of the other two percepts. In fact, listeners perceive only the nonspeech sound and the unambiguous syllable.

Another conceivable interpretation was that while there are indeed two different auditory representations, they are of the same kind, both being representations of pitch and timbre. The two representations might arise either because the sloping resonance was an inherently ambiguous stimulus or because of some auditory interaction between the resonance and the context. But this interpretation does not explain the great difference in discrimination functions, which implies not only different representations but representations of different kinds, and it is made to seem even more implausible by duplex perception.

As for the possibility that the nonspeech and speech percepts are merely alternative representations of an ambiguous stimulus, we note that there is no alternation between them, such as an ambiguous stimulus would evoke; rather, they are experienced simultaneously. Moreover, duplex perception is mandatory and convincing; listeners cannot choose to hear one percept or the other, and they are in no doubt about what they do hear.

Nor can the difference between the two percepts depend on some auditory interaction between the two parts of the stimulus, for in all forms of duplex percep-

tion the acoustic context of the duplexly perceived resonance is held constant. In this connection, the diotic forms of duplex perception are especially telling, because the duplexly perceived part of the stimulus is presented together with the remaining part at both ears and is therefore in the same context at the very earliest stage of the perceptual process.

We are left, then, with the conclusion that the two representations are of genuinely different kinds, having different primitives. Speech perception must therefore be a separate module, independent of the modules for pitch and timbre, that provides to cognition its own peculiarly phonetic representations.

But duplex perception permits another conclusion that we had not appreciated before: Speech perception is also independent of the module for scene analysis. For in all forms of duplex perception, the two parts of the stimulus are treated by scene analysis as coming from two different sources; yet, as we have pointed out, perception of the unambiguous syllable depends critically on information in both perceived sources, and the ambiguous syllable is not also heard. This is not what would be expected if the pitch and timbre representations of the two different sources, derived from the segregated streams of acoustic information that scene analysis provides, were somehow being cognitively combined or reinterpreted. If that were the case, perception should, once again, be triplex—that is, listeners should be able to hear the ambiguous syllable as well as the nonspeech sound and the unambiguous syllable. Thus, we must conclude that speech perception makes no use of the representations of the qualities of particular sources. Rather, it appears simply to use all the phonetically relevant information in the signal and to have its own specifically phonetic criteria—very different from those of scene analysis—for what is to count as one event and what as more than one. Thus we have yet another reason to consider speech perception to be a separate module. Quite apart from the differences in primitives that show the speech perception module to be independent of those for pitch and timbre, its failure to respect differences among sound sources shows it to be independent of the scene analysis module.

DIFFERENT KINDS OF PERCEPTUAL PRIMITIVES: OPEN AND CLOSED MODULES

Each of the several auditory modules has its own domain, its own sources of information, and its own primitives. In that sense, every module is different from all others. Nevertheless, modules can be grouped into classes on the basis of common properties. Thus the several auditory modules (and no doubt the modules of some other modalities as well) respond to the dimensions of the signal in two characteristically different ways, producing percepts we have previously called homomorphic and heteromorphic (Mattingly and Liberman, 1988). For convenience we refer to the modules that produce homomorphic percepts as "open" and to those that produce heteromorphic percepts as "closed."

Open modules are adapted to the perception of a wide variety of environmental events, including a vast array that evolution could not have anticipated: passing cars, breaking glass, and rattling chains. Accordingly, each of these modules represents to cognition an auditory quality—pitch, loudness, or timbre—corresponding more or less directly to a simple dimension of the acoustic signal—frequency, inten-

sity, or spectral shape. It is left to a further, cognitive stage to associate these qualities with some particular class of events.

Closed modules comprise a variety of specializations, including, for example, scene analysis, echo ranging in bats (Suga, 1984), and phonetic perception. What these closed modules have in common, and what distinguishes them from the open modules, is that the percepts are heteromorphic: The dimensions of the percept do not correspond directly to the dimensions of the signal; the signal dimensions are merely the data from which the very different, indeed incommensurate, dimensions of the percept are derived. The closed scene-analysis module, for instance, responds to the narrow range of interaural time disparities that is ecologically appropriate for sounding objects at different positions of azimuth. What is perceived, however, is the location of the source, not temporal disparity as such. The bat's echo-ranging module measures the delay between the originally emitted cry and its reflection, but what is perceived is presumably the distance of the reflecting object, not an echoing bat cry. We claim that the percepts produced by the phonetic module are similarly heteromorphic. This module tracks the changing center frequencies of formants, but what is perceived is a sequence of phonetic events, not changing timbre or a medley of changing pitches. Of course, such homomorphic qualities as pitch and timbre are also associated with the phonetic event and somehow specify the voice quality and identity of the speaker, but they are not the primitives out of which the phonetic percepts are formed.[2]

It is possible that there are closed modules for the perception of other special classes of ecological events as well as speech sounds. Perception of rhythm or of the quality of the impact of one body on another might well be modular. Thus, while there can be no biological specialization for rattling chains as such, the cognitive identification of rattling chains might depend on closed as well as open modules. Unfortunately, very little is known about these matters; the perception of nonspeech sounds is not usually studied from an ecological point of view (see, however, Warren and Verbrugge, 1984).

The distinction we have made between open and closed modules is implied in certain neurobiological accounts of the source localization system (Knudsen and Konishi, 1978; Yin and Kuwada, 1984). These stress that in contrast to many other perceptual processes (those we call "open"), source localization cannot rest on a direct projection from the sensory epithelium to some perceptual map in the brain. Thus, Konishi (1986) says that it requires a "central synthesis" and proceeds to show what this means neurobiologically.

THE EFFECTS OF EXPERIENCE ON OPEN AND CLOSED MODULES

All modules, both open and closed, are presumably susceptible to change under environmental influences, but there is, in our view, an important difference between them in how these influences are constrained. This reflects a difference in the kinds of tasks for which the two kinds of modules are specialized and also a differ-

[2] Thus, in much the same way (and for much the same reason) that the neurobiological module for stereopsis is color-blind (see, e.g., Livingstone and Hubel, 1987), we suppose that the phonetic module is tone-deaf.

ence in the level, precognitive or cognitive, at which environmentally induced changes occur. The processes we earlier characterized as cognitive deal with the outputs of the modules, not with their internal workings. It is accordingly a cognitive process by which the great variety of ecologically arbitrary acoustic events acquire meaning. Breaking glass, for example, becomes connected in experience with the particular combination of homomorphic auditory primitives it evokes. But the open modules that yield these primitives are not themselves altered, certainly not in ways that would make them respond heteromorphically, as some closed modules do, to the properties of the event as such. Indeed, they must not be so altered, lest they become so well adapted to perceiving breaking glass as to become that much less well adapted to perceiving other acoustic events—say, rattling chains—that must depend on different combinations of the same homomorphic primitives.

Much the same cognitive account would have to be offered for speech perception, especially for the way it adjusts to different languages, if we were to accept the conventional assumption that speech and breaking glass depend on the same modules and share the same primitives. For in this case, learning the phonetic categories of a language would require that the outputs of the open modules be connected to the appropriate phonetic names or prototypes; the homomorphic primitives themselves would necessarily remain unchanged. Therefore, the learned phonetic response would forever require a translation, however automatic, from the perceptual primitives of the open modules to the cognitively established phonetic categories of some particular language. For surely there is no way a listener could ever come by experience to replace homomorphic chirps, for example, with heteromorphic consonants, any more than an echo-ranging bat could, in any reasonable account, ever come by experience to replace the homomorphic echoes of other bats' cries with the heteromorphic distances the echoes of its own cries presumably evoke. Either of these replacements would require a kind of perceptual alchemy, a transmutation of one category of perceptual primitives into another.

But if speech has its own closed module and, accordingly, its own heteromorphic primitives, the effects of experience can be precognitive. That is, experience can act on the modular mechanisms themselves, hence on the perceptual representations they immediately produce; perception of the indefinitely many sounds that excite the open modules is in no way affected. It is, then, the module itself that can "learn" and so come to deploy its phonetic resources in accordance with the particular requirements of the languages to which it is exposed. Such modification would qualify as epigenetic; it would be no different in principle from the way the closed module for scene analysis, for example, presumably adjusts to changes in interaural time disparities as the head grows bigger. In this kind of development, it can hardly be that the animal learns to "translate" old perceived disparities into new locations, or new perceived disparities into old locations, if only because the scene analysis module never did perceive disparities homomorphically as disparities. It must rather be that this module accommodates its processes, and hence its heteromorphic perceptual output (location of a source), to the changing environmental circumstances. We should suppose that, given proper conditions, the phonetic module accommodates itself in much the same epigenetic way.

Similar considerations apply to the development of phonetic perception in evolution. For the conventional assumption that speech and nonspeech sounds share a

common set of processes and primitives entails a constraint on evolutionary adaptation identical to the one that applies in ontogenesis: Changes in the open modules that might be appropriate for speech sounds would be inappropriate for most others. It is therefore easy to see why, among the open auditory modules and their primitives, there are apparently few, if any, adaptations that are specific to the human species or to the speech this species commands. But if, as we believe, there is a closed phonetic module, the processes of phonetic perception were free to go where evolution happened to take them.

There remains a constraint that applies only to modules that serve for communication: What counts as structure in production must count as structure in perception, or else communication does not occur. How then was this parity established and maintained as speech developed in the race and indeed as it develops once again in the child? Here the prevailing view that speech and nonspeech are managed by a common set of perceptual modules leads to an awkward conclusion. For if the percepts produced by phonetically significant articulations are no different in kind from those produced by breaking glass, then parity between sender and receiver can have been established only by convention, in which case the evolution of the system must have been cultural rather than biological. Phonetic communication would then have been governed by rules like those of such arbitrary codes as Morse or semaphore. To use the code, one would have acquired the rules, either by instruction from an adept or, as is so often assumed about natural language, by associative learning or inductive reasoning. But if phonetic perception rests on a module that is narrowly adapted to phonetic structures, the relation to production can be truly organic. One can suppose, then, that the system is so adapted as to cause the listener and speaker to be immediately in the same (phonetic) domain; there is no need for a convention to establish the relation between phonetic element and nonphonetic percept, and hence no need for a cognitive translation across this relation. Given this resolution of the parity problem, it is easier to see speech in its proper light as a product of biological evolution. We have developed the basis of this notion elsewhere (Mattingly and Liberman, 1988), but it bears so directly on the nature of the phonetic specialization and its primitives that we summarize it here.

All the utterances of a language are made up of only two or three dozen phonetic segments—that is, consonants and vowels. These few elements can form a vast number of words (and a vaster number of sentences) because they are ordered in relatively long strings. But producing these strings at rates high enough to make the linguistic enterprise feasible is not accomplished by ordinary means; rather, it rests on a specialization that, as much as any other, characterizes our species. This specialization, called coarticulation, speeds production by overlapping gestures for successive segments that are executed by different articulatory organs and by merging gestures that are executed by the same organ. This required that the gestures used in speech be a distinct set, different for the most part from those people make when they lick their lips, chew, move food around in the mouth, and so on. It also required that the gestures be coordinated in both space and time, so that the underlying phonetic structure would be properly maintained. In that connection, we suppose that coarticulation evolved primarily under opposing pressures: to overlap and merge so as to gain as much speed as possible, but not to allow this strategy to destroy all information about the order of segments.

Coarticulation also speeds the perception process, because it folds information about successive phonetic segments into a single stretch of sound and so evades the limitation on rate that would otherwise be set by the temporal resolving power of the ear. But another, equally significant, consequence of coarticulation is that the shape of the vocal tract, and hence the nature of the sound, is determined simultaneously by the identity of the several segments that are being coarticulated. There is, accordingly, a complex relation between acoustic signal and the appropriate percept, a relation that is peculiar to speech. But this matters little to cognition so long as the phonetic segments are defined as gestures (or, more properly, as intended gestures) and so long as there is, at the precognitive stage, a process specialized to recover these gestures from the acoustic signal. The phonetic module is just such a specialization, as we see it, and the gestures that define the phonetic elements are the primitives it presents to cognition. What evolved, then, is a single specialization that comprises complementary processes, one for executing and coordinating phonetic gestures and one for dealing with the acoustic consequences.

Thus parity between sender and receiver is built into the very nature of the system and is automatically maintained as the system develops, whether in the race or in the individual. Accordingly, production and perception of phonetic structures are accomplished in the precognitive domain. To speak a word or to perceive it, one need not know how it is spelled—that is, what sequence of consonants and vowels it comprises—or how the gestures of a speaker are matched to the percepts of a listener; both these requirements are met in the precognitive work of the phonetic specialization.[3]

THE ARCHITECTURE OF AUDITORY MODULES

The various perceptual modules in Fodor's (1983) account simply operate in parallel, each taking the information it needs from the input signal and presenting cognition with the results of its analysis. However, we have already assumed one departure from this simple scheme. Since auditory qualities are attributed separately to each of the sound sources in the auditory scene, scene analysis must segregate the acoustic information into separate streams according to source (Bregman, 1978). This is as much as to say that with respect to the flow of acoustic information, the closed scene analysis module is in series with and precedes each of the open modules. On the other hand, there is no reason that the various open modules should not be in parallel with one another. They represent different properties of acoustic sources to cognition, so, for a particular source, they naturally use the same piece of acoustic information without conflict.

[3] Spelling became an occasionally cognitive matter only after some of our ancestors discovered that words are, in fact, compounded of a limited stock of consonants and vowels. That discovery, which is not often made spontaneously by preliterate children or illiterate adults, was, of course, the precondition for the invention of alphabetic transcription (Liberman et al., 1974; Morais et al., 1979, 1984; Mattingly, 1985). It is in precisely this sense that reading and writing are properly regarded as a secondary, cognitive elaboration of a primary, precognitive faculty.

There is, however, a serious question that arises more generally about the relation between closed and open modules. For, as we have pointed out, closed and open modules are concerned with different kinds of events. Yet, as is evident from examples we offered earlier, a particular closed module and a particular open module may use the same acoustic information, though in very different ways. Therefore, if the open and the closed module were simply in parallel, we should expect duplex perception to be the rule under normal ecological conditions. Whenever a piece of acoustic information was available that both the closed module and the open module could use, both modules would respond, producing both a homomorphic and a heteromorphic percept. The question is how such duplexity is avoided.

Consider, for example, the case of the echo-ranging bat. Investigators have been at pains to explain how it is that out of all possible sounds to which the bat is sensitive, including especially the cries of other bats, the echo-ranging module manages to respond only to the echoes of the bat's own cries. One explanation, by Suga (1984), takes account of the fact that the first harmonic of the bat's cry does not radiate and is therefore available (by bone conduction) only to the bat that emitted the cry. Suga proposes, then, that the echo-ranging module responds only when the returning echo can somehow be appropriately matched to this harmonic. Apart from the fact that this proposal finds empirical support, it is most appealing because the selection of stimuli to be processed is an automatic consequence of the very processes that the echo-ranging module is specialized for. But this can explain only why stimuli that are inappropriate for echo ranging do not engage the echo-ranging module. It does not explain, in response to the question we raised, why stimuli that are appropriate for echo ranging do not also engage the open modules, causing the bat to perceive not only the (heteromorphic) distance of its prey, but also the (homomorphic) cry from which the distance percept was derived. Because these open modules respond to an indefinitely numerous variety of sounds, including in particular the echo-ranging cries of other bats, they could hardly be expected to discriminate selectively against the echoes of the bat's own cries. Given a parallel arrangement of the modules, such selectivity would require that the open modules be equipped with gates through which the inappropriate information could not pass or inhibitors that would nullify whatever response such information might evoke. But these gates and inhibitors would have to identify just the information that is relevant to echo ranging, and to do that they would need exactly those special abilities with which the echo-ranging module itself is endowed.

The same question arises with even greater force in the case of the phonetic module. There we have seen that the representations of the phonetic module are independent of those of the various open modules. But we have also seen that the phonetic module and the open modules can use not just similar acoustic information, as in the case of echo ranging, but exactly the same information. Moreover, it is obvious that in ordinary ecological conditions the phonetic module and the open modules regularly operate at the same time. A person listening to speech hears not only the consonants and vowels represented by the closed phonetic module, but also nonphonetic properties of the speech signal, represented by the open modules, that enable him to judge the speaker's sex, voice quality, and vocal effort. But there is no actual duplexity: The listener does not hear a medley of chirps and buzzes. Any gating or inhibiting mechanism that might be supposed to account for this

would need to have the same special properties that must be attributed to the phonetic module itself.

What is required to prevent duplex perception, both for echo ranging and for speech perception, is not a separate gating mechanism but an architectural arrangement that allows a closed module to preempt the acoustic information relevant to events that concern it, preventing this information from reaching the open modules at all (Mattingly and Liberman, 1985). Indeed, just this kind of arrangement can be observed in the relation between the closed scene analysis module and the various open modules, for, as we have observed, scene analysis must precede the various open modules in series. However, scene analysis does not simply pass on all the acoustic information; it preempts some of it in the very process of defining sound sources. Consider, in this connection, one of the cues that scene analysis uses for this purpose: disparity in time of arrival of similar signals at the two ears. A sufficiently great disparity is taken to mean that the two signals, despite their similarities, must correspond to two separate sources. The listener duly perceives these two sources, one at each ear, as temporally disparate. A small temporal disparity, however, is used as evidence of the azimuth position of one source. In this case, the listener hears this one source at a particular location but not the temporal disparity of the input signals as such. Perception of this disparity as disparity would obviously be of no value to the listener and would serve only to confuse him: The listener wants to know about genuine temporal disparities between objectively different events in the acoustic environment and about the location of these events, but not about disparities in the time of arrival at his two ears of similar signals resulting from the same event. Evidently, information about these latter disparities is preempted by the scene analysis module.

THE PREEMPTIVENESS OF THE PHONETIC MODULE

We turn now to some experimental evidence for the notion we have just advanced: that the phonetic module does, indeed, come first, preempting only those aspects of the signal that are phonetically relevant. The experiment makes use of one of the forms of duplex perception we described earlier, in which the 50-msec transitions of the third formant (the cue that distinguishes [da] and [ga]) and the rest of the synthetic pattern (Figure 3) are presented diotically (Whalen and Liberman, 1987). But now, instead of simply observing that these transitions are perceived duplexly when their intensities are sufficiently high, we follow the course of the percepts as, beginning near zero, the intensity is increased over a wide range. Another difference between this experiment and the one described earlier is that the 50-msec transitions are now not proper formants moving through a harmonic series, but sinusoids with the same frequency trajectories as the formants. In isolation, these sinusoidal transitions sound like whistles rather than chirps, and are thus even more nonspeechlike. There is, then, in these patterns a discordance at the surface of the signal, since the sinusoid is unrelated to the harmonics in the remainder of the pattern. It is, incidentally, of some interest that once again a considerable surface discordance is ignored in phonetic perception: Listeners perceive the same [da] or [ga] they would have perceived had the transition cues been not sinusoids, but

proper resonances like those that form the remainder of the pattern. We suppose, as do Whalen and Liberman, that the basis for perception of coherent [da]s and [ga]s is the underlying phonetic structure the phonetic module is specialized to reveal. In any case, the sinusoidal transition in place of the resonance is not critical to those aspects of the results that bear on preemptiveness; these results are also found when proper resonances are used (D. H. Whalen, personal communication).

As the intensity of the sinusoidal transition is gradually increased from near zero, a point is reached at which it becomes phonetically effective, causing the listeners to perceive [da] or [ga] appropriately. At this point, perception is simplex: Listeners do not also perceive the nonspeech whistles the sinusoidal transitions produce in isolation. With further increases in the intensity of the transition, perception remains simplex. (Within this range, listeners are at chance levels when they try to associate the whistles produced by isolated transitions with the [da]s and [ga]s these same transitions produce when integrated into the phonetic percept.) As the upper limit of this range is reached, perception becomes duplex: The transitions continue to serve their phonetic function—that is, they continue to produce [da] and [ga] appropriately—but now they begin also to produce nonspeech whistles. (Within this range, listeners can match the whistles that are part of the duplex percept with those that are perceived when the sinusoidal transitions are presented in isolation.) At first these whistles are barely audible, but they become progressively louder as the intensity of the transitions is further increased. Over this range, the [da]s and [ga]s that are also produced by the transitions remain clearly perceptible and apparently constant in loudness.

These findings imply that the transition is used first to provide information for the phonetic module, which produces the heteromorphic syllables [da] or [ga]. Then, when the phonetic module has had its fill, as it were, it passes the remainder on to the open modules; these convert the remainder into homomorphic whistles. Thus it is that phonetic processes have first claim on the relevant stimulus information, using it for their own phonetic purposes and effectively preventing it from engaging the open modules at all. Because the phonetic module preempts the information in this way, listeners do not normally perceive speech duplexly.

Since the phonetic module exploits the entire signal without regard to decisions about number of sources or their location, preemption of the input to the open modules should occur not only in diotic, but also in dichotic, duplexity. There is some evidence that it does: Even when the third-formant transition has an intensity too low to allow subjects to detect a chirp at the ear to which it was presented, it is nevertheless sufficient to determine whether [da] or [ga] is heard at the ear receiving the remainder of the syllable (Bentin and Mann, 1983; S. Bentin and B. Repp, personal communication).

A REMAINING PROBLEM

According to the account of auditory architecture we have so far given, scene analysis represents an array of sources to cognition and segregates the acoustic information according to source. These segregated streams are then available to the open modules, which attribute pitch, timbre, and loudness to each source. The phonetic module, independently of the scene analysis module, uses all the available

and relevant information to form phonetic percepts; information not thereby preempted also becomes available to the open modules.

Thus the scene analysis module precedes the open modules in series; so too does the phonetic module. But how do these two closed modules relate to each other? Are they in parallel or in series? A similar question arises for the bat. For if, as we have assumed, the bat hears the echo-ranging cries of other bats together with other ambient sounds, it must hear them as separate sources and must therefore have its own scene analysis module preceding the open modules. How then is the scene analysis module related to the echo-ranging module? In both cases, a parallel arrangement between the closed modules must be ruled out, for it would simply defeat their preempting functions. The bat's open modules would receive information about the cries of other bats, unsegregated by source, through the echo-ranging module, and information about its own cries and their echoes through the scene-analysis module. Thus, in addition to the ecologically appropriate percepts, it would hear its own echo-ranging signals and a noise that was the sum of all the echo-ranging signals of its conspecifics. Similarly, the human listener's open modules would receive unsegregated information about voice quality and ambient sounds from the phonetic module, together with information about formant transitions through the scene-analysis module, so that the listener would hear not only the appropriate percepts but also an extra voice, degraded by ambient signals, and an assortment of chirps and buzzes. Clearly, the two closed modules must be in series, in both bats and humans, if the open modules are not to receive two versions of the acoustic input, neither of which is ecologically appropriate.

How, then, are the closed modules ordered with respect to each other? In the case of the bat, we suggest that echo ranging comes before scene analysis, for the emitted sonar signal and its echo are already defined sources and the position in azimuth of the reflecting object is determined by the orientation of the bat's body that maximizes the loudness of the echo. Echo ranging has no need to know about the general auditory scene; on the other hand, scene analysis has no use for signals that originate not in the bat's environment, but in the bat itself. Perhaps further neurophysiological investigation will provide a definitive answer.

In the case of humans, we conjecture that in similar fashion the closed phonetic module comes before scene analysis. We have seen in the phenomenon of duplex perception that the phonetic module ignores the representations by the open modules of the qualities of the particular sources that scene analysis has defined. We have also seen in duplex perception that phonetic perception makes no use of, and need have no access to, the segregated streams of information provided by scene analysis. So, if the phonetic module came after scene analysis, it would, very unparsimoniously, be reintegrating and preempting phonetically relevant signal information that had just been separated by source, yet still be obliged to pass along segregated streams of phonetically irrelevant information to the open modules. On the other hand, if scene analysis comes after the phonetic module, no similar difficulties arise: Scene analysis simply segregates the acoustic information that has not been preempted by the phonetic module, and the open modules operate on the resulting streams.

This account makes the empirical prediction that scene analysis will have different input information when the phonetic module is active than when it is not. It may be possible to confirm this prediction experimentally.

The notions we have advanced here are relevant to questions that are not commonly asked. Still, many students of auditory perception, accepting that the auditory system is a collection of components of the sort we call modules, might grant the validity of our question about the architecture that causes listeners to assign sources to locations and pitches to sources. Not so for our question about where in this architecture a phonetic module fits in. This question is pointless if there is no such thing as a phonetic module, and few believe that there is. Counting ourselves among the believers, we have here asked the question and found evidence that is consistent with an answer. That there is such an empirically plausible, if unconventional, answer testifies, we think, to the validity of the question and supports the assumptions about a phonetic module out of which it arises. It also suggests that it will be rewarding to inquire more broadly into the modular architecture of perceptual systems.

NOTE

This is an expanded and revised version of a paper by Liberman and Mattingly that appeared in Science (1989, Vol. 243, pp. 489–494).

ACKNOWLEDGMENTS

The writing of this chapter was supported by National Institutes of Health Grant NICHD-HD-01994 to Haskins Laboratories. We are grateful to Peter Eimas, Carol Fowler, Patrick Haggard, Bruno Repp, Lawrence Rosenblum, Michael Studdert-Kennedy, and Douglas Whalen for helpful comments on earlier drafts.

REFERENCES

Bentin, S., and V. A. Mann (1983) Selective effects of masking of speech and nonspeech in the duplex perception paradigm. *Haskins Lab. Status Rep. Speech Res.* 76:65–85.

Bregman, A. S. (1978) The formation of auditory streams. In *Attention and Performance*, Vol. 7, J. Requin, ed., pp. 63–75, Erlbaum, Hillsdale, New Jersey.

Bregman, A. S. (1987) The meaning of duplex perception: Sounds as transparent objects. In *The Psychophysics of Speech Perception*, M. E. H. Schouten, ed., pp. 95–111, Nijhoff, Dordrecht, The Netherlands.

Broadbent, D. E. (1955) A note on binaural fusion. *Q. J. Exp. Psychol.* 7:46–47.

Crowder, R. G., and J. Morton (1969) Pre-categorical acoustic storage (PAS). *Percept. Psychophys.* 5:365–373.

Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychol. Rev.* 83:114–140.

Darwin, C. J. (1981) Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Q. J. Exp. Psychol.* 33A:185–207.

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

Fujisaki, H., and T. Kawashima (1970) Some experiments on speech perception and a model for the perceptual mechanism. *Ann. Rep. Engin. Res. Inst. (Tokyo)* 29:207–214.

Knudsen, E. I., and M. Konishi (1978) A neural map of auditory space in the owl. *Science* 200:795–797.

Konishi, M. (1986) Centrally synthesized maps of sensing space. *Trends Neurosci.* 9:163–168.

Kuhl, P. K. (1981) Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *J. Acoust. Soc. Am.* 70:340–349.

Liberman, A. M. (1979) Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In *Proceedings of the Ninth International Congress of Phonetic Sciences*, Vol. 2, E. Fischer-Jorgensen, J. Rischel, and N. Thorsen, eds., pp. 468–473, Univ. Copenhagen Press, Copenhagen.

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

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967) Perception of the speech code. *Psychol. Rev.* 74:431–461.

Liberman, A. M., I. G. Mattingly, and M. Turvey (1972) Language codes and memory codes. In *Coding Processes in Human Memory*, A. W. Melton and E. Martin, eds., pp. 307–334, Winston, Washington, D.C.

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

Liberman, I. Y., D. Shankweiler, F. W. Fischer, and B. Carter (1974) Explicit syllable and phoneme segmentation in the young child. *J. Exp. Child Psychol.* 18:201–212.

Livingstone, M., and D. Hubel (1987) Psychophysical evidence for separate channels for the perception of form, color, movement and depth. *J. Neurosci.* 7:3416–3468.

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

Mattingly, I. G. (1985) Did orthographies evolve? *Remed. Spec. Ed.* 6:18–23.

Mattingly, I. G. (1987) Dichotic duplex perception and modularity. *J. Acoust. Soc. Am.* 82(S1):S120.

Mattingly, I. G., and A. M. Liberman (1969) The speech code and the physiology of language. In *Information Processing in the Nervous System*, K. N. Leibovic, ed., pp. 97–117, Springer-Verlag, New York.

Mattingly, I. G., and A. M. Liberman (1985) Verticality unparalleled. *Behav. Brain Sci.* 8:24–26.

Mattingly, I. G., and A. M. Liberman (1988) Specialized perceiving systems for speech and other biologically significant sounds. In *Auditory Function: Neurobiological Bases of Hearing*, G. M. Edelman, W. E. Gall, and W. M. Cowan, eds., pp. 775–793, Wiley, New York.

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

Miller, J. D. (1977) Perception of speech sounds in animals: Evidence for speech processing by mammalian auditory mechanisms. In *Recognition of Complex Acoustic Signals*, T. H. Bullock, ed., pp. 49–58, Dahlem Konferenzen, Berlin.

Morais, J., L. Cary, J. Alegria, and P. Bertelson (1979) Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* 7:323–331.

Morais, J., M. Cluytens, and J. Alegria (1984) Segmentation abilities of dyslexics and normal readers. *Percept. Mot. Skills* 58:221–222.

Oden, G. C., and D. W. Massaro (1978) Integration of featural information in speech perception. *Psychol. Rev.* 85:172–191.

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

Rand, T. C. (1974) Dichotic release from masking for speech. *J. Acoust. Soc. Am.* 55:678–680.

Repp, B., and S. Bentin (1984) Parameters of spectral/temporal fusion in speech perception. *Percept. Psychophys.* 36:523–530.

Repp, B., C. Milburn, and J. Ashkenas (1983) Duplex perception: Confirmation of fusion. *Percept. Psychophys.* 33:333–337.

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

Suga, N. (1984) The extent to which bisonar information is represented in the bat auditory cortex. In *Dynamic Aspects of Neocortical Function*, G. M. Edelman, W. E. Gall, and W. M. Cowan, eds., pp. 315–373, Wiley, New York.

Warren, W., and R. Verbrugge (1984) Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *J. Exp. Psychol. (Hum. Percept.)* 10:704–712.

Whalen, D. H., and A. M. Liberman (1987) Speech perception takes precedence over nonspeech. *Science* 237:169–171.

Yin, T. C. T., and S. Kuwada (1984) Neuronal mechanisms and binaural interaction. In *Dynamic Aspects of Neocortical Function*, G. M. Edelman, W. E. Gall, and W. M. Cowan, eds., pp. 263–313, Wiley, New York.