27 January 1989, Volume 243, pp. 489–494

SCIENCE

# A Specialization for Speech Perception

ALVIN M. LIBERMAN AND IGNATIUS G. MATTINGLY

# A Specialization for Speech Perception

ALVIN M. LIBERMAN AND IGNATIUS G. MATTINGLY

**The processes that underlie perception of consonants and vowels are specifically phonetic, distinct from those that localize sources and assign auditory qualities to the sound from each source. This specialization, or module, increases the rate of information flow, establishes the parity between sender and receiver that every communication system must have, and provides for the natural development of phonetic structures in the species and in the individual. The phonetic module has certain properties in common with modules that are "closed" (for example, sound localization or echo ranging in bats) and, like other members of this class, is so placed in the architecture of the auditory system as to preempt information that is relevant to its special function. Accordingly, this information is not available to such "open" modules as those for pitch, loudness, and timbre.**

PERCEIVING SPEECH IS GENERALLY ASSUMED TO BE NO different from perceiving sounds of other kinds (*1*). All of auditory perception is supposed to depend on various specializations, each one adapted to analyze the acoustic signal in a distinct way and to produce for cognition a correspondingly distinct representation. One of these specializations, "auditory scene analysis" (*2*), parses the signal, representing to cognition an array of localized sound sources; other specializations assign to each source appropriate values for such primitive auditory qualities as pitch, loudness, and timbre. The representations of, say, a squeaking door and a stop consonant differ only in the particular mix of values for these primitives. It is a later, cognitive stage that identifies the one mix as a squeaking door, the other as a stop consonant.

In the less conventional view that we mean to promote (*3*), the specifically phonetic aspects of speech perception are the articulatory gestures of which all linguistic utterances are ultimately composed (*4, 5*). Recurrent and phonologically significant patterns of these gestures, misleadingly called "speech sounds" or "phonetic segments," are the basis for the consonant and vowel symbols of phonetic transcriptions. Accordingly, the gestures stand apart from the paraphonetic aspects of speech—for example, voice quality and affective tone—which are presumably like nonspeech sounds in the nature of their perceptual primitives and in the specialized processes that evoke them. Perception of the gestures is different, for it is controlled by a "phonetic module" (*6*), a specialization for speech that has its own modes of signal analysis and its own primitives (*7*). Thus, phonetic perception is immediate; no cognitive translation from patterns of pitch, loudness, and timbre is required.

Several kinds of evidence have been offered in support of the claim that such a phonetic module does exist (*4, 5*). Here, we will be concerned only with a kind that seems especially telling and exemplary. Then, taking a broader view of the claim, we will describe the function that the phonetic module serves, say how it compares to other modules of the auditory system, and speculate about where it fits in the architecture they form.

## Evidence for a Phonetic Module: Duplex Perception

There is a species of psychoacoustic experiment that, in effect, dissects audition into its component processes, thus testing the hypothesis that the phonetic processes are distinct from the others. In this kind of experiment, two simultaneously presented parts of a stimulus are made discordant, or in some other way acoustically inconsistent with one another, with the result that they are heard as separate sound sources; but the information required for the perception of a particular speech sound is divided between the parts. The consequence is that one source is perceived as a nonspeech sound that, not surprisingly, depends on information in one of the two parts, but the other is perceived as the particular speech sound that depends on information in both. Thus, one of the two parts contributes to both the speech and nonspeech percepts at the same time. This phenomenon is called "duplex perception" (*8–14*).

Appropriate raw materials for an experiment on duplex perception are the control patterns, shown in Fig. 1, for two synthetic consonant-vowel syllables. Acoustic waveforms computed from these patterns consistently yield percepts that may be phonetically transcribed as [da] and [ga]. The bars in the figure represent vocal-tract resonances, called "formants," which vary in center frequency as the articulators assume particular configurations. The sloping resonances at the beginning of the pattern ("transitions") reflect the movements of the articulators as consonant and vowel gestures combine to form syllables.

In natural speech, the acoustic information that cues a phonetic gesture is dispersed, both in time and frequency, and there are many perceptually significant differences between the patterns for [da] and [ga]. For our experimental purposes, however, we have omitted or neutralized all but the difference in the transitions of the third (highest) formant. Fixed parts of the pattern, including, in particular, the first- and second-formant transitions, indicate that a stop consonant is being produced, and exclude stops other than [d] or [g]. Thus, given the full syllabic pattern, the perceived difference between [d] and [g] depends entirely on the differing transitions of the third formant (*15*). When presented in isolation, however, these transitions do not evoke speech percepts at all but, rather, nonspeech chirps of differing quality (*16*). The point of the experiment is to test

A. M. Liberman is a member of the research staff and former president of Haskins Laboratories, New Haven CT 06511, professor of psychology, emeritus, at the University of Connecticut, and professor of linguistics, emeritus, at Yale University. I. G. Mattingly is a member of the research staff of Haskins Laboratories and professor of linguistics at the University of Connecticut.

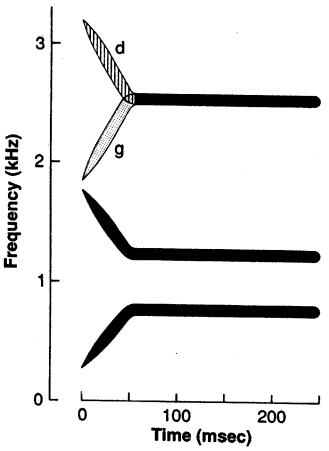


**Fig. 1.** Patterns that show how the perceived difference between [da] and [ga] can be made to depend on the slope of the initial third-formant transition.

whether these two ways of perceiving the same transition—as speech and as nonspeech—do, indeed, depend on different modules.

For the purposes of the experiment, each stimulus consists of two parts, as shown schematically in Fig. 2. One part, which is variable from presentation to presentation, is chosen from the series of third-formant transitions seen in Fig. 2A. These differ by equal steps in the frequencies at which they begin, covering a range that would, in full syllabic context, produce [ga] at the lower end of the scale and [da] at the higher, with a sharp break between the two at a point near the middle (*10, 11*). In isolation, each transition sounds, as we have said, like a nonspeech chirp, distinguishable from the others by its characteristic timbre. The other part of the stimulus, which is constant, is the remainder of the syllable, as shown in Fig. 2B. In isolation, this remainder sounds like a consonant-vowel syllable, but, lacking the critical third-formant transition, it is ambiguous, being judged sometimes as [da], sometimes as [ga]. These two parts are presented dichotically, the transition at one ear, the remainder at the other. As a consequence, the third formant moves from one ear to the other at 50 milliseconds, producing a sharp discontinuity in interaural intensity.

The perceptual result accords with the general account of this kind of experiment that we have already offered. Listeners hear two sounds, one at each ear. At the ear receiving the transition, they hear a nonspeech chirp, just as they do when the transition is presented in isolation. At the ear receiving the remainder of the syllable, they hear [da] or [ga]. But, surprisingly, these latter percepts are not ambiguous, as they are when the remainder is presented in isolation; rather, they are unambiguously determined to be [da] or [ga] by the slope of the transition at the other ear, just as they are when an undivided syllable is presented in the normal way. Yet this percept required that information be combined across two parts of the stimulus that are heard as different sound sources, and one of these parts (the third-formant transition) evoked, simultaneously, the nonspeech chirp and the perceived difference between [da] and [ga].

But such duplex perception speaks to our claim about an independent phonetic module only if the two percepts—chirp and consonant—are wholly distinct representations. We must, therefore, eliminate the possibility that the timbre of the transition, though represented only once, is being cognitively interpreted in two different ways, as if listeners were simply following a rule that this representation is to be called a chirp when presented in isolation, but a stop consonant when in combination with the remainder of the acoustic pattern. Two facts satisfy any concern we might have on this score. One is that, in duplex perception, listeners do not hear the ambiguous syllable that is evoked when the remainder of the pattern is presented in isolation (*9–11*). Thus, it cannot be that they are cognitively combining this percept with the chirp to get the unambiguous [da] or [ga]. The other relevant fact is that listeners are at chance when they try to match the isolated transitions to undivided [da]'s and [ga]'s; apparently they can neither hear [da] or [ga] in the chirps, nor chirps in the [da]'s and [ga]'s (*11, 14*).

We must, of course, also be sure that the two sides of the duplex percept are not merely two representations of the same kind, being composed of different combinations of the same primitives. They could be so interpreted if, as with the familiar visual examples, duplexity were the result of a shifting of attention between two representations of an ambiguous stimulus. But this interpretation is ruled out by the fact that the speech and nonspeech percepts are simultaneous and mandatory.

Additional evidence that the speech and nonspeech percepts are different kinds of representations comes from a further experiment on duplex perception, in which listeners were required to discriminate, on any basis, between two successive chirps heard at one ear, and, separately, between two successive speech sounds heard at the other, the two third-formant transitions in both cases being three steps apart on the series in Fig. 2A (*10*). As shown in Fig. 3, the discrimination functions for the nonspeech chirps and the speech are grossly different, though the stimuli that provided the only basis for the discrimination were identical. The function for the chirps is approximately linear, and conforms reasonably well to what is expected, given the results of psychoacoustic research on sloping resonances (*17*). On the speech side, however, the function is sharply peaked, reflecting a strong tendency to hear the stimuli in the nearly categorical manner that has been found to characterize the perception of phonetic structures (*18*).

Like the fact that transitions could not be matched to full syllables, this result of the discrimination test shows that listeners do not have access to any representation common to the speech and nonspeech aspects of the duplex percept. But it also strongly implies that the two representations they do have access to are of different kinds, being formed of different primitives—pitch and timbre in the one case, phonetic categories in the other. For the stimuli that evoked the two percepts were identical pairs of transitions, they were presented in a perfectly constant context, and they were discriminated according to the same psychophysical procedure. Yet the resulting functions had markedly different shapes, each one appropriate, it would seem, to the kind of representation on which it was based.

Thus, duplex perception supports the claim that phonetic repre-

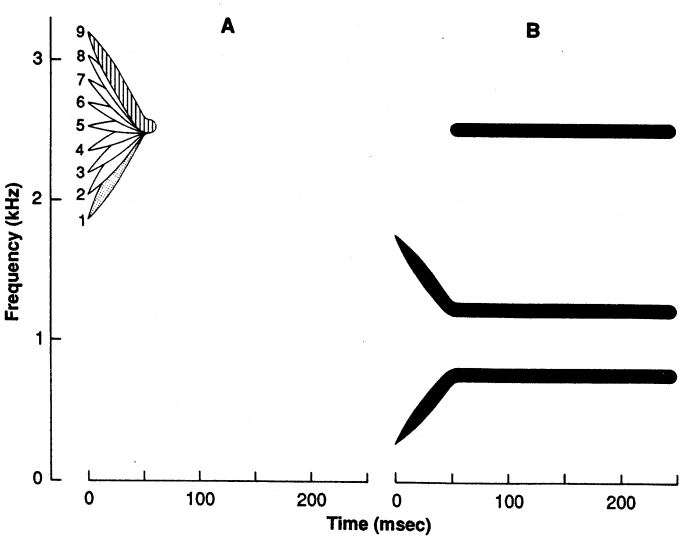


**Fig. 2.** Dichotic stimuli yielding duplex perception, derived from the patterns of Fig. 1. (**A**) A series of third-formant transitions covering the range from [ga] to [da]. (**B**) The constant remainder of the syllable.

sentations are formed by a distinct module, independently of the modules for pitch, loudness, and timbre. But it also shows that this phonetic module is independent of the module for scene analysis. For the [da] and [ga] percepts can only have been formed by combining information across two parts of a stimulus that are treated by scene analysis as separate sources at separate locations. This contrasts with the behavior of the modules for pitch, loudness, and timbre, since they did not combine information in this way, but rather attributed their auditory properties separately to each of the sources that scene analysis defined; hence the chirp-like character of one source and the voice-like character of the other. Thus, the phonetic module, but not the others, ignored what scene analysis had done, responding instead to a coherence that existed across both parts of the stimulus, though only in the phonetic domain, and that provided the only basis for assigning "da-ness" and "ga-ness" appropriately to the voice-like sources. The implication is that this module has its own, specifically phonetic criteria, different from those used by scene analysis, for determining what counts as one event and what counts as more than one. This seems the more remarkable when one takes into account how fundamental to auditory perception are the processes that assign sounds to localized sources. Apparently, the phonetic module is so independent that it somehow avoids those processes.

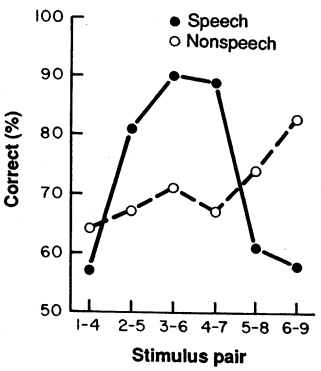


**Fig. 3.** Discrimination functions for the speech and nonspeech sides of the duplex percept. The stimulus pairs are numbered to correspond to the formant transitions in Fig. 2. Chance performance is 50% correct. [Adapted from (*10*)]

## The Function of a Phonetic Module

Why should phonetic gestures be treated in a special way? Why should language, so often regarded as a cognitive capacity of the highest order, turn out to operate, like scene analysis or pitch perception, at a level that is clearly precognitive? The answer lies in the special means by which the phonetic gestures manage their communicative function. Although they are, to be sure, appropriate for producing audible sounds, they are also more specifically adapted to serve as the structural elements of phonology (*4, 19*), a part of natural human grammatical capacity that, together with syntax, distinguishes language from all other forms of communication. The specific function of phonology is to make possible a vocabulary comprising vastly more than the number of holistically different sounds that humans can efficiently produce and perceive. This it does by providing a system for combining and permuting a few dozen gestures, specifically phonetic objects that belong, thus, to a natural class. But the system works in practice only because there is a specialization for producing these phonetic objects, that is, for translating the abstract gestural structures we call words and sentences into neuromotor commands for the articulator movements of particular utterances (*4*).

What is remarkable about this specialization is its capacity for parallel transmission of phonetic information. Such transmission is accomplished by coarticulation, that is, by overlapping the movements for gestures of different articulators and by merging into one continuous movement different gestures of the same articulator. The gestures that evolved for effective coarticulation are a distinct set, different, for the most part, from those people make when they lick their lips, chew, move food around the mouth, and so forth. These gestures are special, too, in the way that they are coordinated, both in space and in time. For the overlap and merging are thorough enough to permit the gestures to be produced at high rates, yet so controlled as to preserve information about their identities and their structural relation to one another.

While coarticulation greatly increases the rate at which phonetic information can be transmitted, it necessarily complicates the process of perception, because it precludes any simple correspondence between successive phonetic gestures and successive acoustic segments. Any particular stretch of a speech signal will commonly contain information about more than one gesture, and the acoustic information specifying a particular gesture will vary according to the phonetic context (*20*).

One sees, now, the function of a specialization for speech perception. Adapted specifically to cope with the peculiar complications of speech, it processes the acoustic signal so as to recover the coarticulated gestures that produced it. These gestures are the primitives that the mechanisms of speech production translate into actual articulator movements, and they are also the primitives that the specialized mechanisms of speech perception recover from the signal. Indeed, there is but one specialization with two complementary processes, one for computing the articulator movements and one for dealing with the acoustic consequences (*4, 5*).

This aspect of the phonetic specialization is important for its relevance to a truism that applies to all communication systems: what counts as structure in production must count as structure in perception, else communication does not occur. If not by the kind of Janus-faced specialization we have described, how is this parity established and maintained as the system develops in the species and as it develops anew in each member? The question is the more worth asking, because the conventional view of speech implies an unparsimonious answer. For if, in speech, the auditory and motor representations are distinct from each other, having in common only that neither is specifically phonetic, then both must be translated into phonetic representations by matching them to phonetic prototypes or otherwise assigning them to appropriate phonetic categories, much as the conventional view assumes (*1*). These prototypes and labels might have been established by agreement, as they obviously are in the case of invented communication systems like Morse code. Or, alternatively, they might have taken root as something like the "innate ideas" that some students of language invoke (*21*). In either case, however, they are neither motoric nor perceptual; they are, rather, cognitive, and their role is just to relate nonphonetic modes of acting and perceiving to each other and to language. But if speech has anything like the characteristics we have attributed to it, then there is no need for agreements or innate ideas and the cognitive translations they participate in. At every stage of phylogenetic and ontogenetic development, the single precognitive specialization for production and perception provides a common currency of specifically phonetic primitives, hence a sufficient basis for the parity between sender and receiver that must exist (*4, 5*).

Experiments on the biology of parity in phonetic communication are, for obvious reasons, hard to carry out, but there is at least some direct evidence that production and perception do, indeed, have common and specific neural loci (*22*). The biology of communication in nonhuman animals is, of course, more available to an experimenter, and, though such communication does not rest on a phonetic base, it is nonetheless subject to the requirement of parity. One is, therefore, not surprised to find, in the communication

systems of a variety of species, that the biological bases of perception and production are associated in an especially intimate way (*23*).

## Speech and the Rest of the Auditory System: Open and Closed Modules

The modules of the auditory system (and, presumably, of other modalities as well) can be divided into two classes, which we have called "open" and "closed," according to the kind of representations that they characteristically produce and the way they respond to environmental influences. Among the open modules are those for pitch, loudness, and timbre. These are adapted for the perception of an indefinitely large number of acoustic events, including many that evolution could not have anticipated. Accordingly, each responds straightforwardly to the dimensions of the signal—pitch to frequency, loudness to intensity, and timbre to spectral shape—so the representations they produce deserve to be called "homomorphic." These modules correspond roughly to those aspects of perceptual systems that, in Konishi's neurobiological classification, produce "projectional maps"—that is, central neural maps that preserve the spatial relations of the responses at the sensory epithelium (*24*).

Like all modules, those of the open class are influenced by environmental circumstances—witness the well-known effects of sensory deprivation—but such influences cannot be governed by acoustic events as such. Long experience with a variety of, say, squeaking doors must not make the open modules better adapted to that kind of event, lest they become that much less well adapted to other events that depend on a different mix of values of the same homomorphic primitives. Thus, such experience cannot change the internal working of the module, only the way its outputs are associated with the particular events that they come to signify.

The closed modules comprise a variety of specializations, including, for example, scene analysis, echo ranging, and phonetic perception. In Konishi's scheme, these belong to a class that requires "central synthesis" for formation of the appropriate neural map; direct projection will not suffice (*24*). In our terms, the members of this class have in common that the percepts they produce are "heteromorphic": the dimensions of the percept do not correspond directly to 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 a sounding object at different positions of azimuth (*25*). 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 emitted cry and its reflection (*26*), but what is perceived is presumably the distance of the reflecting object, not an echoing bat cry. Similarly, in our view, the phonetic 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.

As for environmental influences, the closed modules can respond, as the open modules cannot, by adapting their internal mechanisms, and hence their heteromorphic primitives, to just those events, or derived properties of events, they are concerned with; the homomorphic primitives that must be used for everything else are in no way affected. Consider, for example, how the sound-localization aspect of the scene-analysis module must adjust to changes in interaural time disparities as the head grows bigger. It can hardly be that the animal learns to translate old disparities into new locations, if only because its sound-localizing module never did perceive the disparities homomorphically as disparities. It must rather be that the module adjusts its internal processes, and hence its heteromorphic output (location of a source); thus, it is the module itself that learns. We should suppose that in the acquisition of any particular language, the phonetic module adjusts its internal processes and its hetermorphic representations in much the same epigenetic way. The child need never "translate" homomorphic auditory representations into the phonetic categories his language happens to represent.

Similar considerations apply, of course, to the development of phonetic perception in evolution. For the conventional assumption that speech and nonspeech 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. But if, as we speculate, speech is managed by a closed module, its processes were free to go as evolution took them.

## Architectural Relations Between Open and Closed Modules

The homomorphic and heteromorphic representations that characterize open and closed modules are sometimes formed in response to signals in the same physical range. This is most obviously the case in speech, where, as we have seen in the phenomenon of duplex perception, exactly the same stimulus that causes the open module for timbre to represent homomorphic chirps causes the closed phonetic module to produce the difference between the heteromorphic representations [da] and [ga]. Why, then, are not all speech percepts duplex in this way? Why, when listeners hear [da] and [ga], do they not also hear chirps?

A similar question arises in the case of the bat. Given that the closed echo-ranging module represents the echoes of the bat's cries heteromorphically as objects at certain distances, why do the open modules not also represent them homomorphically as bat cries? Presumably the open modules do not, even though they respond to the physically similar cries of other bats.

Duplex perception would, of course, be prevented in such cases if the open modules had gates through which the unwanted signals could not pass, or inhibitory processes that would nullify whatever responses they might evoke. But there are no superficial properties of the signal that such gates or inhibitors could use. They would, therefore, need the same capacities to respond to underlying properties that the closed modules are specialized for—obviously, an unparsimonious arrangement.

A more parsimonious solution is an architecture that allows a closed module to preempt just the information that concerns it, thus preventing this information from reaching the open modules at all (*4, 5*). Indeed, precisely this kind of arrangement seems to characterize the relation between the closed scene-analysis module and the various open modules. For, as Bregman has made clear, scene analysis must segregate the acoustic information into separate streams according to source, if pitch, timbre, and loudness are to be properly assigned (*2*). This is as much as to say that, with respect to the flow of information, this closed module is in series with, and precedes, each of the open modules. But scene analysis does not simply pass on all the information; rather, it preempts some of it in the very process of defining sound sources. Thus, a sufficiently great interaural disparity in time is taken to mean that the two signals correspond to two sources, regardless of their physical similarity, and the disparity is perceived as disparity. An appropriately small disparity, however, is used as evidence of the azimuth position of one source. Listeners hear this one source, but not the disparity, for that has been preempted in the process of localization. A similar architecture may define the relation of open and closed modules

more generally and serve to resolve the competition between them.

In the case of the phonetic module, evidence for preemptiveness has been reported in an example of duplex perception somewhat different from the one described earlier (*14*). The stimuli in this case are like those of the earlier example, except that the critical third-formant transitions are not resonances, but sinusoids that follow the center frequencies of the resonances. In isolation, these sinusoids sound like brief whistles, and, like the isolated resonances, they cannot be matched to [da] and [ga]. The point of the experiment was to see what happens as these sinusoids are increased in intensity from a level near zero, the sinusoids and resonances that form the remainder of the pattern being presented at both ears.

Within a certain range of intensities at the lower end of the scale, the sinusoids have an effect that is exclusively phonetic: listeners perceive [da] or [ga] appropriately; they do not also perceive whistles or any other kind of nonspeech that can be reliably associated with the sinusoids. In itself, this is of interest, since it offers further testimony to the ability of the phonetic module to respond to phonetic coherence, even though this may require ignoring a considerable discordance at the acoustic surface. Ignored in this case are the gross differences in fundamental frequency, spectrum, and harmonic structure between the sinusoids that critically distinguish [da] from [ga] and the resonances that form the remainder of the pattern.

With further increases in the intensity of the sinusoids, a point is reached at which they begin to serve a double purpose: listeners perceive [da] or [ga] appropriately, as before, but also one or another whistle, which they can reliably match to the whistles produced by the sinusoids in isolation. At first, this whistle is faint, but it grows steadily in loudness as the intensity of the sinusoid is further increased; meanwhile, perception of [da] and [ga] remains unchanged. As in the earlier example of duplex perception, the information in the transitions simultaneously produces speech and nonspeech percepts. In this case, however, it is apparent that the phonetic module is preemptive: it has first claim on the information in the sinusoids, allowing only the unwanted residue to evoke responses in the modules for pitch, loudness, and timbre.

The relation between the information the phonetic module receives and the information it passes on is not as yet clear. It certainly would not do to view the phonetic module as simply removing phonetically crucial portions of the represented signal (the formant transitions, for example), for the listener needs the paraphonetic information these portions also contain. Perhaps the action of the module is to be thought of as a kind of inverse filtering that undoes the effects of the resonant cavities of the vocal tract, leaving paraphonetic information about the excitation of the cavities as well as information about other ambient sounds. In the experiment just described, the residue for low intensities of the sinusoid is perceived just as laryngeal excitation, while at higher intensities a nonphonetic source, the whistle, also becomes obvious.

## A Remaining Problem

According to our account of auditory architecture, the closed scene-analysis module represents an array of sources to cognition and segregates acoustic information according to source. Given these separate streams of information, open modules then attribute pitch, timbre, and loudness to each source. The phonetic module, independently of scene analysis, uses all the relevant information available to form phonetic percepts; information not so preempted becomes available to the open modules.

Thus, in the human case, the scene-analysis module precedes the open modules in series; so, too, does the phonetic module. But what is the architectural relation of the two closed modules themselves? A similar question arises for the bat. If the bat hears the echo-ranging cries of other bats, along with other ambient sounds, it must hear them as separate sources, and must, therefore, have its own scene-analysis module preceding the open modules. What, then, is the architectural relation of scene analysis and echo ranging?

A parallel arrangement of closed modules must, presumably, be ruled out in both cases, for such an arrangement would obviously defeat the preempting functions of these modules: acoustic information preempted by the phonetic module (or by the echo-ranging module) would reach the open modules through the scene-analysis module, and conversely. Thus, the closed modules must be in series, both in man and in bat; only the ordering within the series is in question.

In the bat, we suggest that echo ranging comes before scene analysis. Scene analysis has no concern with signals that originate in the bat itself; their preemption would simplify its task. Echo ranging has no need to know about the auditory scene: the emitted sonar signal and its echo are already defined sources, and the bat determines the position in azimuth of the reflecting object by pointing its head so as to minimize the interaural disparity of the echo signals (*27*).

In man, we have conjectured that, in similar fashion, the closed phonetic module precedes the closed module for scene analysis (*12*). In the examples of duplex perception offered here, the phonetic module makes no use of the separation of sources provided by scene analysis. If the phonetic module came after scene analysis, it would be reintegrating 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. Such an arrangement is not very parsimonious. 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.

Unfortunately for such conjectures, it is the unparsimonious alternative that the evidence so far seems to favor. This evidence is owed to Darwin (*28, 29*), who has adduced a number of examples in which phonetic integration does not occur, although it might be expected to if the phonetic module precedes scene analysis in series. In one such example, a vowel is first synthesized without a particular harmonic it would naturally contain, with the result that it is perceived as different in quality from a synthetic vowel not thus depleted. If a tone equal in amplitude and frequency to the missing harmonic is added synchronously to the depleted vowel, the sum is of course perceived as the undepleted vowel, and the tone is not separately heard. But if the depleted vowel is short enough, and its onset follows that of the tone by some tens of milliseconds, the depleted vowel and, separately, the tone itself are heard (*28*). It is as if scene analysis, preceding the phonetic module, had defined the asynchronous tone and vowel as separate events, which the phonetic module must then either use for their entire durations or not use at all. In light of effects such as this, it may be necessary to reconsider our simple account of the serial ordering of the two closed modules.

## Broader Issues

Taken in their most general terms, the questions raised here are not necessarily limited to the phonetic domain; they can, rather, be extended in two directions. One looks toward the other aspects of language, where investigators have for some time been exploring the possibility that syntax, like phonetics, is part of a distinct, precognitive module, and not, as more commonly assumed, one among

many expressions of a general capacity for cognitive computation (7). The other direction leads to any perceptual system that can be characterized as a group of modules; there it might prove rewarding to ask, further, how the modules can usefully be classified and what the architectural arrangements among them might be (*24, 30*).

## REFERENCES AND NOTES


1. R. G. Crowder and J. Morton, *Percept. Psychophys.* **5**, 365 (1969); K. N. Stevens, in *Auditory Analysis and Perception of Speech*, G. Fant and M. A. Tatham, Eds. (Academic Press, New York, 1975), pp. 303–330; J. D. Miller, in *Recognition of Complex Acoustic Signals*, T. H. Bullock, Ed. (Dahlem Konferenzen, Berlin, 1977), pp. 49–58; G. C. Oden and D. W. Massaro, *Psychol. Rev.* **85**, 172 (1978); P. K. Kuhl, *J. Acoust. Soc. Am.* **70**, 340 (1981).
2. A. S. Bregman, in *Attention and Performance*, J. Requin, Ed. (Erlbaum, Hillsdale, NJ, 1978), vol. 7, pp. 62–74.
3. See A. M. Liberman and I. G. Mattingly, "Signal and sense: Local and global order in perceptual maps," paper presented at the Fifth Annual Symposium of the Neurosciences Institute, Stockholm, Sweden, 31 May to 5 June 1987.
4. ———, *Cognition* **21**, 1 (1985).
5. I. G. Mattingly and A. M. Liberman, in *Auditory Function: Neurobiological Bases of Hearing*, G. M. Edelman, W. E. Gall, W. M. Cowan, Eds. (Wiley, New York, 1988), pp. 775–793.
6. More properly, a *linguistic* module that not only perceives phonetic gestures, but also recognizes words and parses sentences. Our concerns, however, have been chiefly phonetic.
7. See J. Fodor, *The Modularity of Mind* (MIT Press, Cambridge, MA, 1983).
8. T. C. Rand, *J. Acoust. Soc. Am.* **55**, 678 (1974); A. M. Liberman, in *Proceedings of the Ninth International Congress of Phonetic Science*, E. Fischer-Jorgensen, J. Rischel, N. Thorsen, Eds. (Univ. of Copenhagen Press, Copenhagen, 1979), vol. 2, pp. 468–473; J. E. Cutting, *Psychol. Rev.* **83**, 114 (1976); B. Repp and S. Bentin, *Percept. Psychophys.* **36**, 523 (1984).
9. A. M. Liberman, D. Isenberg, B. Rakerd, *Percept. Psychophys.* **30**, 133 (1981).
10. V. A. Mann and A. M. Liberman, *Cognition* **14**, 211 (1983).
11. B. Repp, C. Milburn, J. Askenas, *Percept. Psychophys.* **33**, 333 (1983).
12. I. G. Mattingly, *J. Acoust. Soc. Am.* **82** (Suppl. 1), 120 (1987).
13. A. S. Bregman, in *The Psychophysics of Speech Perception*, M. E. H. Schouten, Ed. (Nijhoff, Dordrecht, 1987), pp. 95–111.
14. D. Whalen and A. M. Liberman, *Science* **237**, 169 (1987).
15. K. S. Harris, H. S. Hoffman, A. M. Liberman, P. C. Delattre, F. S. Cooper, *J. Acoust. Soc. Am.* **30**, 122 (1958).
16. I. G. Mattingly, A. M. Liberman, A. M. Syrdal, T. Halwes, *Cogn. Psychol.* **2**, 131 (1971).
17. P. T. Brady, A. N. House, and K. N. Stevens, *J. Acoust. Soc. Am.* **33**, 1357 (1961).
18. A. M. Liberman, K. S. Harris, H. S. Hoffman, B. C. Griffith, *J. Exp. Psychol.* **61**, 379 (1961); A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, *Psychol. Rev.* **74**, 431 (1967); B. Repp and A. M. Liberman, in *Categorical Perception*, S. Harnad, Ed., (Cambridge Univ. Press, Cambridge, 1987), pp. 89–112.
19. C. P. Browman and L. Goldstein, *Phonol. Yearb.* **3**, 219 (1986).
20. F. S. Cooper, P. C. Delattre, A. M. Liberman, J. Borst, L. Gerstman, *J. Acoust. Soc. Am.* **24**, 597 (1952); G. Fant, *Logos* **5**, 3 (1962).
21. N. Chomsky, *Cartesian Linguistics* (Harper & Row, New York, 1966).
22. G. Ojemann and C. Mateer, *Science* **205**, 1401 (1979).
23. R. Hoy and R. C. Paul, *ibid.* **180**, 82 (1973); R. Hoy, J. Hahn, R. C. Paul, *ibid.* **195**, 82 (1977); H. C. Gerhardt, *ibid.* **199**, 992 (1978); J. S. McCasland and M. Konishi, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7815 (1983); D. Margoliash, *J. Neurosci.* **3**, 1039 (1983); H. Williams and F. Nottebohm, *Science* **229**, 279 (1985); M. J. Ryan and W. Wilczynski, *Science* **240**, 1786 (1988).
24. M. Konishi, *Trends Neurosci.* **9**, 163 (1986).
25. E. Hafter, in *Dynamic Aspects of Neocortical Function*, G. M. Edelman, W. E. Gall, W. M. Cowan, Eds. (Wiley, New York, 1984), pp. 425–448.
26. N. Suga, in *ibid.*, pp. 315–373.
27. J. Simmons, in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds. (Springer, New York, 1987), pp. 214–225.
28. C. J. Darwin, *J. Acoust. Soc. Am.* **76**, 1636 (1984).
29. ——— and N. S. Sutherland, *Q. J. Exp. Psychol.* **36A**, 193 (1984).
30. M. Livingstone and D. Hubel, *Science* **240**, 740 (1988); J. M. Wolfe, *Psychol. Rev.* **93**, 269 (1986); R. Blake and R. P. O'Shea, *ibid.* **95**, 151 (1988); J. M. Wolfe, *ibid.*, p. 155.


31. Support from NIH grant H-01994 to Haskins Laboratories is gratefully acknowledged.