OFFPRINTS FROM AUDITORY FUNCTIOIN
Edited by G. M. Edelman, W. E. Gall and W. M. Cowan


# Chapter 26

# Specialized Perceiving Systems for Speech and Other Biologically Significant Sounds

IGNATIUS G. MATTINGLY
ALVIN M. LIBERMAN

**ABSTRACT**

*Perception of speech rests on a specialized mode, narrowly adapted for the efficient production and perception of phonetic structures. This mode is similar in some of its properties to the specializations that underlie, for example, sound localization in the barn owl, echolocation in the bat, and song in the bird.*

Our aim is to present a view of speech perception that runs counter to the conventional wisdom; [redacted]speech perception is to humans as sound localization is to barn owls. This is not merely to suggest that humans are preoccupied with listening to speech, much as owls are with homing in on the sound of prey. It is to offer a particular hypothesis: Like sound localization, speech perception is a coherent system in its own right, specifically adapted to a narrowly restricted class of ecologically important events. In this important respect, speech perception and sound localization are more similar to each other than is either to the processes that underlie the perception of such ecologically arbitrary events as squeaking doors, rattling chains, or whirring fans.

To develop the unconventional view, we contrast it with its more conventional opposite, say why the less conventional view is nevertheless the more plausible, and describe several properties of the speech-perceiving system that the unconventional view reveals. We compare speech perception with other specialized perceiving systems that also treat acoustic signals, including not only sound localization in the owl, but also song in the bird and echolocation in the bat. Where appropriate, we develop the neurobiological implications, but we do not try here to fit them to the vast and diverse literature that pertains to the human case.

Through most of this chapter we construe speech, in the narrow sense, as referring only to consonants and vowels. Then, at the end, we briefly say how our view of speech might nevertheless apply more broadly to sentences. Following the instructions of the editors, we discuss primarily issues and principles; however,

we do offer the results of a few experiments, not so much to prove our argument as to illuminate it. [For full accounts of these experiments and the many others that support the claims we will be making below, see Liberman et al. (1967), Liberman and Mattingly (1986), and the studies referred to therein.]

## TWO VIEWS OF SPEECH PERCEPTION: GENERALLY AUDITORY VERSUS SPECIFICALLY PHONETIC

The conventional view derives from the common assumption that mental processes are not specific to the real-world events to which they are applied. Thus perception of speech is taken to be in no important way different from perception of other sounds. [Not surprisingly, there are a number of variations on the "conventional view"; they are discussed in Liberman and Mattingly (1986).] In all cases, it is as if the primitive auditory consequences of acoustic events were delivered to a common register (the primary auditory cortex?), from whence they would be taken for such cognitive treatment as might be necessary in order to categorize each ensemble of primitives as representative of squeaking doors, stop consonants, or some other class of acoustic events. On any view, there are, of course, specializations for each of the several auditory primitives that together make up the auditory modality, but there is surely no specialization for squeaking doors as such, and, on the conventional view, none for stop consonants either.

Our view is different on all counts. Seen our way, speech perception takes place in a specialized phonetic mode, different from the general auditory mode and served, accordingly, by a different neurobiology. Contrary to the conventional assumption, there is then a specialization for consonants and vowels as such. This specialization yields only phonetic structures; it does not deliver to a common auditory register those sensory primitives that might, in arbitrarily different combinations, be cognitively categorized as any of a wide variety of ordinary acoustic events. Thus specialization for perception of phonetic structures begins prior to such categorization and is independent of it.

The phonetic mode is not auditory, in our view, because the events it perceives are not acoustic. They are, rather, gestural. For example, the consonant [b] is a lip-closing gesture; [h] is a glottis-opening gesture. Combining lip-closing and glottis-opening yields [p]; combining lip-closing and velum-lowering yields [m], and so on. Despite their simplistic labels, the gestures are in fact quite complex: As we shall see, a gesture usually requires the movements of several articulators, and these movements are most often context-sensitive. A rigorous definition of a particular gesture has, therefore, to be fairly abstract. Nevertheless, it is the gestures that we take to be the primitives of speech perception, no less than of speech production. Phonetic structures are patterns of gestures, then, and it is just these that the speech system is specialized to perceive.

## THE PLAUSIBLE FUNCTION OF A SPECIFICALLY PHONETIC MODE

But why should consonants and vowels be gestures, not sounds, and why should it take a specialized system to perceive them? To answer these questions, it is

helpful to imagine the several ways in which phonetic communication might have been engineered.

Accepting that Nature made a firm commitment to an acoustic medium, we can suppose that she might have defined the phonetic segments—consonants and vowels—in acoustic terms. This, surely, is what common sense suggests, and, indeed, what the conventional view assumes. The requirements that follow from this definition are simply that the acoustic signals be appropriate to the sensitivities of the ear, and that they provide the invariant basis for the correspondingly invariant auditory percept by which each phonetic segment is to be communicated. The first requirement is easy enough to satisfy, but the second is not. For if the sounds are to be produced by the organs of the vocal tract, then strings of acoustically defined segments require strings of discrete gestures. Such strings can be managed, of course, but only at unacceptably slow rates. Indeed, we know exactly how slow, because speaking so as to produce a segment of sound for each phonetic segment is what we do when we spell. Thus, to articulate the consonant–vowel syllables [di] and [du], for example, the speaker would have to say something like [də] [i] and [də] [u], converting each consonant and each vowel into a syllable. Listening to such spelled speech, letter by painful letter, is not only time-consuming but also maddeningly hard.

Nature might have thought to get around this difficulty by abandoning the vocal tract in favor of a to-be-developed set of sound-producing devices, specifically adapted for creating the drumfire that communication via acoustic segments would require if speakers were to achieve the rates that characterize speech as we know it, rates that run at eight to ten segments per second, on average, and at double that for short stretches. But this would have defeated the ear, severely straining its capacity to identify the separate segments and keep their order straight.

Our view is that Nature solved the problems of rate by avoiding the acoustic strategy that gives rise to them. The alternative was to define the phonetic segments as gestures, letting the sound go pretty much as it might, so long as the acoustic consequences of the different gestures were distinct. On its face, this seems at least a reasonable way to begin, for it takes into account that phonetic structures are not really objects of the acoustic world anyway; they belong, rather, to a domain that is internal to the speaker, and it is the objects of this domain that need to be communicated to the listener. But the decisive consideration in favor of the gestural strategy is surely that it offers critical advantages for rate of communication, both in production and in perception. These advantages were not to be had, however, simply by appropriating movements that were already available, for example, those of eating and breathing. Rather, the phonetic gestures and their underlying controls had to be developed, presumably as part of the evolution of language. Thus, as we will argue later, speech production is as much a specialization as speech perception; as we will also argue, it is indeed the same specialization.

In production, the advantage of the gestural strategy is that, given the relative independence of the muscles and organs of the vocal tract and the development of appropriately specialized controls, gestures belonging to successive segments in the phonetic string can be executed simultaneously or with considerable overlap. Thus the gesture for [d] is overlapped with com-

ponent gestures for the following vowel, whether [i] or [u]. By just such coarticulation, speakers achieve the high rates at which phonetic structures are in fact transmitted, rates that would be impossible if the gestures had to be produced seriatim.

In perception, the advantage of the gestural strategy is that it provides the basis for evading the limit on rate that would otherwise have been set by the temporal resolving abilities of the auditory system. This, too, is a consequence of coarticulation. Information about several gestures is packed into a single segment of sound, thereby reducing the number of sound segments that must be dealt with per unit of time.

But the gain for perception is not without cost, for if information about several gestures is transmitted at the same time, the relation between these gestures and their acoustic vehicles cannot be straightforward. It is, to be sure, systematic, but only in a way that has two special and related consequences. First, there is no one-to-one correspondence in segmentation between phonetic structure and signal; information about the consonant and the vowel can extend from one end of the acoustic syllable to the other. Second, the shape of the acoustic signal for each particular phonetic gesture varies according to the nature of the concomitant gestures and the rate at which they are produced. Thus the cues on which the processes of speech perception must rely are context conditioned. For example, the perceptually significant second-formant transition for [d] begins high in the spectrum and rises for [di], but begins low in the spectrum and falls for [du].

How might the complications of this unique relation have been managed? Consider first the possibility that no further specialization is provided, the burden being put on the perceptual and cognitive equipment with which the listener is already endowed. By this strategy, the listener uses ordinary auditory processes to convert the acoustic signals of speech to ordinary auditory percepts. But then, having perceived the sound, the listener must puzzle out the combination of coarticulated gestures that might have produced it, or, failing that, learn ad hoc to connect each context-conditioned and eccentrically segmented token to its proper phonetic type. However, the puzzle is so thorny as to have proved so far to be beyond the capacity of scientists to solve; and given the large number of acoustic tokens for each phonetic type, ad hoc learning might well have been endless. Moreover, listening to speech would have been a disconcerting experience at best, for the listener would have been aware not only of phonetic structure but also of the auditory base from which phonetic structure would have had to be recovered. We gain some notion of what this experience would have been like when we hear, in isolation from their contexts, the second-formant transitions that cue [di] and [du]. As would be expected on psychoacoustic grounds, the transition for [di] sounds like a rising glissando on high pitches (or a high-pitched chirp); the transition for [du] like a falling glissando on low pitches (or a low-pitched chirp). If the second-formant transition is combined with the concomitant transitions of other formants, the percept becomes a "bleat" whose timbre depends on the nature of the component transitions. Fluent speech, should it be heard in this auditory way, would thus be a rapid sequence of qualitatively varying bleats. The plight of the listener who had to base a cognitive

analysis of phonetic structure on such auditory percepts would have been like that of a radio operator trying to follow a rapid-fire sequence of Morse code dots and dashes, only worse, because, as we have seen, the "dots and dashes" of the speech code take as many different acoustic forms as there are variations in context and rate.

The other strategy for recovering phonetic structure from the sound—the one that must have prevailed—was to use an appropriate specialization. Happily, this specialization was already at hand in the form of those arrangements, previously referred to, that made it possible for speakers to articulate and coarticulate phonetic gestures. These must have incorporated in their architecture all the constraints of anatomy, physiology, and phonetics that organize the movements of the speech organs and govern their relation to the sound, so access to this architecture should have made it possible, in effect, to work the process in reverse—that is, to use the acoustic signal as a basis for computing the coarticulated gestures that caused it. It is just this kind of perception–production specialization that our view assumes. Recovering phonetic structure requires, then, no prodigies of conscious computation or arbitrary learning. To perceive speech, a person has only to listen, for the specialization yields the phonetic percept immediately. This is to say that there is no conscious mediation by an auditory base. The gestures for consonants and vowels, as perceived, are themselves the distal objects; they are not, like the dots and dashes of Morse code (or the squeak of the door), at one remove from it. But perception is immediate in this case (and in such similar cases as, for example, sound localization), not because the underlying processes are simple or direct, but only because they are well suited to their unique and complex task.

## SOME PROPERTIES OF THE PHONETIC MODE COMPARED WITH THOSE OF OTHER PERCEPTUAL SPECIALIZATIONS

Every perceptual specialization must differ from every other in the nature of the distal events it is specialized for, as it must, too, in the relation between these events and the proximal stimuli that convey them. At some level of generality, however, there are properties of these specializations that invite comparison. Several of the properties that are common, perhaps, to all perceiving specializations—for example, "domain specificity," "mandatory operation," and "limited central access"—have been described by Fodor (1983) and claimed by us to be characteristic of the phonetic mode (Liberman and Mattingly, 1986). We do not review these here, but choose to put our attention on four properties of the phonetic mode that are not so widely shared and that may, therefore, define several subclasses.

### Heteromorphy

The phonetic mode, as we have conceived it, is "heteromorphic" in the sense that it is specialized to yield perceived objects whose dimensionalities are radically

different from those of the proximal stimuli.* Thus the synthetic formant transitions that are perceived homomorphically in the auditory mode as continuous glissandi are perceived heteromorphically in the phonetic mode as consonant or vowel gestures that have no glissandolike auditory qualities at all. But is it not so in sound localization too? Surely, interaural disparities of time and intensity are perceived heteromorphically, as locations of sound sources, and not homomorphically, as disparities, unless the interaural differences are of such great magnitude that the sound-localizing specialization is not engaged. Thus the heteromorphic relation between distal object and the display at the sense organ is not unique to phonetic perception. Indeed, it characterizes not only sound localization but also echolocation in the bat, if we can assume that, as Suga's (1984) neurobiological results imply, the bat perceives not echo time as such but rather something more like the distance it measures. If we look to vision for an example, we find an obvious one in stereopsis, where perception is not of two-dimensionally disparate images but of third-dimensional depth.

To see more clearly what heteromorphy is, let us consider two striking and precisely opposite phenomena of speech perception together with such parallels as may be found in sound localization. In one of these phenomena, two stimuli of radically different dimensionalities converge on a single, coherent percept; in the other, stimuli lying on a single physical dimension diverge into two different percepts. In neither case can the contributions of the disparate or common elements be detected.

***Convergence on a Single Percept: Equivalence of Acoustic and Optical Stimuli.*** The most extreme example of convergence in speech perception was discovered by McGurk and McDonald (1976). As slightly modified for our purpose, it takes the following form. Subjects are repeatedly presented with the acoustic syllable [ba] as they watch the optical syllables [bɛ], [vɛ], [ðɛ], and [dɛ] being silently articulated by a mouth shown on a video screen. (The acoustic and optical syllables are approximately coincident.) The compelling percepts that result are of the syllables [ba], [va], [ða], and [da]. Thus the percepts combine acoustic information about the vowels with optical information about the consonants, yet subjects are not aware—indeed, they cannot become aware—of the bimodal nature of the percept.

This phenomenon is heteromorphy of the most profound kind, for if optical and acoustic contributions to the percept cannot be distinguished, then surely the percept belongs to neither of the modalities, visual or auditory, with which these classes of stimuli are normally associated. Recalling our claim that phonetic perception is not auditory, we add now that it is not visual, either. The phonetic

* Our notion of heteromorphy as a property of one kind of perceiving specialization seems consistent with comments about sound localization by Knudsen and Konishi (1978), who have observed that "[the barn owl's] map of auditory space is an emergent property of higher-order neurons, distinguishing it from all other sensory maps that are direct projections of the sensory surface . . . these space-related response properties and functional organization must be specifically generated through neuronal integration in the central nervous system. . . ." Much the same point has been made by Yin and Kuwada (1984), who say that "the cochlea is designed for frequency analysis and cannot encode the location of sound sources. Thus, the code for location of an auditory stimulus is not given by a 'labeled line' from the receptors, but must be the result of neural interactions within the central auditory system."

mode accepts all information, acoustic or optical, that pertains in a natural way to the phonetic events it is specialized to perceive. Its processes are not bound to the modalities associated with the stimuli presented to the sense organs; rather, they are organized around the specific behavior they serve and thus to their own phonetic "modality."

An analogue to the convergence of acoustic and optical stimuli in phonetic perception is suggested by the finding of neural elements in the optic tectum of the barn owl that respond selectively, not only to sounds in different locations, but also to lights in those same locations (Knudsen, 1984). Do we dare assume that the owl can't really tell whether it heard the mouse or saw it? Perhaps not, but in any case, we might suppose that, as in phonetic perception, the processes are specific to the biologically important behavior. If so, then perhaps we should speak of a mouse-catching "modality."

Putting our attention once more on phonetic perception, we ask: "Where does the convergence occur?" Conceivably, for the example we offered, "auditory" and "visual" processes succeed separately in extracting phonetic units. Thus the consonant might have been visual, the vowel auditory. These would then be combined at some later stage and perhaps in some more cognitive fashion. Of course, such a possibility is not wholly in keeping with our claim that speech perception is a heteromorphic specialization, nor, indeed, does it sit well with the facts now available. Evidence against a late-stage, cognitive interpretation is that the auditory and visual components cannot be distinguished phenomenally, and that convergence of the McGurk–McDonald type does not occur when printed letters, which are familiar but arbitrary indices of phonetic structure, are substituted for the naturally revealing movements of the silently articulating mouth. Additional and more direct evidence showing that the convergence occurs at an early stage, before phonetic percepts are formed, is available from a recent experiment by Green and Miller (1985; see also Summerfield, 1979). The particular point of this experiment was to test whether optically presented information about rate of articulation affects placement on an acoustic continuum of a boundary known to be rate-sensitive, such as the one between [bi] and [pi]. Before the experiment proper, it was determined that viewers could estimate rate of articulation from the visual information alone, but could not tell which syllable, [bi] or [pi], had been produced; we may suppose, therefore, that there was no categorical phonetic information in the optical display. Nevertheless, in the main part of the experiment, the optical information about rate did affect the acoustic boundary for the phonetic contrast; moreover, the effect was consistent with what happens when the information about rate is entirely acoustic. We should conclude then that the visual and auditory information converged at some early stage of processing before anything like a phonetic category had been extracted. This is what we should expect of a thoroughly heteromorphic specialization to which acoustic and optical stimuli are both relevant, and it fits as well as may be with the discovery in the owl of bimodally sensitive elements in centers as low as the optic tectum.

***Convergence on a Coherent Percept: Equivalence of Different Dimensions of Acoustic Stimulation.*** Having seen that optical and acoustic information can be indistinguishable when, in heteromorphic specialization, they specify the same distal

object, we turn now to a less extreme and more common instance of convergence in speech perception: The convergence of the disparate acoustic consequences of the same phonetic gesture, measured most commonly by the extent to which these can be "traded," one for another, in evoking the phonetic percept for which they are all cues. If, as such trading relations suggest, the several cues are truly indistinguishable, and therefore perceptually equivalent, we should be hard put, given their acoustic diversity, to find an explanation in auditory perception. Rather, we should suppose that they are equivalent only because the speech perceiving system is specialized to recognize them as products of the same phonetic gesture.

A particularly thorough exploration of such equivalence was made with two cues for the stop consonant [p] in the word *split* (Fitch et al., 1980). To produce the stop and thus to distinguish *split* from *slit*, a speaker must close and then open his lips. The closure causes a period of silence between the noise of the [s] and the vocalic portion of the syllable; the opening produces particular formant transitions at the beginning of the vocalic portion. Each of these—the silence and the transition—is a sufficient cue for the perceived contrast between *split* and *slit*. Now the acid test of their equivalence would be to show that the *split–slit* contrast produced by the one cue cannot be distinguished from the contrast produced by the other. Unfortunately, to show this would be to prove the null hypothesis. So equivalence was tested, somewhat less directly, by assuming that truly equivalent cues would either cancel each other or summate, depending on how they were combined. The silence and transition cues for *split–slit* passed the test: Patterns that differed by two cues weighted in opposite phonetic directions (one biased *for* [p], the other against) were harder to discriminate than patterns that differed by the same two cues weighted in the same direction (both biased *for* [p]).

A similar experiment done subsequently on the contrast between *say* and *stay* (Best et al., 1981) yielded similar results but with an important addition. In one part of this later experiment, the formants of the synthetic speech stimuli were replaced by sine waves made to follow the formant trajectories. As had been found previously, such sine wave analogues are perceived under some conditions as complex nonspeech sounds—chords, glissandi, and the like—but under others as speech (Remez et al., 1981). For those subjects who perceived the sine wave analogues as speech, the discrimination functions were much as they had been in both experiments with the full-formant stimuli. But for subjects who perceived the patterns as nonspeech, the results were different: Patterns that differed by two cues were about equally discriminable, regardless of the direction of a bias in the phonetic domain; these two-cue patterns were both more discriminable than those differing by only one. Thus the silence cue and the transition cue are equivalent only when they are perceived in the phonetic mode as cues for the same gesture.

If we seek parallels for such equivalences in the sound-locating faculty, we find one, perhaps, in data obtained with human beings. There, binaural differences in time and in intensity are both cues to location in azimuth, and there, also, it has been found that the two cues truly cancel each other, though not completely (Hafter, 1984).

We consider equivalences among stimuli—whether between stimuli belonging to different modalities, as traditionally defined, or between stimuli that lie on

different dimensions of the same modality—to be of particular interest, not only because they testify to the existence of a heteromorphic specialization, but also because they provide a way to define its boundaries.

*Divergence into Two Percepts: Nonequivalence of the Same Dimension of Acoustic Stimulation in Two Modes.* We have remarked that a formant transition (taken as an example of a speech cue) can produce two radically different percepts: a glissando or chirp when the transition is perceived homomorphically in the auditory mode as an acoustic event, or a consonant, for example, when it is perceived heteromorphically in the phonetic mode as a gesture. But it will not have escaped notice that the acoustic context was different in the two cases—the chirp was produced by a transition in isolation, the consonant by the transition in a larger acoustic pattern—and the two percepts, of course, were not experienced at the same time. It would surely be a stronger argument for the existence of two neurobiologically distinct processes, and for the heteromorphic nature of one of them, if, with acoustic context held constant, a transition could be made to produce both percepts in the same brain and at the same time. Under normal conditions, such maladaptive "duplex" perception never occurs, presumably because the underlying phonetic and auditory processes are so connected as to prevent it. (In a later section, we will consider the form this connection might take.) By resort to a most unnatural procedure, however, experimenters have managed to undo the normal connection and so produce a truly duplex percept (Rand, 1974; Liberman, 1979). Into one ear—it does not matter critically which one—the experimenter puts one or another of the third-formant transitions (called the "isolated transition") that lead listeners to perceive two otherwise identical formant patterns as [da] or [ga]. By themselves, these isolated transitions sound like chirps, and listeners are at chance when required to label them as [d] or [g] (Repp et al., 1983). Into the other ear is put the remaining, constant portion of the pattern (called the "base"). By itself, the base sounds like a consonant–vowel syllable, ambiguous between [da] and [ga]. But if the two stimuli are presented dichotically and in approximately the proper temporal arrangement, then, in the ear stimulated by the base, listeners perceive [da] or [ga], depending on which isolated transition was presented, while in the other ear they perceive a chirp. The [da] or [ga] is not different from what is heard when the full pattern is presented binaurally, nor is the chirp different from what is heard when the transition is presented binaurally without the base.

It is, perhaps, not to be wondered at that the dichotically presented inputs fuse to form the "correct" consonant–vowel syllable, since there is a strong underlying coherence. What is remarkable is that the chirp continues to be perceived, though the ambiguous base syllable does not. This is to say that the percept is precisely duplex, not triplex. Listeners perceive in the only two modes available: the auditory mode, in which they perceive chirps, and the phonetic mode, in which they perceive consonant–vowel syllables.

The sensitivities of these two modes are very different, even when stimulus variation is the same. This was shown with a stimulus display, appropriate for a duplex percept, in which the third-formant transition was the chirp and also the cue for the perceived difference between [da] and [ga] (Mann and Liberman, 1983). Putting their attention sometimes on the "speech" side and sometimes on the "chirp" side of the duplex percept, subjects discriminated various pairs

of stimuli. The resulting discrimination functions were very different, though the transition cues had been presented in the same context, to the same brain, and at the same time: The function for the chirp side of the duplex percept was linear, implying a perceived continuum, while the function for the phonetic side rose to a high peak at the location of the phonetic boundary (as determined for binaurally presented syllables), implying a tendency to categorize the percepts as [da] or [ga].

These results with psychophysical measures of discriminability are of interest because they support our claim that heteromorphic perception in the phonetic mode is not a late-occurring interpretation (or match-to-prototype) of auditory percepts that were available in a common register. Apparently, heteromorphic perception goes deep.

The facts about heteromorphy reinforce the view expressed earlier that the underlying specialization must become distinct from the specializations of the homomorphic auditory system at a relatively peripheral stage. In this respect, speech perception in the human is like echolocation in the bat. Both are relatively late developments in the evolution of human and bat, respectively, and both apparently begin their processing independently of the final output of auditory specializations that are older.

## Generative Detection

Since there are many other environmental signals in the same frequency range to which the speech-perceiving system must be sensitive, we should wonder how speech signals as a class are detected, and what keeps this system from being jammed by nonspeech signals that are physically similar. One possibility is that somewhere in the human brain there is a preliminary sorting mechanism that directs speech signals to the heteromorphic speech-perceiving system and other signals to the collection of homomorphic systems that deal with environmental sounds in general. Such a sorting mechanism would necessarily rely not on the deep properties of the signal that are presumably used by the speech-perceiving system to determine phonetic structure but on superficial properties like those that man-made speech-detection devices exploit: quasi-periodicity, characteristic spectral structure, and syllabic rhythm, for example.

The idea of a sorting mechanism is appealing because it would explain not only why the speech perceiving system is not jammed but, in addition, why speech is not also perceived as nonspeech—a problem to which we have already referred and to which we will return. Unfortunately, this notion is not easy to reconcile with the fact that speech is perceived as speech even when its characteristic superficial properties are masked or destroyed. Thus speech can be high-pass filtered, low-pass filtered, infinitely clipped, spectrally inverted, or rate adjusted, and yet remain more or less intelligible. Even more remarkably, intelligible speech can be synthesized in very unnatural ways: for example, as already mentioned, with a set of frequency-modulated sinusoids whose trajectories follow those of the formants of some natural utterance. Evidently, information about all these signals reaches the speech perceiving system and is processed by it, even though they lack some or all of the characteristic superficial properties on which the sorting mechanism we have been considering would have to depend.

The only explanation consistent with these facts is that there *is* no preliminary sorting mechanism; it is, instead, the speech perceiving system itself that decides between speech and nonspeech, exploiting the phonetic properties that are intrinsic to the former and only fortuitously present in the latter. Presumably, distorted and unnatural signals like those we have referred to can be classified as speech because information about phonetic structure is spread redundantly across the speech spectrum and over time; thus much of it is present in these signals even though the superficial acoustic marks of speech may be absent. On the other hand, isolated formant transitions, which have the appropriate acoustic marks but, out of context, no definite phonetic structure, are, as we have said, classified as nonspeech. In short, the signal is speech if and only if the pattern of articulatory gestures that must have produced it can be reconstructed. We call this property "generative detection," having in mind the analogous situation in the domain of sentence processing. There superficial features cannot distinguish grammatical sentences from ungrammatical ones. The only way to determine the grammaticality of a sentence is to parse it—that is, to try to regenerate the syntactic structure intended by the speaker.

Is generative detection found in the specialized systems of other species? Consider first the mustached bat, whose echolocation system relies on biosonar signals (Suga, 1984, and this volume). The bat has to be able to distinguish its own echolocation signals from the similar signals of conspecifics. Otherwise, not only would the processing of its own signals be jammed, but many of the objects it located would be illusory, because it would have subjected the conspecific signals to the same heteromorphic treatment it gives its own. According to Suga, the bat probably solves the problem in the following way. The harmonics of all the biosonar signals reach the CF-CF and FM-FM neurons that determine the delay between harmonics $F_2$ and $F_3$ of the emitted signals and their respective echoes. But these neurons operate only if $F_1$ is also present. This harmonic is available to the cochlea of the emitting bat by bone conduction but is weak or absent in the radiated signal. Thus the output of the CF-CF and FM-FM neurons reflects only the individual's own signals and not those of conspecifics. The point is that, as in the case of human speech detection, there is no preliminary sorting of the two classes of signals. Detection of the required signal is not a separate stage but inherent in the signal analysis. However, the bat's method of signal detection cannot properly be called generative, because, unlike speech detection, it relies on a surface property of the input signal.

Generative detection is, perhaps, more likely to be found in the perception of song by birds. While, so far as we are aware, no one has suggested how song detection might work, it is known about the zebra finch that pure tones as well as actual song produce activity in the neurons of the song motor nucleus HVc (Williams, 1984; Williams and Nottebohm, 1985), a finding that argues against preliminary sorting and for detection in the course of signal analysis. Moreover, since the research just cited also provides evidence that the perception of song by the zebra finch is motoric, generative detection must be considered a possibility until and unless some superficial acoustic characteristic of a particular song is identified that would suffice to distinguish it from the songs of other avian species. Generative detection in birds seems the more likely, given that some species—the winter wren, for example—have hundreds of songs that a conspecific

can apparently recognize correctly even if it has never heard them before (Konishi, 1985). It is therefore tempting to speculate that the wren has a grammar that generates possible song patterns, and that the detection and parsing of conspecific songs are parts of the same perceptual process.

While generative detection may not be a very widespread property of specialized perceiving systems, what does seem to be generally true is that these systems do their own signal detection. Moreover, they do it by virtue of features that are also exploited in signal analysis, whether these features are simple superficial characteristics of the signal, as in the case of echolocation in the bat, or complex reflections of distal events, as in the case of speech perception. This more general property might be added to those that Fodor (1983) has identified as common to all perceptual modules.

**Preemptiveness**

As we have already hinted, our proposal that there are no preliminary sorting mechanisms leads to a difficulty, for without such a mechanism, we might expect that the general-purpose, homomorphic auditory systems, being sensitive to the same dimensions of an acoustic signal as a specialized system, would also process special signals. This would mean that the bat would not only use its own biosonar signals for echolocation, but would also hear them as it presumably must hear the similar biosonar signals of other bats; the zebra finch would perceive conspecific song not only as song but also as an ordinary environmental sound; and human beings would hear chirps and glissandi as well as speech. We cannot be sure with nonhuman animals that such double processing of special-purpose signals does not in fact occur, but certainly it does not for speech, except under the extraordinary and throroughly unecological conditions, described earlier, that induce "duplex" perception. We should suppose, however, that except where complementary aspects of the same distal object or event are involved, as in the perception of color and shape, double processing would be maladaptive, for it would result in the perception of two distal events, one of which would be irrelevant or spurious. For example, almost any environmental sound may startle a bird, so if a conspecific song were perceived as if it were also something else, the listening bird might well be startled by it.

The general-purpose homomorphic systems themselves can have no way of defining the signals they should process in a way that excludes special signals, since the resulting set of signals would obviously not be a natural class. But suppose that the specialized systems are somehow able to preempt signal information relevant to the events that concern them, preventing it from reaching the general-purpose systems at all. The bat would then use its own biosonar signals to perceive the distal objects of its environment but would not also hear them as it does the signals of other bats; the zebra finch would hear song only as song; and human beings would hear speech as speech but not also as nonspeech.

An arrangement that would enable the preemptiveness of special-purpose systems is serial processing, with the specialized system preceding the general-purpose systems (Mattingly and Liberman, 1985). The specialized system would not only detect and process the signal information it requires but would also provide an input to the general-purpose systems from which this information had been removed. In the case of the mustached bat, the mechanism proposed

by Suga (1984) for the detection of the bat's own biosonar signals would also be sufficient to explain how the information in these signals, but not the similar information in the conspecific signals, could be kept from the general-purpose system. Though doubtless more complicated, the arrangements in humans for isolating phonetic information and passing on nonphonetic information would have the same basic organization. We suggest that the speech perceiving system not only recovers whatever phonetic structure it can but also filters out those features of the signal that result from phonetic structure, passing on to the general-purpose systems all of the phonetically irrelevant residue. If the input signal includes no speech, the residue will represent all of the input. If the input signal includes speech as well as nonspeech, the residue will represent all of the input that was not speech plus the laryngeal source signal (as modified by the effects of radiation from the head), the pattern of formant trajectories that results from the changing configuration of the vocal tract having been removed. Thus the perception not only of nonspeech environmental sounds but also of nonphonetic aspects of the speech signal, such as voice quality, is left to the general-purpose systems.

Serial processing appeals to us for three reasons. First, it is parsimonious. It accounts for the fact that speech is not also perceived as nonspeech, without assuming an additional mechanism and without complicating whatever account we may eventually be able to offer of speech perception itself. The same computations that are required to recover phonetic structure from the signal also suffice to remove all evidence of it from the signal information received by the general-purpose system.

Second, by placing the speech processing system ahead of the general-purpose systems, the hypothesis exploits the fact that while nonspeech signals have no specific defining properties at all, speech signals form a natural class, with specific, though deep, properties by virtue of which they can be reliably assigned to the class.

Third, serial processing permits us to understand how precedence can be guaranteed for a class of signals that has special biological significance. It is a matter of common experience that the sounds of bells, radiators, household appliances, and railroad trains can be mistaken for speech by the casual listener. On the other hand, mistaking a speech sound for an ordinary environmental sound is comparatively rare. This is just what we should expect on ethological grounds, for, as with other biologically significant signals, it is adaptive that the organism should put up with occasional false alarms rather than risk missing a genuine message. Now if speech perception were simply one more cognitive operation on auditory primitives, or if perception of nonspeech preceded it, the organism would have to learn to favor speech, and the degree of precedence would depend very much on its experience with acoustic signals generally. But if, as we suggest, speech precedes the general-purpose system, the system for perceiving speech need only be reasonably permissive as to which signals it processes completely for the precedence of speech to be insured.

### Commonality Between the Specializations for Perception and Production

So far we have been concerned primarily with speech perception, and we have argued that it is controlled by a system specialized to perceive phonetic gestures.

But what of the system that controls the gestures? Is it specialized, too, and how does the answer to that question bear on the relation between perception and production?

A preliminary observation is that there is no logical necessity for speech production to be specialized merely because speech perception appears to be. Indeed, our commitment to an account of speech perception in which the invariants are motoric deprives us of an obvious argument for the specialness of production. For if the perceptual invariants were taken to be generally auditory, it would be easy to maintain that only a specialized motoric system could account for the ability of every normal human being to speak rapidly and yet to manipulate the articulators so as to produce just those acoustically invariant signals that the invariant auditory percepts would require. But if the invariants are motoric, as we claim, it could be that the articulators do not behave in speech production very differently from the way they do in their other functions. In that case, there would be nothing special about speech production, though a perceptual specialization might nevertheless have been necessary to deal with the complexity of the relation between articulatory configuration and acoustic signal. However, the perceptual system would then have been adapted very broadly to the acoustic consequences of the great variety of movements that are made in chewing, swallowing, moving food around in the mouth, whistling, licking the lips, and so on. There would have been few constraints to aid the perceptual system in recovering the gestures, and nothing to mark the result of its processing as belonging to an easily specifiable class of uniquely phonetic events. However, several facts about speech production strongly suggest that it is, instead, a specialized and highly constrained process.

It is relevant, first, that the inventory of gestures executed by a particular articulator in speech production is severely limited, both with respect to manner of articulation (i.e., the style of movement of the gesture) and place of articulation (i.e., the particular fixed surface of the vocal tract that is the apparent target of the gesture). Consider, for example, the tip of the tongue, which moves more or less independently of, but relative to, the tongue body. In nonphonetic movements of this articulator, there are wide variations in speed, style, and direction, variations that musicians, for example, learn to exploit. In speech, however, the gestures of the tongue tip, though it is perhaps the most phonetically versatile of the articulators, are restricted to a small number of manner categories: stops (e.g., [t] in *too*), flaps ([D] in *butter*), trills ([r] in spanish *perro*), taps ([ɾ] in Spanish *pero*), fricatives ([θ] in *thigh*), median approximants ([ɹ] in *red*) and lateral approximants ([l] in *law*). Place of articulation for these gestures is also highly constrained, being limited to dental, alveolar, and immediately postalveolar surfaces (Ladefoged, 1971; Catford, 1977). These restricted movements of the tongue tip in speech are not, in general, similar to those it executes in nonphonetic functions (though perhaps one could argue for a similarity between the articulation of the interdental fricative and the tongue-tip movement required to expel a grape seed from the mouth. But, as Sapir (1925) observed about the similarity between an aspirated [w] and the blowing out of a candle, these are "norms or types of entirely distinct series of variants"). Speech movements are, for the most part, peculiar to speech; they have no obvious nonspeech functions.

The peculiarity of phonetic gestures is further demonstrated in consequence of the fact that, in most cases, a gesture involves more than one articulator.

Thus the gestures we have just described, though nominally attributed to the tongue tip, actually require also the cooperation of the tongue body and the jaw to ensure that the tip will be within easy striking distance of its target surface (Lindblom, 1983). The requirement arises because, owing to other demands on the tongue body and jaw, the tongue tip cannot be assumed to occupy a particular absolute rest position at the time a gesture is initiated. Cooperation between the articulators is also required in such nonphonetic gestures as swallowing, but the particular cooperative patterns of movement observed in speech are apparently unique, even though there may be nonspeech analogues for one or another of the components of such a pattern.

Observations analogous to those just made about the tongue tip could be made with respect to each of the other major articulators: the tongue body, the lips, the velum, and the larynx. That the phonetic gestures possible for each of these articulators form a very limited set that is drawn upon by all languages in the world has often been taken as evidence for a universal phonetics (e.g., Chomsky and Halle, 1968). (Indeed, if the gestures were not thus limited, a general notation for phonetic transcription would hardly be possible.) That the gestures are eccentric when considered in comparison with what the articulators are generally capable of is evidence that speech production does not merely exploit general tendencies for articulator movement but depends, rather, on a system of controls specialized for language.

A further indication of the specialness of speech production is that certain of the limited and eccentric set of gestures executed by the tongue tip are paralleled by gestures executed by other major articulators. Thus stops and fricatives can be produced not only by the tongue tip but also by the tongue blade, the tongue body, the lips, and the larynx, even though these various articulators are anatomically and physiologically very different from one another. Nor, to forestall an obvious objection, are these manner categories mere artifacts of the phonetician's taxonomy. They are truly natural classes that play a central role in the phonologies of the world's languages. If these categories were unreal, we should not find that in language *x* vowels always lengthen before all fricatives, that in language *y* all stops are regularly deleted after fricatives, or that in all languages the constraints on the sequences of sounds in a syllable are most readily described according to manner of articulation (Jespersen, 1920). And when the sound system of a language changes, the change is frequently a matter of systematically replacing sounds of one manner class by sounds of another manner class produced by the same articulators. Thus the Indo-European stops [p],[t],[k],[q] were replaced in Primitive Germanic by the corresponding fricatives [f],[θ],[x],[X] ("Grimm's law").

Our final argument for the specialness of speech production depends on the fact of gestural overlap. Thus in the syllable [du], the tongue-tip closure gesture for [d] overlaps the lip-rounding and tongue-body-backing gestures for [u]. Even more remarkably, two gestures made by the same articulator may overlap. Thus in the syllable [gi], the tongue-body-closure gesture for [g] overlaps the tongue-body-fronting gesture for [i], so that the [g] closure occurs at a more forward point on the palate than would be the case for [g] in [gu]. As we have already suggested, it is gestural overlap, making possible relatively high rates of information transmission, that gives speech its adaptive value as a communication system. But if the strategy of overlapping gestures to gain speed is not to defeat itself,

the gestures can hardly be allowed to overlap haphazardly. If there were no constraints on how the overlap could occur, the acoustic consequences of one gesture could mask the consequences of another. In a word such as *twin*, for instance, the silence resulting from the closure for the stop [t] could obscure the sound of the approximant [w]. Such accidents do not ordinarily occur in speech, because the gestures are apparently phased so to provide the maximum amount of overlap consistent with preservation of the acoustic information that specifies either of the gestures (Mattingly, 1981). This phasing is most strictly controlled at the beginnings and ends of syllables, where gestural overlap is greatest, and most variable in the center of the syllable, where less is going on (Tuller and Kelso, 1984). Thus, to borrow Fujimura's (1981) metaphor, the gestural timing patterns of consonants and consonant clusters are icebergs floating on a vocalic sea. Like the individual gestures themselves, these complex temporal patterns are peculiar to speech and could serve no other ecological purpose.

We would conclude, then, that speech production is specialized, just as speech perception is. But if this is so, we would argue further that these two processes are not two systems, but, rather, modes of one and the same system. The premise of our argument is that because speech has a communicative function, what counts as phonetic structure for production must be the same as what counts as phonetic structure for perception. This truism holds regardless of what one takes phonetic structure to be, and any account of phonetic process has to be consistent with it. Thus, on the conventional account, it must be assumed that perception and production, being taken as distinct processes, are both guided by some cognitive representation of the structures that they deal with in common. On our account, however, no such cognitive representation can be assumed if the notion of a specialized system is not to be utterly trivialized. But if we are to do without cognitive mediation, what is to guarantee that at every stage of ontogenetic (and for that matter phylogenetic) development, the two systems will have identical definitions of phonetic structure? The only possibility is that they are directly linked. This, however, is tantamount to saying that they constitute a single system, in which we would expect representations and computational machinery not to be duplicated, but to coincide insofar as the asymmetry of the two modes permits.

To make this view more concrete, suppose, as we have elsewhere suggested (Mattingly and Liberman, 1969; Liberman et al., 1972; Liberman and Mattingly, 1986), that the speech production–perception system is, in effect, an articulatory synthesizer. In the production mode, the input to the synthesizer is some particular, abstractly specified gestural pattern from which the synthesizer computes a representation of the contextually varying articulatory movements that will be required to realize the gestures, and from this articulatory representation, the muscle commands that will execute the actual movements, some form of "analysis by synthesis" being obviously required. In the perceptual mode, the input is the acoustic signal from which the synthesizer computes—again by analysis by synthesis—the articulatory movements that could have produced the signal, and from this articulatory representation, the intended gestural pattern. The computation of the muscle commands from articulatory movement is peculiar to production, and the computation of articulatory movement from the signal is peculiar to perception. What is common to the two modes, and carried out

by the same computations, is the working out of the relation between abstract gestural pattern and the corresponding articulatory movements.

We earlier alluded to a commonality between modes of another sort when we referred to the finding that the barn owl's auditory orientation processes use the same neural map as its visual orientation processes do. Now we would remark the further finding that this arrangement is quite one-sided: The neural map is laid out optically, so that sounds from sources in the center of the owl's visual field are more precisely located and more extensively represented on the map than are sounds from sources at the edges (Knudsen, 1984). This is of special relevance to our concerns, because, as we have several times implied, a similar one-sidedness seems to characterize the speech specialization: Its communal arrangements are organized primarily with reference to the processes of production. We assume the dominance of production over perception because it was the ability of appropriately coordinated gestures to convey phonetic structures efficiently that determined their use as the invariant elements of speech. Thus it must have been the gestures, and especially the processes associated with their expression, that shaped the development of a system specialized to perceive them.

More comparable, perhaps, to the commonality we see in the speech specialization are examples of commonality between perception and production in animal communication systems. Evidence for such commonality has been found for the tree frog (Gerhardt, 1978); the cricket (Hoy and Paul, 1973; Hoy et al., 1977); the zebra finch (Williams, 1984; Williams and Nottebohm, 1985); the white-crowned sparrow (Margoliash, 1983); and the canary (McCasland and Konishi, 1983). Even if there were no such evidence, however, few students of animal communication would regard as sufficiently parsimonious the only alternative to commonality: that perception and production are mediated by cognitive representations. But if we reject this alternative in explaining the natural modes of nonhuman communication, it behooves us to be equally conservative in our attempt to explain language, the natural mode of communication in human beings. Just because language is central to so much that is uniquely human, we should not therefore assume that its underlying processes are necessarily cognitive.

## THE SPEECH SPECIALIZATION AND THE SENTENCE

As a coda, we here consider, though only briefly, how our observations about perception of phonetic structure might bear more broadly on perception of sentences. Recalling first the conventional view of speech perception—that it is accomplished by processes of a generally auditory sort—we find its extension to sentence perception in the assumption that coping with syntax depends on a general faculty too. Of course, this faculty is taken to be cognitive, not auditory, but like the auditory faculty, it is supposed to be broader than the behavior it serves. Thus it presumably underlies not just syntax but all the apparently smart things people do. For an empiricist, this general faculty is a powerful ability to learn, and so to discover the syntax by induction. For a nativist, it is an intelligence that knows what to look for because syntax is a reflection of how the mind works. For both, perceiving syntax has nothing in common with perception of

speech, or, a fortiori, with perception of other sounds, whether biologically significant or not. It is as if language, in its development, had simply appropriated auditory and cognitive processes that are themselves quite independent of language and, indeed, of each other.

The parallel in syntax to our view of speech is the assumption that sentence structures, no less than speech, are dealt with by processes narrowly specialized for the purpose. On this assumption, syntactic and phonetic specializations are related to each other as two components of the larger specialization for language. We should suppose then that the syntactic specialization might have important properties in common, not only with the phonetic specialization but also with the specializations for biologically significant sounds that occupy the contributors to this volume.

## ACKNOWLEDGMENTS

The writing of this chapter was supported by a grant to Haskins Laboratories (NIH-NICHD-HD-01994). We are grateful to Harriet Magen and Nancy O'Brien for their help with references and to Alice Dadourian for invaluable editorial assistance and advice. We received shrewd comments and suggestions from Carol Fowler, Masakazu Konishi, Eric Knudsen, Daniel Margoliash, Bruno Repp, Michael Studdert-Kennedy, Nobuo Suga, and Douglas Whalen. Some of these people have views very different from those expressed here, but we value their criticisms all the more for that.

## REFERENCES

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

Catford, J. C. (1977) *Fundamental Problems in Phonetics*, Indiana Univ. Press, Bloomington.

Chomsky, N., and M. Halle (1968) *The Sound Pattern of English*, Harper & Row, New York.

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

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

Fujimura, O. (1981) Temporal organization of speech as a multi-dimensional structure. *Phonetica* **38**:66–83.

Gerhardt, H. C. (1978) Temperature coupling in the vocal communication system of the gray tree frog *Hyla versicolor*. *Science* **199**:992–994.

Green, K. P., and J. L. Miller (1985) On the role of visual rate information in phonetic perception. *Percept. Psychophys.* **38**:269–276.

Hafter, E. R. (1984) Spatial hearing and the duplex therory: How viable is the model? In *Dynamic Aspects of Neocortical Function*, G. M. Edelman, W. E. Gall, and W. M. Cowan, eds., pp. 425–448, Wiley, New York.

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

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

Jespersen, O. (1920) *Lehrbuch der Phonetik*, Teubner, Leipzig.

Knudsen, E. I. (1984) Synthesis of a neural map of auditory space in the owl. In *Dynamic Aspects of Neocortical Function*, G. M. Edelman, W. E. Gall, and W. M. Cowan, eds., pp. 375–396, Wiley, New York.

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

Konishi, M. (1985) Birdsong: From behavior to neuron. *Annu. Rev. Neurosci.* **8**:125–170.

Ladefoged, P. (1971) *Preliminaries to Linguistic Phonetics*, Univ. Chicago Press, Chicago.

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 IXth International Congress of Phonetic Sciences*, Vol. 2, E. Fischer-Jørgensen, J. Rischel, and N. Thorsen, eds., pp. 468–473, Univ. Copenhagen, Copenhagen.

Liberman, A. M., and I. G. Mattingly (1986) 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.

Lindblom, B. (1983) Economy of speech gestures. In *The Production of Speech*, P. MacNeilage, ed., pp. 217–245, Springer, New York.

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

Margoliash, D. (1983) Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *J. Neurosci.* **3**:1039–1057.

Mattingly, I. G. (1981) Phonetic representation and speech synthesis by rule. In *The Cognitive Representation of Speech*, T. Myers, J. Laver, and J. Anderson, eds., pp. 415–420, North-Holland, Amsterdam.

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, New York.

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

McCasland, J. S., and M. Konishi (1983) Interaction between auditory and motor activities in an avian song control nucleus. *Proc. Natl. Acad. Sci. USA* **78**:7815–7819.

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

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

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

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

Sapir, E. (1925) Sound patterns in language. *Language* **1**:37–51. Reprinted in *Language, Culture and Personality: Selected Writings of Edward Sapir*, D. G. Mandelbaum, ed., pp. 33–45, Univ. California Press, Berkeley, 1963.

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.

Summerfield, Q. (1979) Use of visual information for phonetic perception. *Phonetica* **36**:314–331.

Tuller, B., and J. A. S. Kelso (1984) The relative timing of articulatory gestures: Evidence for relational invariants. *J. Acoust. Soc. Am.* **76**:1030–1036.

Williams, H. (1984) *A Motor Theory of Bird Song Perception*, unpublished doctoral dissertation, Rockefeller University, New York.

Williams, H., and F. N. Nottebohm (1985) Auditory responses in avian vocal motor neurons: A motor theory for song perception in birds. *Science* **229**:279–282.

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.