Reprinted from AMERICAN PSYCHOLOGIST, Vol. 37, No. 2, February 1982
Printed in U. S. A.

365

On Finding That Speech Is Special

ALVIN M. LIBERMAN *Haskins Laboratories, New Haven, University of Connecticut, and Yale University*

ABSTRACT: *A largely unsuccessful attempt to communicate phonologic segments by sounds other than speech led my colleagues and me to ask why speech does it so well. The answer came the more slowly because we were wedded to a "horizontal" view of language, seeing it as a biologically arbitrary assemblage of processes that are not themselves linguistic. Accordingly, we expected to find the answer in general processes of auditory perception, to which the acoustic signal had been made to conform by appropriate regulation of the movements of articulation. What we found was the opposite: specialized processes of phonetic perception that had been made to conform to the acoustic consequences of the way articulatory movements are regulated. The distinctively linguistic function of these specializations is to provide for efficient perception of phonetic structures that can also be efficiently produced. To assume that a phonetic specialization exists accords well with a "vertical" view of language, in which the underlying activities are seen as coherent and distinctive. Recent evidence for such special processes comes from experiments designed to investigate the integration of cues.*

I welcome this opportunity to talk about a subject that has been too much taken for granted. The subject is perception of phonetic segments, the consonants and vowels that lie near the surface of language. My aim is to promote the hypothesis that perception of those segments rests on specialized processes. These support a phonetic mode of perception; they serve a distinctively linguistic function; and they are part of the larger specialization for language.

The phonetic specialization is apparently adapted to the singular code by which phonetic structure is connected to sound, a code that owes its character to the way the segments of the structure are articulated and coarticulated by the organs of the vocal tract. Not surprisingly, then, phonetic processes incorporate a link between perception and production. With that as key, an otherwise opaque code becomes perfectly transparent: The diverse, continuous, and tangled sounds of speech are automatically perceived as a scant handful of discrete and variously ordered segments. Moreover, the segments are given in perception as distinctively phonetic objects, without the encumbering auditory baggage that would make them all but useless for their proper role as vehicles of language.

But we do take speech and its acoustic nature for granted, so much so that it is hard to see why perception of phonetic segments should require processes of an other-than-auditory sort and even harder, perhaps, to imagine what it might mean to perceive those segments as phonetic objects, free of a weighty burden of auditory particulars. It may help, then, to begin by recounting my experience with an attempt to transmit phonologic information by purely auditory means. That experience exposed the problem that a phonetic specialization might solve, though it did not, of course, reveal how the solution is achieved, nor did it show that the solution requires specialized processes. Evidence bearing on those matters is reserved for later sections.

Perceiving Phonologic Segments in the Auditory Mode: An Assumption That Failed

In the mid-1940s I began, together with colleagues at Haskins Laboratories, to design a reading ma-

This paper is based on a Distinguished Scientific Contribution Award address given at the meeting of the American Psychological Association, Los Angeles, August 25, 1981. Preparation of the paper, and much of the research on which it is based, was supported by National Institute of Child Health and Human Development Grant HD 01994 and by Biomedical Research Support Grant RR05596.

At Haskins Laboratories it is hard to know what is owed and to whom. I would, however, especially acknowledge my debt to Franklin S. Cooper, with whom I have been closely associated for 35 years. For help with this paper, I thank Louis Goldstein, Isabelle Liberman, Virginia Mann, Sharon Manuel, Ignatius Mattingly, Patrick Nye, Bruno Repp, and Michael Studdert-Kennedy.

I am grateful to J. A. Fodor for making available to me an early draft of his monograph, "The Modularity of Mind," which I found particularly relevant and stimulating.

Requests for reprints should be sent to Alvin M. Liberman, c/o Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510.

0003-066X/82/3702-0148$00.75

chine for the blind (Cooper, 1950; Nye, 1963; Studdert-Kennedy & Cooper, 1966). This was, or was to have been, a device that would scan print and use its contours to control an acoustic signal. At the outset we assumed that our machine had only to produce, for each letter, a pattern of sound that was distinctively different from the patterns for other letters. Blind users would presumably learn to associate the sounds with the letters and thus, in time, come to read. The rationale, largely unspoken, was an assumption about the nature of speech—to wit, that the sounds of speech represent the phonemes (roughly, the letters of the alphabet) in a straightforward way, one segment of sound for each phoneme. Accordingly, the perception of speech was thought to be no different from the perception of other sounds, except that there was in speech a learned association between perceived sound and the name of the corresponding phoneme. Why not expect, then, that arbitrary but distinctive sounds would serve as well as speech, provided only that the users had sufficient training?

Given that expectation we were ill prepared for the disappointing performance of the nonspeech signals our early machines produced. So we persisted, seeking to increase the perceptual distinctiveness of the sound alphabet and the ease with which its units would form into words and sentences. But our best efforts were unavailing. No matter how we patterned them, the sounds evoked a clutter of auditory detail that subjects could not readily organize and identify. This discouraged the subjects, but not me, for I had faith that the difficulty would ultimately yield to practice and the principles of learning. What loomed as a far more serious failing was that modest increases in rate caused the unit sounds to dissolve into an imperspicuous buzz. Indeed, this happened at rates barely one tenth those at which the discrete units of phonetic structure can be conveyed by speech.

Having thus come to the conclusion that we should try to learn from speech, we began to study it. But our hope at that early stage was only that we might find principles of auditory perception, hitherto unnoticed, that the language system had somehow managed to exploit.[1] These would not only be interesting in their own right but also useful in enabling us to overcome the practical difficulty we had been having, since the auditory principles we hoped to find could presumably be applied to the design of nonspeech sounds that our reading machine might be made to produce.

What I did not for a long time understand was that our practical difficulty lay not in our having failed to find the right principles of auditory perception but, much deeper, in our having failed to see that the principles we sought were simply not auditory. Perhaps I would have arrived at that understanding earlier had I not been in the grip of a misleading assumption that had decisively shaped my thinking about speech, language, and, indeed, almost anything else I might have found psychologically interesting. I was the more misled because the assumption reflected what I took to be the received view; in any case I had never thought to question it.

In casting about for a word to characterize the view I speak of, I hit on *horizontal* as being particularly appropriate, only to discover that J. A. Fodor (Note 1) had chosen the same word to describe what I take to be much the same view. Apparently, we have here a metaphor whose time has come. As applied to language, the metaphor is intended to convey that the underlying processes are arranged in layers, none of them specific to language. On that horizontal orientation, language is accounted for by reference to whatever combination of processes it happens to engage. Hence our assumption, in the attempt to find a substitute for speech, that perception of phonologic segments is normally accomplished, presumably in the first layer, by processes of a generally auditory sort—that is, by processes no different from those that bring us the rustle of leaves in the wind or the rattle of a snake in the grass. To the extent that we were concerned with the rest of language, we must have supposed in like manner that syntactic structures are managed by using the most general resources of cognition or intelligence. There were surely other processes on our minds when we thought about language—attention, memory, learning, for example—the exact number and variety depending on just which aspects of language activity our attention was directed to at the moment. But all the processes we might have invoked had in common that none was specialized for language. We were not prepared to give language a biology of its own, but only to treat it as an epiphenomenon, a biologically arbitrary assemblage of processes that were not themselves linguistic.

[1] At one point we assumed that these principles were so general as to extend to perception in all modalities. Indeed, we carried out experiments designed to explore the possibility that patterns could be preserved across vision and audition, provided the stimulus coordinates were properly transformed (Cooper, Liberman, & Borst, 1951).

The opposite view—the one to which I now incline—is, by contrast, vertical. Seen this way, language does have its own biology. It is a coherent system, like echolocation in the bat, comprising distinctive processes adapted to a distinctive function. The distinctive processes are those that underlie the grammatical codes of syntax and phonology; their distinctive function is to overcome the limitations of communicating by agrammatic means. To appreciate those limitations we need only consider how little we could say if, as in an agrammatic system, there were a straightforward relation between message and signal, one signal, however elaborately patterned, for each message. In such a system the number of messages to be communicated could be no greater than the number of holistically and distinctively different signals that can be efficiently produced and perceived; and surely that number is very small, especially when the signal is acoustic. What the processes of syntax and phonology do for us, then, is encode an unlimited number of messages into a very limited number of signals. In so doing they match our message-generating capabilities to the restricted resources of our signal-producing vocal tracts and our signal-perceiving ears. As for the phonetic part of the phonologic domain, which is the subject of this paper, I suggest that it, too, partakes of the distinctive function of grammatical codes and that it is, accordingly, also special. (For further discussion, see Liberman, 1970; Liberman & Studdert-Kennedy, 1978; Mattingly & Liberman, 1969.)

The Special Function of the Phonetic Mode

To produce a large, indeed an infinite, number of messages with a small number of signals, a syntax would, in principle, suffice. Without a phonology, however, each smallest unit of an utterance would necessarily be a word, so a talker would have to make do with a very small vocabulary. The obvious function of the phonologic domain is, then, to construct words out of a few meaningless units and, thus, to make possible the large vocabularies that human beings like to deploy. But the words of the vocabulary are presumed to be found in the deeper reaches of the phonology, where they are represented by the abstract phonemes that stand beneath the many phonetic variations at the surface, variations associated with phonetic context, word boundaries, rate of articulation, lexical stress, phrasal stress, idiolect, and dialect, to name the most obvious sources. What remains in speaking is, of course, to derive the surface phonetic structures and, then, to transmit them by using the organs of articulation to produce and modify sounds. Transmitting those structures as sounds and at high rates becomes the distinctive function of the phonetic mode.

At average rates of speaking, talkers produce and listeners perceive about eight to ten segments per second. In the extreme the rate may go to 25 or 30 per second, at least for short stretches. Plainly such rates would be impossible if each segment were represented, as in the acoustic alphabets of our early reading machines, by a segment of sound. The organs of the vocal tract cannot make unit gestures that fast, and even if they could, the rate of delivery of the resulting units of sound would overreach the temporal resolving power of the ear. The trick, then, is to evade the limitations on the rate at which discrete segments of sound can be transmitted and perceived while yet preserving the discrete phonetic segments those sounds must convey.

The vocal tract solves its part of the problem by breaking the two or three dozen phonetic segments into a smaller number of features, assigning each feature to a gesture that can be made more or less independently, and then turning the articulators loose, as it were, to do what they can. A consequence is that gestures corresponding to features of successive segments are produced at the same time, or else greatly overlapped, according to the constraints and possibilities inherent in the masses to be moved and in the neuromuscular arrangements that move them. This is to say that the character of speech is determined largely by the nature of the mechanisms that do the speaking. But it could hardly be otherwise. For even if nature had devised articulators that could make successive unit gestures at rapid rates—putting aside that this would presumably have destroyed the utility of the vocal tract for such other purposes as eating and breathing—the resulting drumfire of sound would, as noted earlier, defeat the ear. At all events the nature of the articulatory process produces a relation between phonetic segment and sound—the singular code referred to in the introduction—that must take first place in any attempt to investigate and understand the perception of speech.

One characteristic of the code that should immediately engage our attention follows from the fact that one or another of the articulators is almost always moving. The consequence is that many, perhaps most, of the potential acoustic cues—that

is, aspects of the sound that bear a systematic relation to the phonetic segment—are of a dynamic sort. Witness, for example, the changes in formant frequency caused by the movement from one articulatory position to another and known to be important cues for various consonants—and, indeed, for vowels (Liberman, Delattre, Cooper, & Gerstman, 1954; Mann & Repp, 1980; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957; Strange, Jenkins, & Edman, 1977). How do these time-varying acoustic cues evoke discrete and unitary phonetic percepts that have no corresponding time-varying quality?

Another characteristic of the code, owing again to the way the articulators produce and modulate the sound, is that the acoustic cues are numerous and diverse. In the contrast between the [b] of *rabid* and the [p] of *rapid*, for example, Lisker (1978) has so far identified 16 cues, representing a variety of acoustic types. The many cues are not ordinarily of equal power—some will override others—but power does not appear to be determined primarily by acoustic prominence. How, then, is such a numerous variety of seemingly arbitrary cues bound into a single phonetic percept?

Finally, the processes of articulation and, more particularly, coarticulation cause the potential cues for a phonetic segment to be widely distributed through the signal and merged, often quite thoroughly, with potential cues for other segments. In a syllable like *bag*, to take a simple case, it is likely that a single parameter of the acoustic signal—say the second formant—simultaneously carries information about at least two of the constituent segments and, in some places, all three (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Liberman, 1974). Indeed, it is this characteristic of speech, this encoding of several phonetic segments into one segment of sound, that is, as we have seen, an essential aspect of the processes by which phonetic segments are produced and perceived at high rates. But the result is an acoustic amalgam, not an alphabet. How does the listener recover the string of discrete phonetic segments it encodes?

Of course we might try to evade those questions, and the thorny problems they pose for the auditory mode, by supposing that the articulators produce for each phonetic segment at least one cue that represents that segment quite straightforwardly (Stevens & Blumstein, 1981). Because the relation of that cue to the phonetic segment is transparent to ordinary auditory processes, the listener might respond most attentively just to it, dismissing the others as so much chaff or else learning to accept them as associated with, but wholly incidental to, the real business of talker and listener. Such evasion will be hard to maintain, however, if as we now have reason to think, the typical listener is sensitive to all of the phonetic information in speech sounds (Bailey & Summerfield, 1980).[2] Certainly every potential cue so far tested has proved to be an actual cue, no matter how peculiar seeming its relation to the phonetic segment.

We should suppose, then, that there is in speech perception a process by which the manifold of variously merged, continuous, and time-varying cues is made to form in the listener's mind the discrete and ordered phonetic segments that were produced by the speaker. But it seems hardly conceivable that this could be accomplished by processes of a generally auditory sort. Therefore, I assume, as proposed in the introduction, that the process is a special one—a distinctively phonetic process specifically adapted to the unique characteristics of the speech code. Since that code is opaque except as one understands the special way it comes about, I find it plausible to suppose further that a link between perception and production constrains the process as if by knowledge of what a vocal tract does when it makes linguistically significant gestures (Cooper et al., 1952; Liberman, Delattre, & Cooper, 1952).

A Special Process of the Phonetic Mode: Integration of Cues

Of the many experimental results that bear on the existence and nature of distinctively phonetic processes, none is critical; what tells is the weight of the evidence and the way it converges on certain conclusions. Thus faced with many more results than I could hope to include, I had to choose between picking a closely related few and, alternatively, offering a token of each type. (For recent and comprehensive reviews, see Repp, in press;

[2] In contrast to the remarkable sensitivity of the phonetic mode to all aspects of the acoustic signal that convey phonetic information, there is its equally remarkable insensitivity to those aspects of the signal that do not. Thus, as is well-known from many years of research on synthetic speech, the phonetic component of the percept is usually unaffected by gross variations in those aspects of the signal—for example, bandwidth of the formants—that are beyond the control of the articulatory apparatus and hence necessarily irrelevant for all linguistic purposes (Liberman & Cooper, 1972; Remez, Rubin, Pisoni, & Carrell, 1981). The only effect of such variations is to make the speech sound unnatural or, in the most extreme cases, to make it impossible for the listener to perceive the sound as speech.

Studdert-Kennedy, 1980.) I have chosen the related few, selecting them from recent studies that bear on the three questions raised by the characteristics of the speech code I referred to in the previous section. Aspects of these questions have long been worried about as the problem of segmentation: How is the acoustic signal divided into phonetic segments (Cooper et al., 1952; Fant, 1962; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967)? Recently Repp (1978) and Oden and Massaro (1978) have looked at the other side of the coin, putting attention on the problem of integration: How do cues combine to produce the percept? It suits my purposes to adopt their perspective, and so I will.

INTEGRATION OF A TIME-VARYING SOUND

Frequency sweeps—called *formant transitions*—of the kind shown in Figure 1 can be sufficient cues for the perceived distinction between the stop consonants [d] and [g] in the syllables [da] and [ga] (Harris, Hoffman, Liberman, & Delattre, 1958). But as I asked earlier, how are such frequency sweeps integrated (as information about the phonetic dimension of place) into a unitary percept, [d] or [g], that has about it no hint of a corresponding sweep in pitch? Two interpretations are possible: one, that the integration is accomplished by ordinary auditory processes; the other, that special phonetic processes come into play.

On an auditory interpretation one might suppose, most simply, that this is an instance of low-level sensory integration, something like the well-known integration of intensity and time into the perception of loudness. That possibility is quickly ruled out, however, by the observation that when the transition cues are removed from the pattern and presented alone, as in the part of the figure at lower right, listeners do perceive a rising or falling chirp, almost a glissando, that conforms reasonably to the time-varying percept that psychoacoustic considerations might have led us to expect (Mattingly, Liberman, Syrdal, & Halwes, 1971).

But the auditory theory is not so easily disposed of; it can always fall back on the assumption that the formant transitions collaborate with the rest of the pattern in an interaction of a purely auditory sort, from which the percepts [d] or [g] emerge. It matters little that there is nothing in what we know about perception of complex sounds to suggest that such interaction should occur, for we know very little about perception of complex sounds. Nor does it necessarily matter how implausible it is to suppose that the articulators could so comport themselves as to produce exactly the right combination of sounds, not just in this instance, but in the myriad others that must occur as the articulators accommodate to variations in, for example, phonetic context, rate, and linguistic stress. Such considerations make an explanation based on auditory interaction endlessly ad hoc, but they do not, in principle, rule it out.

A phonetic interpretation, on the other hand, would have it that the integration of the formant transitions into a unitary percept reflects the operation of a device specialized to perceive the sounds in a linguistically appropriate way. As for what is linguistically appropriate, it is plain that perceiving the transitions as rising or falling chirps is not. Language, after all, has no use for that kind of auditory information; it only requires to know whether the segment was [d] or [g]. Indeed, if the chirps and other curious auditory characteristics of speech sounds were heard as such, they would intrude as an intermediate stage of perception that itself had to be interpreted, however automatically. In that case listening to speech would be like listening to the acoustic alphabets of our early reading machines or to Morse code, and that would surely be awkward in the extreme.

What is required, if the time-varying transitions are to be perceived (appropriately) as unitary segments, is that the percept reflect neither the proximal sound nor the more distal movements it betokens but, rather, the still more distal, and presumably more nearly unitary, neural command structure that occasioned the movements. A less timid writer might call that the talker's *phonetic intent*.

But whatever the percept exactly corresponds to, I suppose that nature provided a device that is well adapted to its linguistic function, which is to make available to the listeners just those phonetic objects they need if they are to understand what the speaker said. But nature could not have anticipated the development of synthetic speech and dichotic stimulation, so it is possible to defeat her design in such a way as to discover something about what the design is. To do this we use a method that derives from a discovery by Rand (1974). (See also Isenberg & Liberman, 1978; Liberman, 1979.) Its special feature is a way of presenting patterns of synthetic speech so that an acoustic cue is perceived as a nonspeech sound and, simultaneously, as support for a phonetic percept. The obvious advantage of the method is that it holds the stimulus input constant while yet pro-

ducing two percepts, thus providing a control for auditory interaction. Recently the method has been applied by Mann, Madden, Russell, and Liberman (1981; Mann, Madden, Russell, & Liberman, Note 2) to determine how a time-varying formant transition is integrated into the perception of a stop consonant. The experiment was as follows.

To one ear we presented one or another of the nine formant transitions, as shown at the lower right of Figure 1. By themselves these isolated transitions sound like time-varying chirps—that is, like reasonably faithful auditory reflections of the time-varying acoustic signal. To the other ear we presented the rest of the pattern—the base—that is shown at the lower left of the figure. By itself the base is always perceived as a stop–vowel syllable; most listeners hear it as [da], some, as [ga].

When these two stimuli are presented dichotically, listeners report a duplex percept. On one side of the duplexity, the listeners perceive the syllable [da] or [ga], depending on the identity of the isolated transition. This speech percept is seemingly no different from the one that would have been produced had the base and the isolated transition been electronically mixed and presented in the normal manner. On the other side and at the same time, the listeners perceive a nonspeech chirp, not perceptibly different from what they experience when the transition is presented by itself. Thus, given exactly the same acoustic context and the same brain, the transition is simultaneously perceived in two phenomenally different ways: as critical support for a stop consonant, in which case it is integrated into a unitary percept, and as a nonspeech chirp, in which case it is not.

To go beyond the phenomenology just described, we determined how the transitions would be discriminated, depending on which side of the duplex percept the listener was attending to. For that purpose we sampled the continuum of for-

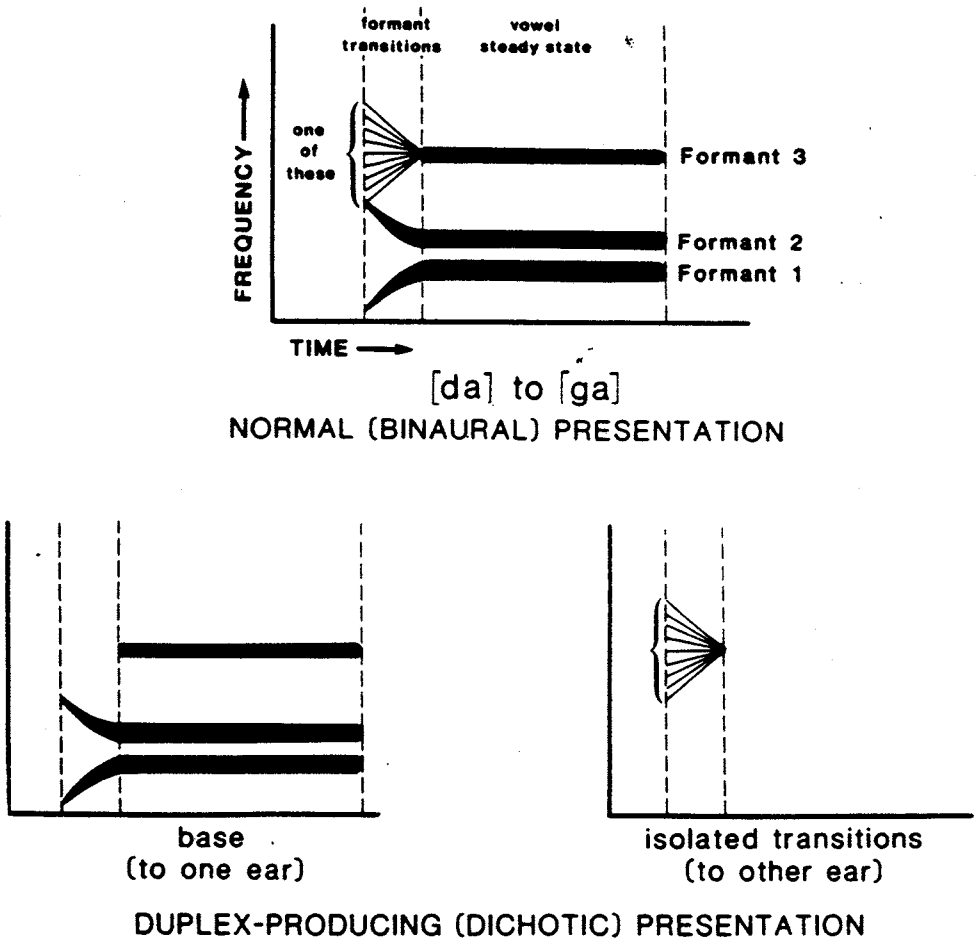

Figure 1. Schematic representation of the stimulus patterns used in the experiment on integration of the time-varying formant transitions (Mann, Madden, Russell, & Liberman, Note 2).

mant transitions by pairs, choosing, as members of each to-be-discriminated pair, stimuli that were three steps apart on the continuum of formant transitions shown in Figure 1. These we presented in an AXB format (A and B being the two stimuli to be discriminated and X being the one or the other) to subjects who were instructed to decide on the basis of *any* perceptible difference whether X was more like A or like B. When the subject's attention was directed to the speech side of the duplex percept, we obtained the results represented in Figure 2 by the solid line; with attention directed to the nonspeech side, we obtained the results shown by the dashed line. The difference is obvious. When the transitions support stop consonants—that is, when they are perceived in the phonetic mode—the discrimination function has a rather high peak, the location of which corresponds closely to the phonetic boundary. This is the familiar tendency toward categorical perception that characterizes segments such as these, a tendency that is itself highly adaptive, since it is only the categorical information—the segment is categorically [d] or [g]—that is most relevant linguistically. When the same transitions are perceived on the nonspeech side of the percept as chirps, the discrimination function, shown as the dashed line and open circles, is different; in fact it is nearly continuous.[3] Thus the discrimination functions confirm the more blatantly phenomenological results described earlier. Both indicate that integration of the formant transition into a phonetic percept is due to a special process that makes available to perception a unitary phonetic object well suited to its role in language.

The same phonetic process that integrates the transitions has other characteristics, of course, including one that has attracted attention for a long time: It adjusts perception to variations in the acoustic signal when those are caused by coarticulatory accommodation to changes in phonetic context; thus it seems to rest on a link between perception and production (Liberman et al., 1952; Mann, 1980; Mann & Repp, 1981). A second part of the experiment just described was designed to examine that perceptual adjustment to phonetic context and to exploit the duplex percept to identify the domain, auditory or phonetic, in which it occurs. To that end we took advantage of an earlier experiment by Mann (1980), in which she had found that placing the syllables [al] or [ar] in front of the [da]–[ga] patterns caused the position of the [da]–[ga] boundary (on the continuum of formant transitions) to shift toward the [g] end for [ar] and

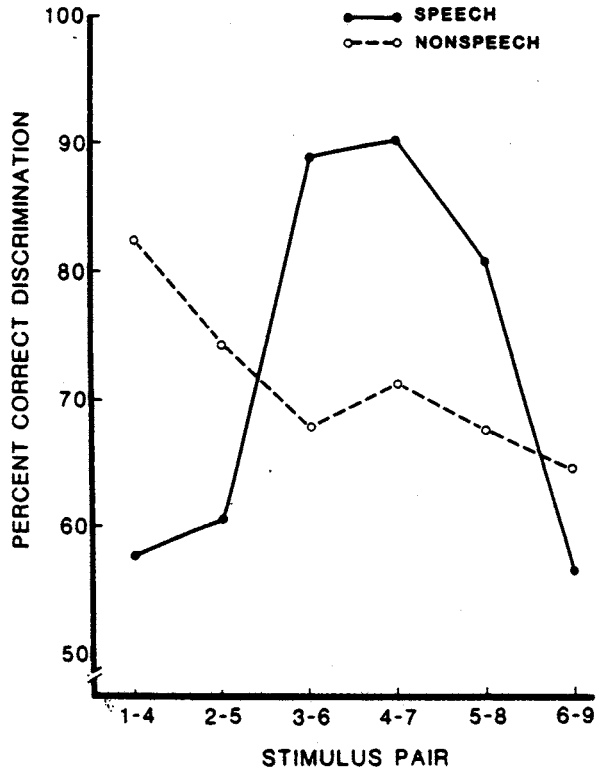

Figure 2. Discriminability of formant transitions when, on the speech side of the duplex percept, they supported perception of stop consonants and when, on the nonspeech side, they were perceived as chirps (from Mann, Madden, Russell, & Liberman, Note 2).

the [d] end for [al]. Since the shift was consistent with the change in [da]–[ga] articulation that can be shown to occur when the syllable [al] or [ar] is spoken immediately before, Mann inferred that this was indeed a case in which the perceptual system had automatically reflected coarticulation and its acoustic consequences.

Our further contribution to Mann's result was simply to repeat her experiment, but with the "duplex" procedure (and with measures of discrimination substituted for the phonetic identifications she had used). The outcome was straightforward. On the speech side of the duplex percept, we (in effect) replicated the earlier result, as shown by the results displayed in Figure 3. Taking the discrimination data obtained with the isolated [da]–[ga] syllables (solid line connecting solid circles) as baseline, we see that placing the syllable [ar] in front caused the discrimination peak (and presum-

[3] When the chirps are discriminated in isolation—that is, not as part of the duplex percept—the function has the same shape, but the level is displaced about 15% higher. The difference in level is presumably due to the distraction produced in the duplex condition by the other side of the percept.

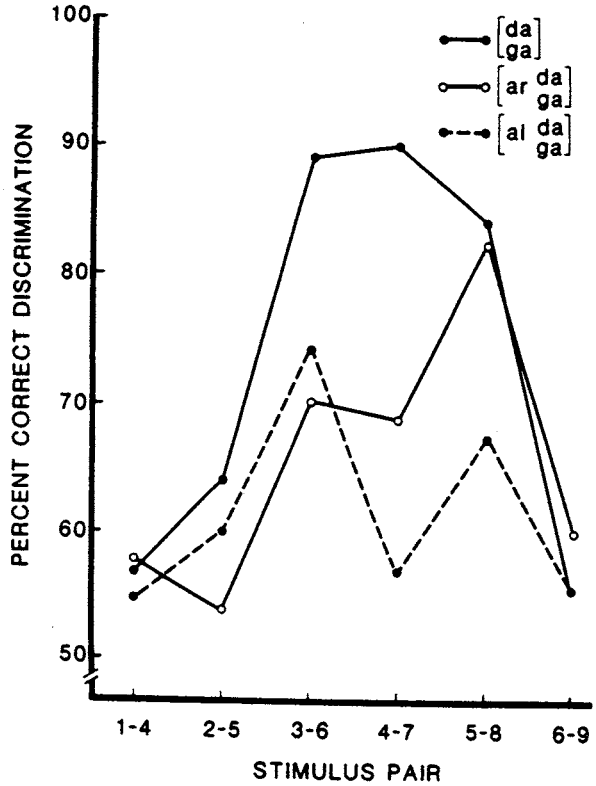

Figure 3. Discriminability of the formant transitions on the speech side of the duplex percept when the target syllables [da] and [ga] were in isolation and when they were presented by the syllables [ar] and [al] (from Mann, Madden, Russell, & Liberman, Note 2).

ably the phonetic boundary) to move to the right, toward the [g] end of the continuum of transitions. When [al] preceded, the peak (and the boundary) apparently shifted in the opposite direction—that is, to the left, toward [d]; for some subjects, indeed, it shifted so far as to move off the stimulus continuum, so for them there is no effective boundary, which explains why the peak is so low. For present purposes, however, the point is simply that there are large effects of prior phonetic context on discrimination of the transitions when those are perceived on the speech side of the duplex percept. On the other hand, as we see in Figure 4, the nonspeech side of the percept is unaffected by phonetic context: Discrimination of the formant transitions was the same whether the base was preceded by [al], by [ar], or by nothing.

Putting the two experiments together, we conclude that, given a single acoustic context, exactly the same formant transitions are perceived in two different modes. In the one mode they evoke nonspeech chirps that have a time-varying quality corresponding approximately to the time-varying stimulus; changes in the transitions are perceived continuously; and perception is unaffected by phonetic context. This is of course the auditory mode. In the other mode the same transitions provide critical support for the perception of stop consonants that lack the time-varying quality of the nonspeech chirps; changes in the transitions are perceived more or less categorically; and perception is markedly affected by phonetic context. This is the phonetic mode.

INTEGRATION OF SOUND AND SILENCE

Perception of a phonetic segment typically depends, as indicated earlier, on the integration of several (many may be a more appropriate word) acoustic cues. Even in the case of [da] and [ga] just described, there was one other cue, silence preceding the transitions, though I did not remark it. To show the effect of such silence—an effect long known to researchers in speech (Bastian, Delattre, & Liberman, 1959)—we must put the stop consonant and its transition cues into some other position, as in the examples [spa] and [sta] shown at the top of Figure 5. As we see there, an important cue for perception of stop consonants—in this case

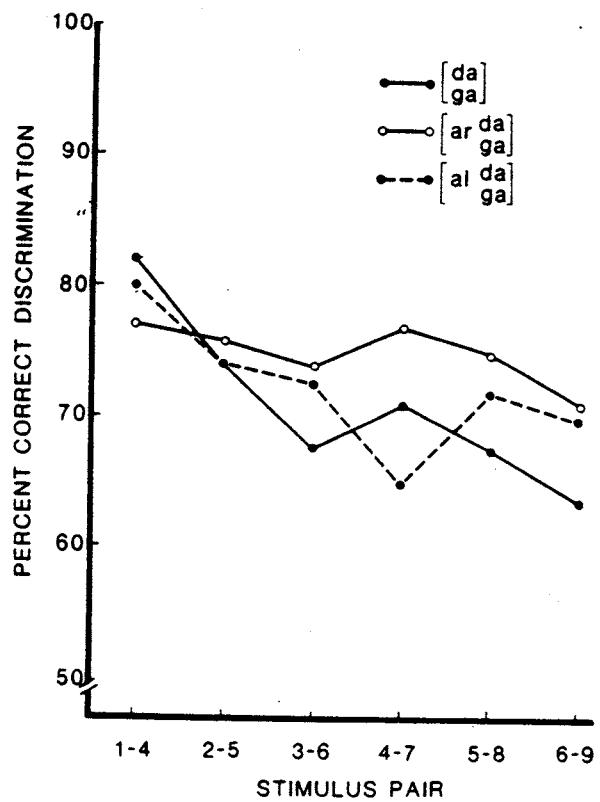

Figure 4. Discriminability of the formant transitions on the nonspeech side of the duplex percept under conditions identical to those represented in Figure 3 (from Mann, Madden, Russell, & Liberman, Note 2).

[p] and [t]—is a short period of silence between the noise of the fricative and the formant transitions that introduce the vocalic part of the syllable (Dorman, Raphael, & Liberman, 1979).

But why is silence necessary, and in which domain, auditory or phonetic, is it integrated with the transition cues to produce stop consonants? On an auditory account we might suppose that there is forward masking of the transition cues by the fricative noise, in which case the role of the intervening silence is to provide time for the transitions to evade masking. Failing that, we could, as always, invoke some previously unnoticed interaction between frequency sweeps (transitions) and silence that is presumed to be characteristic of the way the auditory system works.

A phonetic interpretation, on the other hand, takes account of the fact that presence or absence of silence supplies important phonetic information—to wit, that the talkers closed their vocal tracts, as they must to produce the [p] and [t] in [spa] and [sta], or that they did not, as they do not when saying [sa]. Presumably the processes of the phonetic mode are sensitive to the phonetic significance of the information that silence imparts.

To decide between these interpretations, the phenomenon of duplex perception was again exploited (Liberman, Isenberg, & Rakerd, 1981). As shown in Figure 5, base stimuli that sometimes did and sometimes did not have silence were presented dichotically with transition cues appropriate for [p] or for [t]. Two such dichotically yoked patterns were presented on each trial; subjects were asked to identify the speech percepts and to discriminate the nonspeech chirps. The result was that the subjects fused the transitions with the base and accurately perceived [sa], [spa], or [sta], depending on the presence or absence of silence in the base (to one ear) and the nature of the formant transitions (to the other). But the subjects also perceived the transitions as nonspeech chirps and accurately discriminated them as same or different, regardless of whether there was silence in the base. Thus duplex perception did occur, and silence affected the identification of the speech, but not the discrimination of the nonspeech.

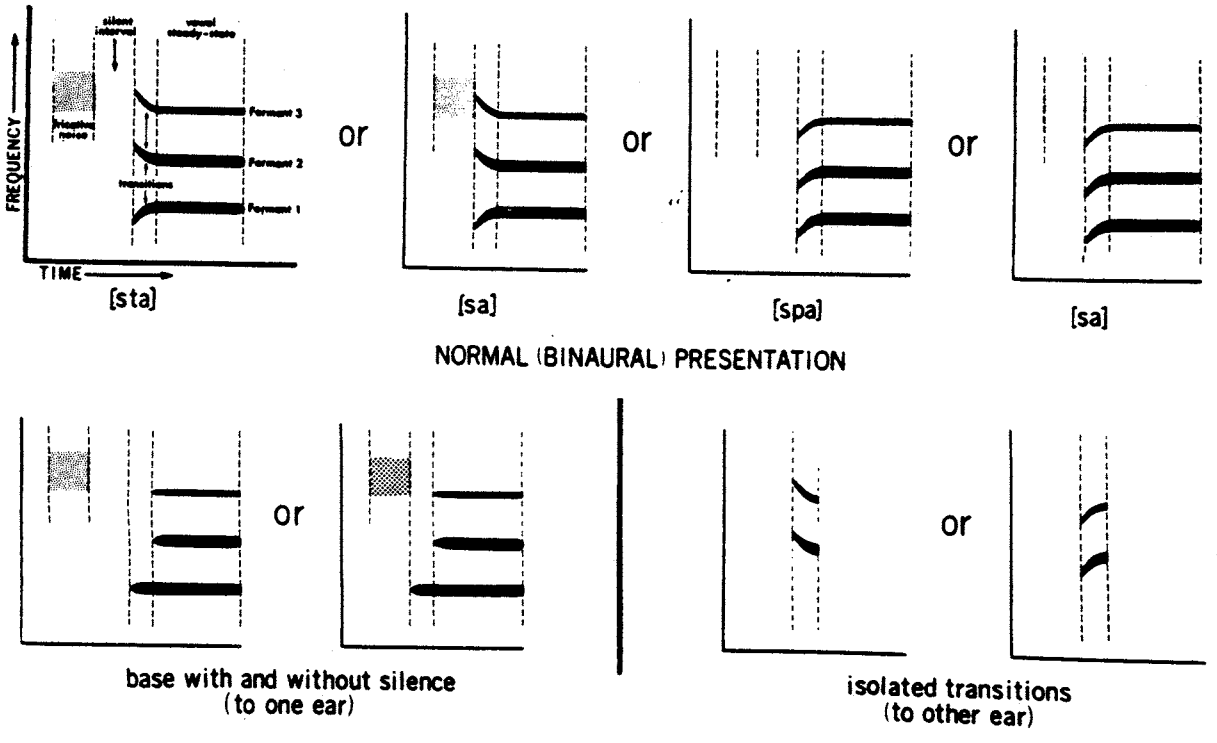

Figure 5. Schematic representations of the stimulus patterns used to determine whether the importance of silence as a cue is owing to auditory or phonetic factors. (From "Duplex Perception of Cues for Stop Consonants: Evidence for a Phonetic Mode," by A. M. Liberman, D. Isenberg, and B. Rakerd, *Perception & Psychophysics*, 1981, *30*, 133–143. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

In a further experiment the investigators provided a more severe test by asking subjects to discriminate their percepts on both sides of the duplexity. For that purpose two dichotically yoked pairs of stimuli were presented on each trial, so arranged as to exhaust all combinations of silence–no silence in the base and [p]–[t] cues in the isolated transitions. Subjects were asked, for each pair of percepts, to rate their confidence that a difference of any kind had been detected. The results are shown in Figure 6. There are but two critical comparisons. The first is in the leftmost third of the figure, in the condition in which there was no silence in either of the two base stimuli presented to the one ear (labeled *No Silence–No Silence*) and the two transition cues to the other ear were different (labeled simply *Different*). On the speech side of the duplexity (open bar), we see that the difference between the transitions was not clearly detected, presumably because, in the absence of silence in either base stimulus, subjects perceived [sa] in both cases. But on the nonspeech side (shaded bar), the same difference *was* detected; here the absence of silence in the base made no difference. The other critical comparison is seen in the bars immediately to the right, in the middle third of the figure, representing the condition that had, in one ear, silence in one base stimulus but not the other and, in the other ear, two transition cues that were the same. On the speech side of the duplex percept, we see that the patterns were perceived as very different, even though the transition cues were the same; presumably this was because one percept, being influenced by the presence of silence, included a stop consonant, whereas the other, being influenced by the absence of silence, did not. The result on the nonspeech side stands in contrast. There the percepts were judged to be not very different, accurately reflecting that they were in fact not different.

Thus in both critical comparisons, silence affected discrimination of the transitions only on the speech side of the duplex percept. Apparently its importance depends on distinctively phonetic processes, and its integration with the transition occurs in the phonetic mode.

The integration of silence and transitions, as in the patterns just described, reinforces the suggestion, made earlier in regard to the integration of the transitions alone, that the perceived object is not to be found in the movements of the speech organs at the periphery but, rather, at some still more distal remove, as suggested by Repp, Liberman, Eccardt, and Pesetsky (1978). To see the point more clearly, we should first take note of a finding that adds another cue for the [p] in [spa]: the shaping of the fricative noise that is caused by the way the vocal tract closes for [p] (Summerfield, Bailey, Seton, & Dorman, 1981). Now we have three acoustic cues that correspond neatly to three corresponding aspects of the articulation. There is, first, the shape of the fricative noise, which signals the closing of the tract; then the silence, which signals the closure itself; and finally the formant transitions, which signal the subsequent opening into the vowel. If these three acoustic cues are integrated into a percept that does not display at least three constituent elements, then the perceived

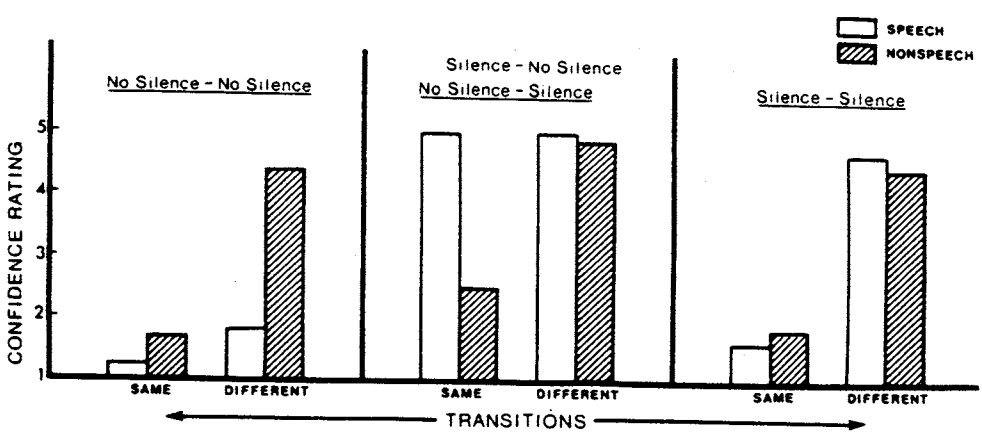

Figure 6. Mean ratings of confidence that the two percepts (speech or nonspeech) were different. (From "Duplex Perception of Cues for Stop Consonants: Evidence for a Phonetic Mode," by A. M. Liberman, D. Isenberg, and B. Rakerd, *Perception & Psychophysics*, 1981, *30*, 133–143. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

object must be upstream from the peripheral articulation. A likely candidate, as suggested earlier, is the unitary command structure from which the various movements at the periphery unfolded.

INTEGRATION OF PERIODIC SOUND AND NOISE

When talkers close their vocal tracts to produce a stop consonant and then open them into a following vowel, the resulting silence and formant transitions are, as we have seen, integrated into a stop consonant. It is surely provocative that similar formant transitions are produced, but without the silence, when talkers almost close their vocal tracts to make the noise of a fricative (e.g., [s]) and then open into the vowel, for in such cases the formant transitions do not support stops; they are, instead, integrated with the noise into the perception of a fricative (Harris, 1958; Mann & Repp, 1980; Whalen, 1981). Such integration is shown in Figure 7, which reproduces the results of a recent experiment by Repp (1981). What we see in the figure are the judgments [ʃ] or [s], made to stimuli that were constructed as follows. The experimental variable, ranged on the abscissa, was the position on the frequency scale of a patch of band-limited noise as it moved between a place appropriate for [ʃ] and one appropriate for [s]. The parameters were the nature of the (following) formant transitions—appropriate, in the one case, for [s] and, in the other, for [ʃ]—and the two vowels [a] and [u]. We see that the transitions (and also the vowels) affected the perception of the fricative.

Though not shown in this particular experiment, I would note parenthetically that patterns like these, but with 50 milliseconds of silence inserted between the fricative noise and the vocalic section, will be perceived, not as fricative–vowel syllables, but as fricative–stop–vowel syllables (Mann & Repp, 1980). That is, inserting 50 milliseconds of silence will cause the formant transitions to be integrated, not into fricatives, but into stops. It is difficult to account for that as an auditory effect, but easy to see how it might reflect a special sensitivity to information about a difference in articulation that changes the phonetic "affiliation" of the acoustic transitions.

In a further, and more severe, test of the integration of transitions and fricative noise that we saw in Figure 7, Repp measured the effect of the formant transitions on the way listeners discriminated variations in the frequency position of the noise patch, using for this purpose the highly sensitive method of fixed standard. He found two

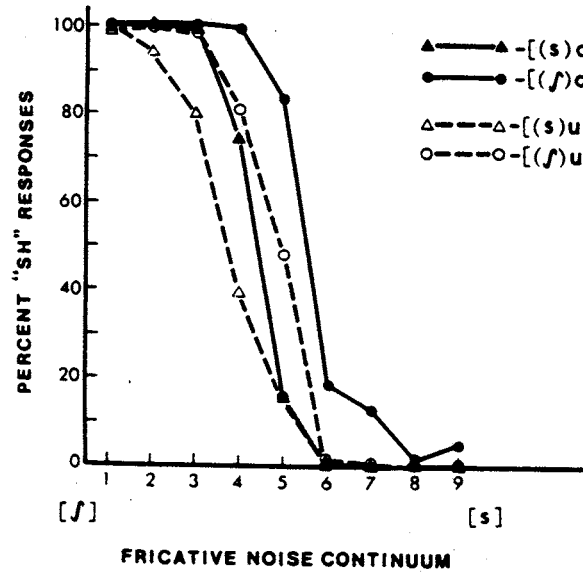

Figure 7. Identification functions for a [ʃ]-[s] noise continuum when connected to [ʃ]-appropriate or [s]-appropriate transitions and the vowels [a] or [u]. (From "Two Strategies in Fricative Discrimination," by B. H. Repp, *Perception & Psychophysics*, 1981, *30*, 217–227. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

distinctly different types of discrimination functions. One clearly showed an effect of the formant transitions and reflected nearly categorical perception; the other just as clearly showed no effect of the formant transitions and represented perception that was nearly continuous. Which type Repp obtained in each particular case apparently depended on the listener's ability to isolate or "stream" the noise—that is, to create an effect similar, perhaps, to the one obtained by Cole and Scott (1973) when they found with fricative–vowel syllables that, as a result of repeated presentation, the noise and vocalic sections would form separate "streams" that had little apparent relation to each other. At all events we have here another instance, though occurring in a different phonetic class and obtained by very different methods, of a single acoustic pattern that is perceived in two distinctly different ways. One reflects the integration of cues in the phonetic mode, the other, the "nonintegration" of the same acoustic elements in the auditory mode.

There is still another method that exploits the possibility of perceiving exactly the same stimulus pattern in two ways and thus enables us to test yet again whether the integration of formant transitions and noise occurs in the phonetic or auditory mode. But now the two ways of perceiving are not

speech versus nonspeech, as in the experiments described thus far, rather, two kinds of speech—namely, fricatives and stops. The relevant experiment is a recent one by Carden, Levitt, Jusczyk, and Walley (1981). Starting with synthetic patterns that produced stop–vowel syllables, they varied the second-formant transitions and found the boundary between [b] and [d]. Then they placed in front of these patterns a fixed patch of band-limited noise, neutralized as between the fricatives [f] and [θ]. In these patterns the formant transitions cue the difference between the fricatives, but because the place of vocal-tract constriction is different for the two fricatives, on the one hand, and the two stops, on the other, the perceptual boundary on the continuum of formant transitions is now displaced. That is, exactly the same formant transitions distinguish the fricatives differently from the way they distinguish the stops. The effect seems most plausibly to be phonetic, reflecting the listener's knowledge, as it were, of the difference in articulatory place of production between the stops [b] and [d], on the one hand, and the fricatives [f] and [θ], on the other. To make sure, Carden and his collaborators presented the patterns with the noise patch to one group of subjects and boldly asked them to perceive stops; then, in precisely reverse fashion, they presented the patterns without the noise patch to a second group with instructions to perceive fricatives. The listeners' judgments reflected boundaries on the continuum of transitions that were appropriate to the class of phonetic segments ([b] vs. [d] or [f] vs. [θ]) they were asked to hear. Thus exactly the same acoustic patterns yielded different boundaries on the continuum of transitions, depending on whether the listeners were perceiving the patterns as stops or as fricatives. Discrimination functions were also obtained, and these confirmed the boundary shift. We see, then, that transition cues like those that integrate with silence to produce a stop consonant will integrate with noise to produce a fricative. In both cases the integration is in the phonetic mode.

THE EQUIVALENCE OF SOUND AND SILENCE WHEN INTEGRATED

Implicit in the discussion so far is the assumption that when acoustic cues integrate to form a phonetic percept, they are perceptually equivalent for that purpose; otherwise it would make no sense to speak of the percept as unitary. It is not implied that the cues are necessarily of equal importance or power, only that their separate contributions are not sensed as separate. But even that implication is of interest from a theoretical point of view because the cues are often very different acoustically, having in common only that they are the common products of the same linguistically significant gesture. Hence their equivalence is to be attributed, most reasonably, to the link between perception and production that presumably characterizes phonetic processes.

But the implied equivalence of diverse cues is so far just that—implied. To test the equivalence more directly was the purpose of several experiments. One of these, by Fitch, Halwes, Erickson, and Liberman (1980), was designed to examine the equivalence of silence and formant transitions in perception of the stop consonant in *split*, as opposed to its absence in *slit*. Synthetic patterns like those shown in Figure 8 were used. The variable was the duration of silence between the fricative noise and the vocalic portion of the syllable; the parameter of the experiment was the nature of the formant transitions at the start of the vocalic sections, set to bias that section toward [lit] in the one case and toward [plit] in the other. When stimuli that had been constructed in this way were presented for identification as *slit* or *split*, the results shown in Figure 9 were obtained. One sees there a trading relation not different in principle from those found by other investigators with other cues. (For a review, see Repp, in press.) The displacement of the two response functions indicates that, for the purpose of producing the [p] in *split*, about 20 milliseconds of silence is equal to appropriate formant transitions.[4] Thus silence is equivalent to sound, but only when both are produced as parts of the same phonetic act.

Of course it might be argued that the *splits* produced by the two different combinations of silence and sound were not really equivalent, but the forced-choice identification procedure, permitting only the responses *slit* or *split*, gave the subjects no opportunity to say so. Against that possibility we carried out another experiment, designed to determine how well the subjects could discriminate selected combinations of the stimuli on any basis whatsoever. The rationale for selection of stimuli was as follows. If the two cues, silence and sound, are truly equivalent in phonetic perception, their

[4] The existence of these trading relations means that the location of a phonetic boundary on an acoustic continuum is not fixed; within limits it will move as the settings of the several cues are changed. The boundary will also move, of course, as a function of phonetic context. (See the foregoing discussion of the effect of preceding context on the [da]–[ga] boundary; also, e.g., Mann & Repp, 1981; Repp & Mann, 1981.)

perceptual effects should be algebraically additive, as it were. Thus, given two synthetic syllables to be discriminated and a base-line level of discriminability determined for pairs of stimuli that differ in only one of the cues, it should be possible to add the second cue to increase or decrease discriminability, as the phonetic "polarity" of the two cues causes their effects to work together or at cross purposes. The cues should "summate," or "cooperate," when they are biased in the same phonetic direction—as when one of the syllables to be discriminated combines a silence cue that is longer by the amount of the "trade" with transition cues of the [plit] type and the other syllable combines a silence cue that is shorter by the amount of the trade with transition cues of the [lit] type. They should cancel each other or conflict when the opposite pairing is made—that is, when the longer silence cue is combined with transition cues of the [lit] type and the shorter silence cue with transition cues of the [plit] type. Pairs of stimuli meeting those specifications, and sampling the continuum of silence durations, were presented for forced-choice discrimination. As shown in Figure 10, discrimination of patterns differing by both cues was in fact either better or worse than patterns that differed by only one, depending on whether the cues were calculated to cooperate or to conflict. Apparently the effects of the two cues did converge on a single perceptual object. By this test, then, the cues may be said to be equivalent and the percept may be said to be truly unitary.

That the equivalence of silence and sound in the above example is due to phonetic processes is supported in an experiment by Best, Morrongiello, and Robson (1981). Indeed, it is supported there more strongly than in the experiment just described, because Best and her collaborators found that the equivalence was manifest only when the stimulus patterns were perceived as speech. As a first step they performed an experiment very similar to the one by Fitch et al. (1980), except that the stimuli were *say–stay* instead of *slit–split* and the tran-

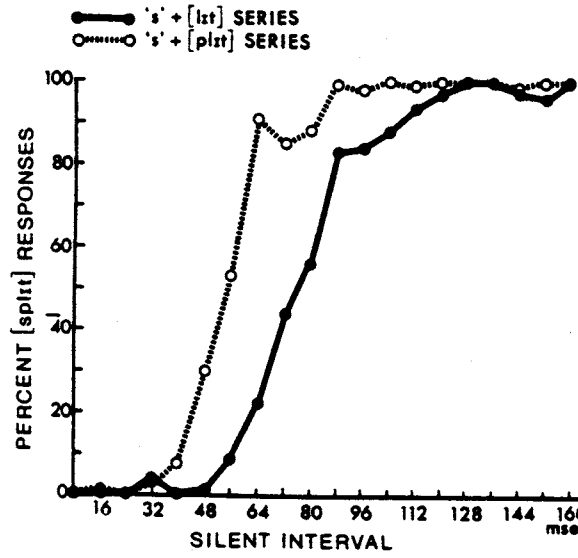

Figure 9. Effect of silent interval on perception of /slit/ vs. /split/ for the two settings of the transition cue. (From "Perceptual Equivalence of Two Acoustic Cues for Stop-Consonant Manner, by H. L. Fitch, T. Halwes, D. M. Erickson, and A. M. Liberman, *Perception & Psychophysics*, 1980, 27, 343–350. Copyright 1980 by the Psychonomic Society, Inc. Reprinted by permission.)

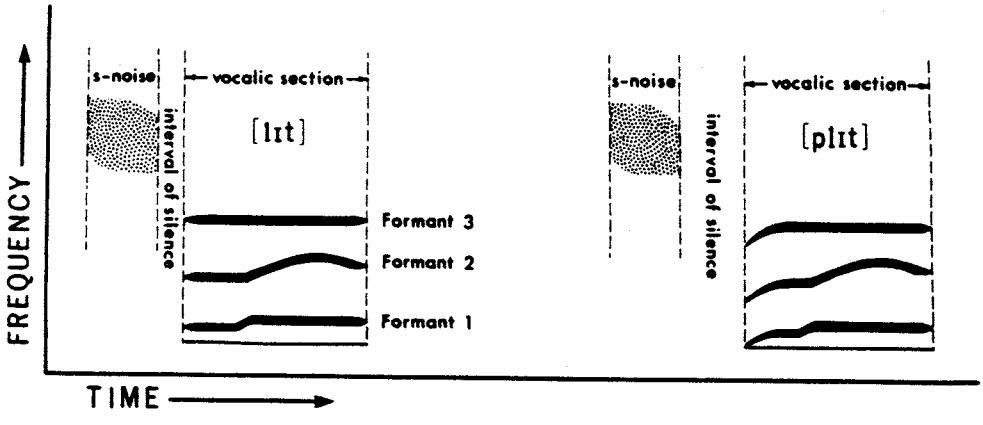

Figure 8. Schematic representation of the patterns used to evaluate the equivalence in stop-consonant perception of silence and formant transitions, showing both settings of the transitions and two representative settings of the silence cue. (From "Perceptual Equivalence of Two Acoustic Cues for Stop-Consonant Manner," by H. L. Fitch, T. Halwes, D. M. Erickson, and A. M. Liberman, *Perception & Psychophysics*, 1980, 27, 343–350. Copyright 1980 by the Psychonomic Society, Inc. Reprinted by permission.)

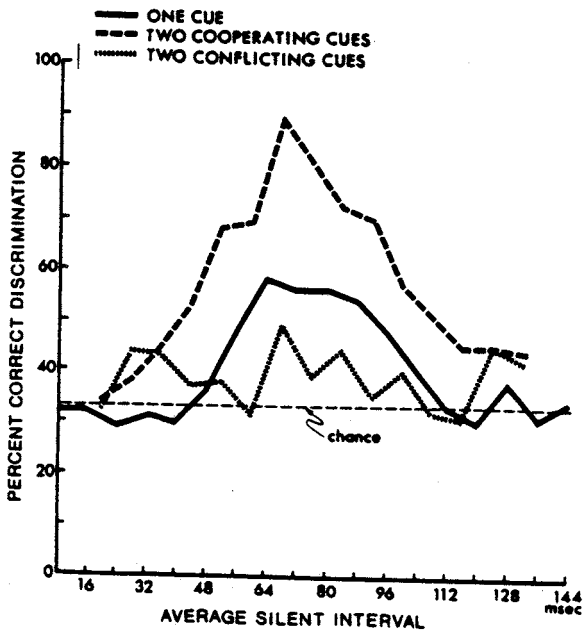

Figure 10. Percent correct discrimination for pairs of stimuli that differ by one cue or by two cues of the same (cooperating cues) or opposite (conflicting cues) phonetic polarities. (From "Perceptual Equivalence of Two Acoustic Cues for Stop-Consonant Manner," by H. L. Fitch, T. Halwes, D. M. Erickson, and A. M. Liberman, *Perception & Psychophysics*, 1980, 27, 343–350. Copyright 1980 by the Psychonomic Society, Inc. Reprinted by permission.)

sition-cue parameter was simply the frequency at which the first formant started. With these stimuli they obtained the identification functions shown in Figure 11. We see there almost exactly the same kind of trading relation between silence and formant transition that had been found in the earlier experiment. In the manner of Fitch et al., they also tested discrimination, finding, just as Fitch et al. had, that the two cues could be made to cooperate or to conflict depending on their phonetic polarities. But now they performed an experiment that proved to be particularly revealing. Borrowing a procedure that had been used successfully for a similar purpose (Bailey, Summerfield, & Dorman, 1977; Dorman, 1979; Lane & Schneider, Note 3), and more recently made the object of further attention (Remez, Rubin, Pisoni, & Carrell, 1981), they replaced the formants of the vocalic portion of the syllable with sine waves, taking care that the sine waves followed exactly the course of the formants they replaced. The sounds that result are perceived by most people, at least initially, as nonspeech patterns of noises and tones. But some spontaneously perceive them as speech, and others perceive them so after it has been suggested to them that they might. It is possible, thus, to obtain identification and discrimination functions for the same stimuli when in the one case, they are perceived as speech and when in the other, they are not. (When perceived as nonspeech the patterns are, of course, not readily identifiable, but identification functions can be obtained by presenting on each trial the target stimulus—that is, the stimulus to be identified—together with the two stimuli at the extremes of the continuum and then asking the subject to say whether the target stimulus is more like one or the other of the extremes. To insure comparability the same procedure is used when the subjects are perceiving the stimuli as speech.) The results are shown in Figure 12. We see in Figure 12a that when the subjects were perceiving the patterns as speech (say–stay listeners), their identification functions exhibited the now-familiar trading relation. But when the same stimuli were perceived as nonspeech, then, as shown in Figures 12b and 12c, two quite different patterns emerged, depending on whether, as inferred from the subjects' descriptions of the sound, they were attending to the transition cue (spectral listeners) or the silence cue (temporal listeners). It is of course precisely because the subjects could not integrate the

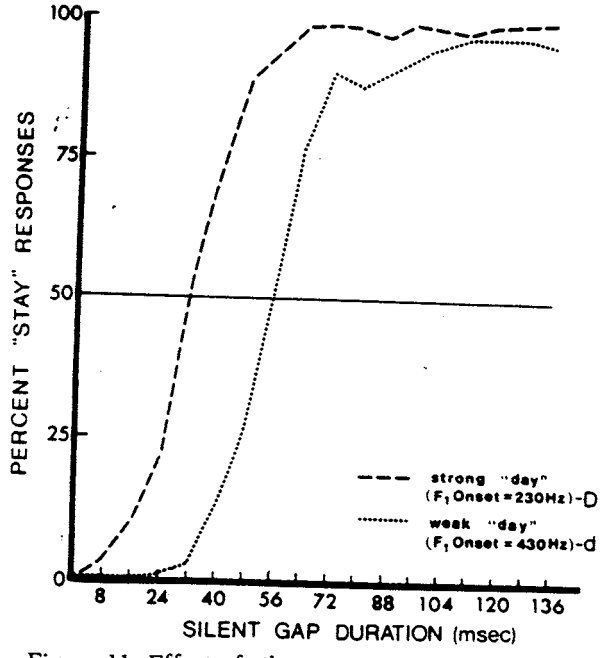

Figure 11. Effect of silent interval on perception of /say/ vs. /stay/ for the two settings of the transition cue. (From "Perceptual Equivalence of Acoustic Cues in Speech and Nonspeech Perception," by C. T. Best, B. Morrongiello, and R. Robson, *Perception & Psychophysics*, 1981, *29*, 191–211. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

cues in the nonspeech percept that they chose, as it were, between the one cue and the other. In any case both of the identification functions in the nonspeech case are different from the one that characterizes the response to exactly the same stimuli when they were perceived as speech. (Discrimination functions obtained with the same stimuli were also different depending on whether the stimuli were perceived as speech, nicely confirming the result obtained with the identification measure.) Thus, with yet another method for obtaining speech and nonspeech percepts from the same stimulus, we again find evidence supporting the existence of a phonetic mode, and we see that the equivalence of integrated cues is to be attributed to the distinctively phonetic processes it incorporates.

THE EQUIVALENCE OF SOUND AND SIGHT WHEN INTEGRATED

Perhaps the most unusual evidence relevant to the issue under discussion comes from a startling discovery by McGurk and MacDonald (1976) about the influence on speech perception of optical information about the talker's articulation. (See also MacDonald & McGurk, 1978; Summerfield, 1979.) When subjects view a film of a talker saying one syllable while a recorded voice says another, then under certain conditions they experience a unitary percept that overrides the conflicting optical and acoustic cues. Thus, for example, when the talker articulated [ga] or [da] and the voice said [ba], most subjects perceived [da]. In that case the effect of the optical stimulus was, at the very least, to determine place of production. When in a subsequent experiment by McGurk and Buchanan (Note 4), the talker was seen to produce the syllables [ba], [va], [ða], [da], [ʒa], [ga], [ha] while the recorded voice said [ba] over and over again, most subjects perceived [ba], [va], [ða], [da], [da], and then for visual [ha], a variety of percepts other than [ba]. Here both place of articulation and manner of articulation were determined by the optical input. (The difficulty of seeing farther back in the vocal tract than [da] presumably accounts for the fact that visual [ʒa], [ga], and [ha] were perceived as having generally more forward places of production.)

Having witnessed a demonstration of the McGurk-MacDonald effect, I take the liberty of offering testimony of my own. I found the effect compelling, but more to the point, I would agree that McGurk and Buchanan (Note 4) have captured my experience when they say, "the majority of listeners have no awareness of bimodal conflict," and then describe the percept as "unified." Surely my percept was unified in the important sense that I could not have decided by introspective analysis that part was visual in origin and part auditory. Even in those cases in which, given conflicting optical and acoustic cues, I experienced two syllables, there was nothing about their quality that would have permitted me to know which I had seen and which I had heard.

By way of interpretation, MacDonald and McGurk (1978) indicated that their results bespeak a connection between perception and production,

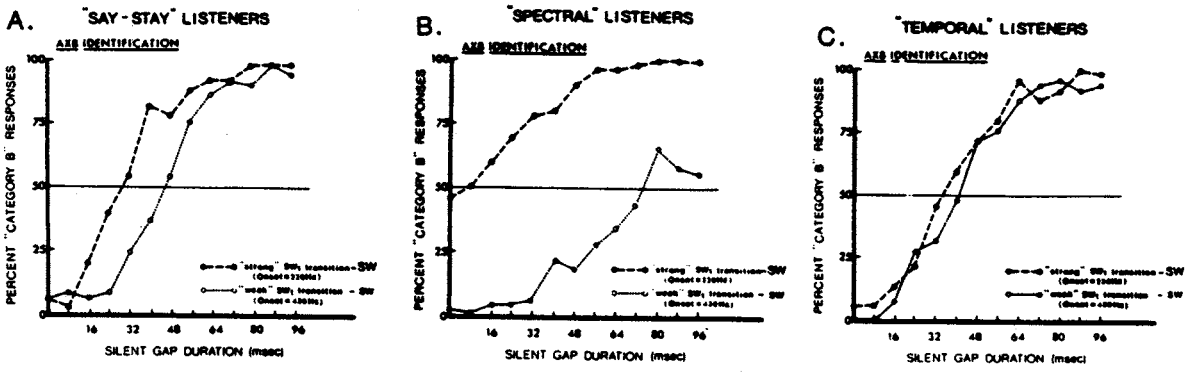

Figure 12. Effect of silent interval on "identification" of sine-wave analogues of *say-stay* stimuli. (Graph A is for those subjects who perceived these stimuli as speech ["say-stay" listeners]. Graphs B and C are for those who perceived them as nonspeech, divided, according to their reports of what the sounds were like, into those who were apparently attending to the transition cue [Graph B, "spectral" listeners] or, alternatively, the silence cue [Graph C, "temporal" listeners]. (From "Perceptual Equivalence of Acoustic Cues in Speech and Nonspeech Perception," by C. T. Best, B. Morrongiello, and R. Robson, *Perception & Psychophysics*, 1981, 29, 191–211. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

and McGurk and Buchanan (Note 4) echo a comment by Summerfield (1979), who observed, after having himself performed several experiments on the phenomenon, that the optical and acoustic signals are picked up in a "common metric of articulatory dynamics." I would agree, though I would of course prefer to call the common metric *phonetic*. But a mode by any other name would bear as weightily on the issue, for the important consideration is that, in any ordinary sense of modality, the speech percept is neither visual nor auditory; it is, rather, something else.

INTEGRATION INTO ORDERED STRINGS

Having so far considered only the perception of individual phonetic segments, we should put some attention on the fact that phonetic segments are normally perceived in ordered strings. This wants explicit treatment, if only because, as the reader may recall, a characteristic of the speech code is that several phonetic segments are conveyed simultaneously by a single segment of sound. As the reader may also recall, it is just this characteristic of the code that enables the listener to evade the limitation imposed by the temporal resolving power of the ear. The further consequence for perception, which we consider now, is that listeners cannot perceive phonetic segment by phonetic segment in left to right (or right to left) fashion; rather, they must take account of the entire stretch of sound over which the information is distributed. Such an acoustic stretch typically signals a phonetic structure that comprises several segments. I offer a brief example, taken from a recent study by Repp et al. (1978) and chosen because the relevant span happens to cross a word boundary.

The experiment dealt with the effect of two cues, silence and noise duration, on perception of the locutions *gray ship, gray chip, great ship*, and *great chip*. Figure 13 is a spectrogram of the words *gray ship*, with which the experiment began. The variable, shown in the figure, was the duration of silence between the two words. Given the results of previous research, we knew that increasing the silence would bias away from the fricative in *ship* and toward the affricate (stop-initiated fricative) in *chip* (Dorman, Raphael, & Isenberg, 1980; Dorman et al., 1979). The parameter, also shown in the figure, was the duration of the fricative noise, known from previous research to be a cue for the same distinction: increases in duration of the noise bias toward fricative and away from affricate (Dorman et al., 1979; Gerstman, 1957).

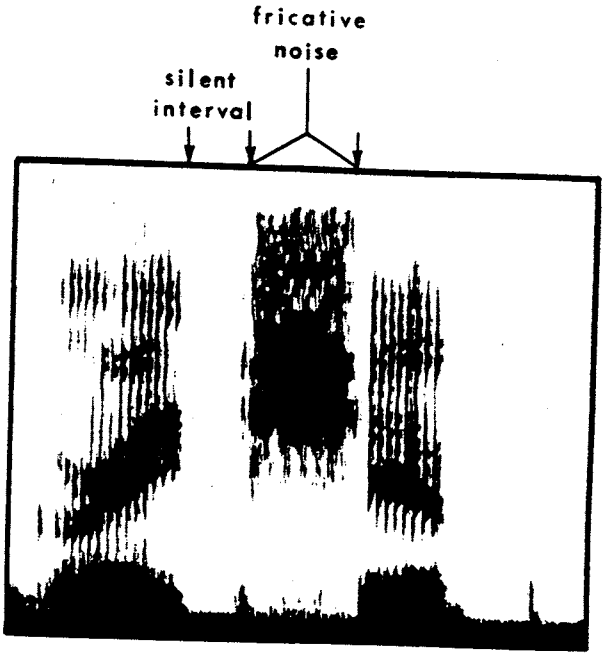

Figure 13. Spectrogram of the words *gray ship*.

Figure 14 displays the results. We see in the graph at the upper left that when the noise duration was relatively short (62 milliseconds), increasing the duration of the silence caused the percept to change from *ship* to *chip*. Thus the effect of silence was to produce a stop-like consonant to its right, much as it had done in the cases of *slit–split* and [sa]–[spa]–[sta], which were dealt with earlier. But as shown in the graph at the lower right, when the duration of the fricative noise was relatively long (182 milliseconds), increases in the duration of the silence caused the perception to change, not to *gray chip* as before, but to *great ship*. That is, increasing the duration of the fricative noise in *ship* put a stop consonant at the end of the preceding word. The effect is superficially right to left. But of course the effect is in neither direction; it is more properly regarded as a matter of apprehending a structure.

Given, then, that the listener must recover several phonetic segments from the same span of sound, we ask three questions about the underlying process. First, how listeners delimit the acoustic span? That is, how do they know when all the information that is to be provided has been provided? There is, after all, no acoustic signal that regularly marks the information boundary. Second, how do listeners store the information as it accumulates? And third, what do they do while waiting? Do they simply resonate, as it were, or do they entertain hypotheses? If the latter, do they

entertain all possible hypotheses? Do they weight them according to the likelihood they are correct? And how quickly do they abandon them as they are proved wrong?

If these questions seem familiar to students of sentence perception, it is because processes in the phonetic and syntactic domains do have something in common. In both cases information is distributed in distinctively linguistic ways through the signal. As a consequence the perceiver must recover distinctively linguistic structures. To that extent the resemblance between processing in the two domains is not superficial. Nor is it, if we take the vertical view of language I earlier espoused, altogether surprising.

Afterwords, Omissions, and Prospects

Having set out years ago to study communication by acoustic alphabets, we might still be so occupied. For acoustic alphabets can be used for communication—witness Morse code—and there are innumerable experiments we could have done had we gone on trying to find the alphabet that works best. But it is not likely, as a practical matter, that we would ever have made a large improvement.

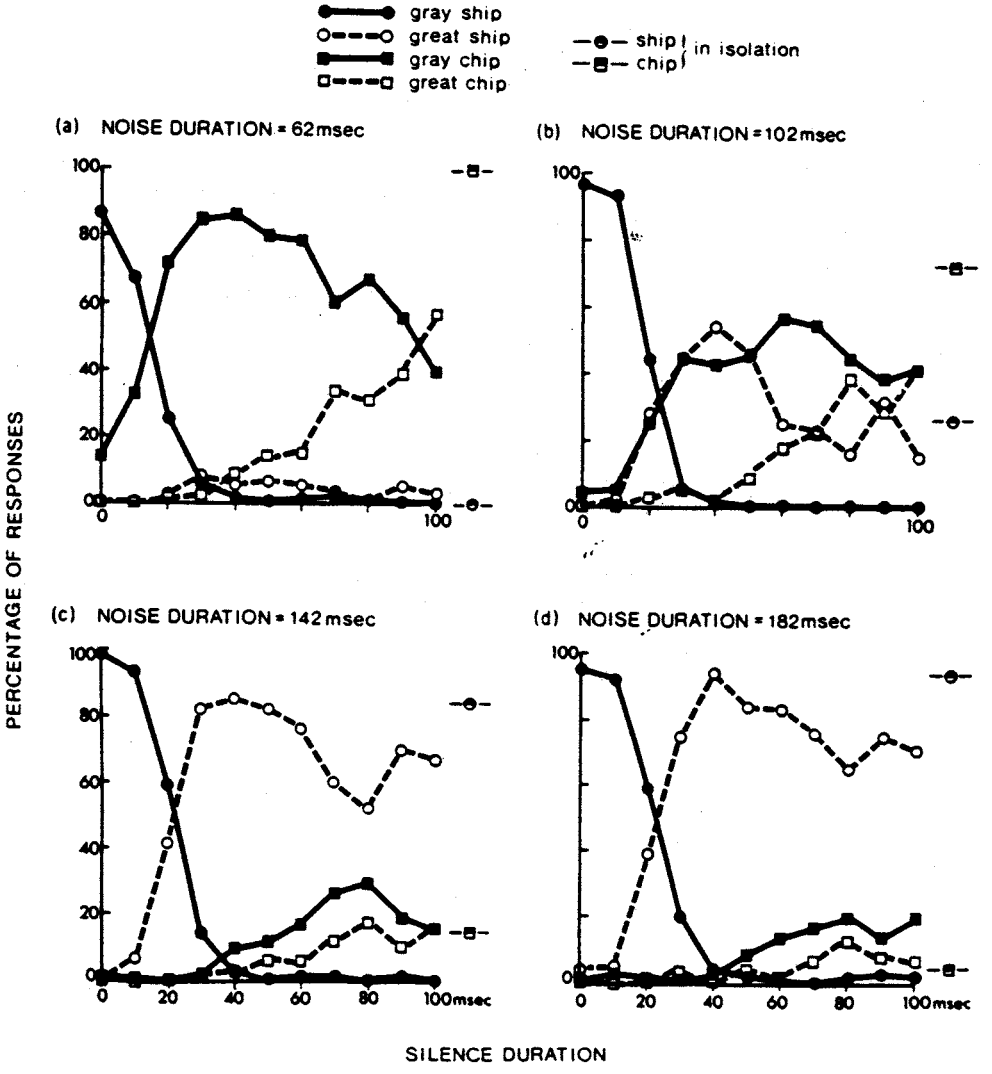

Figure 14. The effect of duration of silence, at each of four durations of fricative mode, on the perception and placement of stop (or affricate) manner. (From "Perceptual Integration of Acoustic Cues for Stop, Fricative, and Affricate Manner," by B. H. Repp, A. M. Liberman, T. Eccardt, and D. Pesetsky, *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 621–637. Copyright 1978 by the American Psychological Association, Inc. Reprinted by permission.)

Nor is it likely, from a scientific point of view, that we would ever have learned anything interesting. Acoustic alphabets cannot become part of a coherent process; I suspect, therefore, that there is nothing interesting to be learned.

But speech was always before us, proof that there is a better way. Inevitably, then, we put our attention there and, in so doing, began to bark up the right tree. It remained only to find that speech and language require to be understood in their own terms, not by reference to diverse processes of a horizontal sort. But once the vertical view is adopted, there is little doubt about what we must try to understand.

There is also little doubt at any stage of the research on speech about how much or how little we do understand, because there is a standard by which progress can be measured; we are not in the position of explaining behavior that we have ourselves contrived. Thus to test what we think we know of the relation between phonetic structure and sound, we have only to see how that knowledge fares when used as a basis for synthesis. In fact it does well enough to enable us to synthesize reasonably intelligible speech, which suggests that we do know something (Klatt, 1980; Liberman, Ingemann, Lisker, Delattre, & Cooper, 1959; Mattingly, 1980). But the speech is not nearly so good as the real thing, which proves, as if proof were needed, that we have something still to learn. Perhaps what we must learn most generally is to accept the hypothesis, alluded to earlier, that human listeners are sensitive to all the phonetically relevant information in the speech signal. If that hypothesis is true, and if the acoustic cues that convey the information are as numerous, various, and intertwined as we now believe them to be, then we should act on our assumption that the key to the phonetic code is in the manner of its production. That requires taking account of all we can learn about the organization and control of articulatory movements. It also requires trying, by direct experiment, to find the perceptual consequences (for the listener) of various articulatory maneuvers (by the speaker). To do that we must of course press forward with the development of a research synthesizer designed to operate from articulatory, rather than acoustic, controls (Abramson, Nye, Henderson, & Marshall, 1981; Mermelstein, 1973; Rubin, Baer, & Mermelstein, 1981). The perfection of such a device, itself an achievement of some scientific consequence, will enable us to find a more accurate, elegant, and useful characterization of the informational basis for speech perception.

It will not have escaped notice that the claim to understanding I have made is, in any case, a modest one. At most, we presume to know something about what phonetic processes do and in what ways they are distinctive and coherent. As for mechanism, however, there is only the assumed link between perception and production, and even there we have no certain, or even clear, idea how such a link might be effected. If we knew more about mechanism, we would presumably be in a better position to design automatic speech recognizers of a nontrivial sort (Levinson & Liberman, 1981). At present, however, we can only claim to understand where the difficulties lie. That is an important step, to be sure, but it is only the first one, and it will almost surely prove to be the easiest.

Since I have taken the position that speech perception depends on biologically specialized processes, I should at last acknowledge that neurological and developmental studies are relevant. For if phonetic processes are distinctive and coherent from a perceptual point of view, we reasonably expect that they are so from a neurological point of view as well. We then look to neuropsychological data to provide further tests of our hypotheses, to refine our characterizations and indeed, to supply new insights into the processes themselves. As for the biology of the matter, we must rely heavily, of course, on developmental studies of speech perception, especially when these include very young infants and comparisons across languages. Such studies enlighten us about what might have developed by evolution in the history of the race and what remains to develop, presumably by epigenesis, in the history of the individual. Of course neither the neuropsychological nor the developmental studies will be useful unless we ask the right questions. But I believe we are learning to do that.

REFERENCE NOTES

1. Fodor, J. A. *The modularity of mind.* Unpublished manuscript, Massachusetts Institute of Technology, 1981.
2. Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. *Integration of time-varying cues and the effects of phonetic context.* Unpublished manuscript, Haskins Laboratories, 1981.
3. Lane, H. L., & Schneider, B. A. *Discriminative control of concurrent responses by the intensity, duration and relative onset time of auditory stimuli.* Unpublished manuscript, Behavior Analysis Laboratory, University of Michigan, 1963.
4. McGurk, H., & Buchanan, L. *Bimodal speech perception: Vision and hearing.* Unpublished manuscript, Department of Psychology, University of Surrey, England, 1981.

REFERENCES

Abramson, A. S., Nye, P. W., Henderson, J. B., & Marshall, C. W. Vowel height and the perception of consonantal na-

sality. *Journal of the Acoustical Society of America*, 1981, *70*, 329–339.

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

Bailey, P. J., Summerfield, Q., & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research*, 1977, *SR-51/52*, 1–25. (ERIC Document Reproduction Service No. ED 147 892)

Bastian, J., Delattre, P., & Liberman, A. M. Silent interval as a cue for the distinction between stops and semivowels in medial position. *Journal of the Acoustical Society of America*, 1959, *31*, 1568. (Abstract)

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

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics*, 1981, *29*, 26–36.

Cole, R. A., & Scott, B. Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, 1973, *27*, 441–449.

Cooper, F. S. Research on reading machines for the blind. In P. A. Zahl (Ed.), *Blindness: Modern approaches to the unseen environment*. Princeton, N.J.: Princeton University Press, 1950.

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

Cooper, F. S., Liberman, A. M., & Borst, J. M. The interconversion of audible and visible patterns as a basis for research on the perception of speech. *Proceedings of the National Academy of Sciences*, 1951, *37*, 318–327.

Dorman, M. F. On the identification of sine-wave analogues of CV syllables. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the Ninth International Congress of Phonetic Sciences* (Vol. 2). Copenhagen, Denmark: University of Copenhagen, 1979.

Dorman, M. F., Raphael, L. J., & Isenberg, D. Acoustic cues for fricative-affricate contrast in word-final position. *Journal of Phonetics*, 1980, *8*, 397–405.

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

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

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

Gerstman, L. J. *Perceptual dimensions for the friction portions of certain speech sounds*. Unpublished doctoral dissertation, New York University, 1957.

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

Harris, K. S., Hoffman, H. S., Liberman, A. M., & Delattre, P. C. Effect of third-formant transitions on the perception of the voiced consonants. *Journal of the Acoustical Society of America*, 1958, *30*, 122–126.

Isenberg, D., & Liberman, A. M. Speech and non-speech percepts from the same sound. *Journal of the Acoustical Society of America*, 1978, *64*(Suppl. 1), S20. (Abstract)

Klatt, D. H. Software for cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 1980, *67*, 971–995.

Levinson, S. E., & Liberman, M. Y. Speech recognition by computer. *Scientific American*, 1981, *244*, 64–76.

Liberman, A. M. The grammars of speech and language. *Cognitive Psychology*, 1970, *1*, 301–323.

Liberman, A. M. The specialization of the language hemisphere. In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program*. Cambridge, Mass.: MIT Press, 1974.

Liberman, A. M. Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the Ninth International Congress of Phonetic Sciences* (Vol. 2). Copenhagen, Denmark: University of Copenhagen, 1979.

Liberman, A. M., & Cooper, F. S. In search of the acoustic cues. In A. Valdman (Ed.), *Papers in phonetics and linguistics to the memory of Pierre Delattre*. The Hague, The Netherlands: Mouton, 1972.

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

Liberman, A. M., Delattre, P. C., & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497–516.

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

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. C., & Cooper, F. S. Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America*, 1959, *31*, 1490–1499.

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

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. W. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of sensory physiology*. Vol. 8: *Perception*. New York: Springer Verlag, 1978.

Lisker, L. Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research*, 1978, *SR-54*, 127–132. (ERIC Document Reproduction Service No. ED161096)

MacDonald, J., & McGurk, H. Visual influences on speech perception processes. *Perception & Psychophysics*, 1978, *24*, 253–257.

Mann, V. A. Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, 1980, *28*, 407–412.

Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. Further investigation into the influence of preceding liquids on stop consonant perception. *Journal of the Acoustical Society of America*, 1981, *69*(Suppl. 1), S91. (Abstract)

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

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

Mattingly, I. G. Phonetic representation and speech synthesis by rule. *Haskins Laboratories Status Report on Speech Research*, 1980, *SR-61*, 15–21. (ERIC Document Reproduction Service No. ED185636)

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

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

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

Mermelstein, P. Articulatory model for the study of speech

production. *Journal of the Acoustical Society of America*, 1973, *53*, 1070–1082.

Nye, P. W. Psychological factors limiting the rate of acceptance of audio stimuli. In L. L. Clark (Ed.), *Proceedings of the International Congress on Technology and Blindness*. New York: American Foundation for the Blind, 1963.

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

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

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

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

Repp, B. H. Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. *Perception & Psychophysics*, 1978, *24*, 471–485.

Repp, B. H. Two strategies in fricative discrimination. *Perception & Psychophysics*, 1981, *30*, 217–227.

Repp, B. H. Phonetic trading relationships and context effects: New experimental evidence for a speech mode of perception. *Haskins Laboratories Status Report on Speech Research, SR-67/68*, in press.

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

Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulations. *Journal of the Acoustical Society of America*, 1981, *69*, 1154–1163.

Rubin, P., Baer, T., & Mermelstein, P. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 1981, *70*, 320–328.

Stevens, K. N., & Blumstein, S. E. The search for invariant acoustic correlates for phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, N.J.: Erlbaum, 1981.

Strange, W., Jenkins, J. J., & Edman, T. R. Identification of vowels in vowel-less syllables. *Journal of the Acoustical Society of America*, 1977, *61*(Suppl. 1), S39. (Abstract)

Studdert-Kennedy, M. Speech perception. *Language and Speech*, 1980, *23*, 45–66.

Studdert-Kennedy, M., & Cooper, F. S. High-performance reading machines for the blind: Psychological problems, technical problems and status. In R. Dufton (Ed.), *Proceedings of the International Conference on Sensory Devices for the Blind*. London: St. Dunstan's, 1966.

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

Summerfield, Q., Bailey, P. J., Seton, J., & Dorman, M. F. Fricative envelope parameters and silent intervals in distinguishing 'slit' and 'split.' *Phonetica*, 1981, *38*, 181–192.

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