Reprint from

INFORMATION PROCESSING IN THE NERVOUS SYSTEM

Edited by K. N. Leibovic

The Speech Code and the Physiology of Language.

I.G.Mattingly & A. M. Liberman

NOT TO BE SOLD

Springer-Verlag
New York · Heidelberg · Berlin
1969

IGNATIUS G. MATTINGLY
Haskins Laboratories and University of Connecticut

ALVIN M. LIBERMAN
Haskins Laboratories, University of Connecticut, Yale University

6 · *The Speech Code and the Physiology of Language**

I. Introduction

To the physiologist who would study language in terms of the interests represented at this symposium, the most obvious linguistic processes—the selection of words to convey meaning and the arrangement of words in sentences—must seem far removed from familiar concepts and methods. Surely, he would prefer to study processes that are physiologically more accessible, but are yet linguistic. We believe that the production and perception of speech, in the narrow sense, is one such process; we suggest, therefore, that the physiologist might do well to start there. The questions we would have him ask can be put very simply: How does a speaker convert the phonetic units—the consonants and vowels—to a stream of sound? On hearing that stream, how does a listener recover the phonetic units?

For the purposes of this paper, we can do as well with one of those questions as with both; in principle, either one will do. We have chosen to deal primarily with the second—the one about speech perception—because we find more data there that speak to the points we want to make.

In the received view, speech perception in our narrow sense is thought to be neither very linguistic nor very interesting. Language is commonly supposed to be structured in levels (syntactic, phonological, phonetic) that represent successive recodings of the information. Each of these levels

* The preparation of this paper, and much of the research on which it is based, has been supported by grants and contracts from the National Institute of Child Health and Human Development, the Office of Naval Research, and the Veterans Administration. Earlier phases of our work were aided by grants from the Carnegie Corporation of New York and the National Science Foundation.

consists of simple units (words, phones) of some kind, organized into successively larger and more complex units (phrases, sentences; syllables, breath groups). It has been the business of the linguist to describe the rules by which the units are organized at each level, and to discover the code by which they are converted from one level to the next.

But the recoding is usually assumed to end with the phones, the empty units that lie at the lowest level of the whole linguistic structure; it is not supposed to include the process we want to talk about here: the conversion from phone to sound and from sound to phone.[1] The traditional view of this conversion is that it is by means of a simple alphabet, each phone being represented by a unit sound. On that view the relation between phone and sound is trivial and outside language proper; the linguist is interested in the sound alphabet only as a concrete base in which he can, when necessary, anchor his abstract assumptions.

An alternative view, which seems to us not very different in its consequences, is that the relationship between phone and sound, while indeed part of language, has a character too irregular to invite systematic description. This would appear to be the position of Chomsky and Miller (1963, p. 318) and Chomsky and Halle (1968, pp. 293–295). Though these investigators assume that a universal phonetics underlies speech production, they believe that speech perception depends not only on the acoustic properties of the signal but also on "the hearer's knowledge of the language as well as on a host of extra-grammatical factors," so that the perception and the sound cannot be rigorously related. They regard speech perception as a "heuristic" process in which a hypothesis about the speaker's utterance is suggested by a sampling of cues in the speech signal. That hypothesis is then tested and corrected by reference to all levels of the grammar—a task that can be carried out by successive iterations of an analysis-by-synthesis procedure—until a plausible reconstruction of the string of phones has been arrived at. But if it is sufficient to have recourse to higher grammatical levels in order to carry out a process that is now viewed as unruly rather than simple, then the process of speech perception must still be considered uninteresting. For an explanation of speech perception, the linguist is, in effect, diverted from serious consideration of speech and directed back to the higher linguistic levels.

Neither do these easy assumptions about speech pose problems of any

[1] The term "phone" is customarily used to mean the way a phonological segment is realized in a particular context. If one assumes that the perceptual, articulatory, and acoustic domains of speech are in one-to-one correspondence, then the phonological realization can be described equivalently as an acoustic event (or class), an articulatory event (or class), or a perceptual unit. But, as will be seen, we believe that these domains of speech are *not* in one-to-one correspondence. We will therefore use the term "phone" to mean the abstract unit which is the output of the phonology. This unit is, in effect, the elementary perceptual unit of speech. Our primary concern, in this paper, is to explore the relationship between the phone, so defined, and its acoustic representation.

special interest for the psycholinguist. If speech is, indeed, a simple alphabet, then the only requirement on its sounds is that they be spaced at comfortable distances apart, both in time and in some perceptual space, so as to be readily discriminated by the listener. If, on the other hand, speech perception is unruly, then the psycholinguist will do better to study the more orderly areas of language—syntax and phonology.

If all these assumptions about speech were true, the physiology of speech perception would be no different from the physiology of auditory perception in general. A physiologist who might be looking for new challenges would have no reason to study speech perception. And if he were, for some reason, already committed to the physiology of language, he would want to start his investigation at a higher level.

We will try to show, however, that the simple assumptions about speech are wrong. Phone and sound are linked by a recoding not different in principle from the more familiar recodings at higher linguistic levels. As a consequence, speech is, like the rest of language, well matched to man and part of his linguistic physiology.

II. Speech and Language as Codes

Linguistic communication requires that a string of phones be transmitted from one person to another. This cannot be done efficiently in any straightforward way. If, to take the simplest case, the phones were represented by an alphabet of discrete sounds, the temporal resolving power of the ear would set a low ceiling on the rate at which we could communicate. Morse "code," which is really an artificial sound alphabet, cannot be understood at rates much higher than five or six characters a second (Cooper, 1950). Many other sound alphabets have been developed, chiefly in the course of research on reading machines for the blind: the highest perceptual rates achieved, after long practice with highly motivated subjects, have been of the order of two characters a second (Freiburger and Murphy, 1961; Coffey, 1963; Studdert-Kennedy and Liberman, 1963; Nye, 1965, 1968; Studdert-Kennedy and Cooper, 1966). At rates of 20 or 30 characters a second, the subject can hardly separate the sounds, let alone identify them. Yet, we know that perception of natural speech at rates as high as 20 or even 30 phones per second is possible (Orr *et al.*, 1965).

To understand the remarkable speeds that can be achieved with speech, we must see that it provides a kind of parallel transmission: cues for two or more successive phones are carried simultaneously, and by the same acoustic event. Thus, the second-formant transition[2] typically carries essen-

[2] A formant is a resonance of the vocal tract represented by a relatively intense band of energy in the speech spectrum. Formants are numbered in order of their position on the frequency scale, the first being the lowest, the second the next higher, and so on. The first two or three carry the important linguistic information.

A formant transition is a rapid change in the frequency of a formant, reflecting articulatory movement.

tial cues about both consonant and adjacent vowel. (For a fuller account of the characteristics of speech perception referred to in this paper, see Liberman, *et al.*, 1967.) This parallel delivery of cues occurs in greater or less degree for most of the phones. As a consequence, the cues for what we hear as a series of discrete phones are typically organized as a syllabic unit in the sound stream. Indeed, this organization is the basis for the perceptual existence of the syllable.

We should note here that the syllable is by no means the largest acoustic structure. Syllables themselves are differentiated with respect to certain acoustic dimensions—duration, fundamental frequency, and intensity—and a series of syllables forms a longer pattern that is a cue for stress. A series of these stress patterns, in turn, forms a still longer acoustic pattern that cues the breath-group, the phonetic counterpart of a clause (Lieberman, 1967; Mattingly, 1968).

The organization of phones into syllables has a very important consequence for efficiency: the limit on rate is set now by the number of syllables per second, not by the number of phones. But the efficiency of speech is achieved at the expense of simplicity, since the parallel delivery of information that produces the syllable is possible only at the cost of a complex relation between the acoustic signal and the phonetic message. Speech is not a simple alphabet, then, but a difficult and demanding code. If a listener can nevertheless perceive speech, it is presumably because he possesses a special device powerful enough to decode the signal and recover the phones.[3]

The speech decoder must deal with several more or less closely correlated complications in the relation between sound and phone. One is that there are not, and, by the nature of the code, there cannot be, commutable acoustic segments that correspond to the phonetic segments. That is, there is no acoustic criterion by which one can delimit segments in the signal that correspond in number or structure to the segments in the message.

Not only are the cues for different phones carried simultaneously by the same acoustic event, but also the cues for the same phone are carried by different acoustic events at different times. A brief period of silence, or of weak low-frequency energy, signals a stop consonant, but which of the stops [b, d, g, p, t, k] is intended is cued by the character of the transition that begins the following vowel (Liberman, *et al.*, 1954; Delattre, *et al.*, 1955). When two vowels are separated by a medial consonant, the character

[3] Not all phones are encoded in the speech signal. Vowels produced in isolation, or in context at slow rates of speech, are represented alphabetically: there is an isolable acoustic segment that corresponds to the vowel phone, and there is no context-conditioned variation. Such phones might presumably be perceived without recourse to the speech decoder, and there are, indeed, striking differences in perception between these phones and those that are, like the stops, always encoded. This is true, in particular, of the characteristics of the speech mode—categorical perception and cerebral lateralization—to be discussed later. For an account of these differences the reader should see Liberman, *et al.*, 1967. Our concern in this paper will be primarily with the phones that are highly encoded and therefore truly linguistic.

of the transition from the first vowel to the consonant is determined not just by the consonant, but by the second vowel (Ohman, 1966); we would expect this transition to have some cue value for the second vowel. Similarly, in a sequence such as [stru] the acoustic quality of the [s] friction, which is a cue for the first phone, is also likely to be affected by a feature belonging to the fourth phone [u]; in that case we might expect to find the perception of the [u] cued to some extent by the characteristics of the [s] friction (Kozhevnikov and Chistovich, 1965, p. 124 ff.).

Parallel delivery of the cues produces yet another and equally serious complication in the relation between sound and phone: the same phone is signaled in different phonetic environments by acoustic cues that are vastly different. In the case of [d], for example, the essential acoustic cue is a formant transition. Before the vowel [i], that formant rises from 2200 to 2600 cps., but before [u], the formant falls from 1200 to 700 cps. (For a fuller account, see Liberman, *et al.*, 1967.) We should expect that the first of these would sound like a rising glissando on high pitches, the second like a falling glissando on low pitches. And when we take these formant transitions out of a speech context, that is, indeed, what we hear. But when the transitions are the only cues for [d] in the syllables [di] and [du], we hear the same initial segment [d]. The kind of context-conditioned variation exemplified here is found, usually in more extreme form, for almost all the consonants.

Within a given context, however, the perceptual boundaries between one phone and another are very well defined, despite the apparent variability of the acoustic signal. Consider, for example, the voiced stops [b, d, g], which are distinguished from each other by the direction and extent of the second-formant transition. If a listener is presented with a series of synthetic voiced stops for which the starting point of this transition is varied along the frequency scale, he will classify these stimuli consistently as [b], [d] or [g], but will be unable to distinguish one member of a class from another. This result is obtained quite generally with speech-like stimuli that differ only along an acoustic continuum that carries an essential encoded cue: a listener typically classifies such stimuli neatly as phones; if he is then asked to discriminate stimuli that lie close together on the continuum, he does very well in those parts of the continuum where his classifying responses indicated a boundary between phones, but very poorly elsewhere. This phenomenon, in which discrimination is no better than identification, has been called categorical perception. (Liberman, *et al.*, 1957; Liberman, *et al.*, 1961; Stevens, *et al.*, in press.)

Thus the phonetic decoding device appears to be capable of recovering the phones from the acoustic signal by (*a*) reassembling the essential cues, which are thoroughly intermixed in the sound stream, (*b*) correcting for an immense amount of contextual variation in the cues, and (*c*) sorting the cues categorically in a mode of perception that is very sensitive to

differences between phones and quite insensitive to differences among tokens of the same phone. We have called this device a decoder because the relation between phonetic message and acoustic signal is highly complex and apparently arbitrary. Yet, we began by noting that to most users of the code—even to students of language—the relationship is seemingly transparent and trivial. Evidently, the speakers of the language have readily available to them a key to the code, a set of rules for the perception of speech, though they are quite unaware of its existence.

We can gain some insight into the nature of the key if we look at the utterances [di] and [du] in articulatory terms. For [di] the tongue moves toward the alveolar ridge and remains there for an instant, closing the vocal tract. Then the tongue moves slightly downward, releasing the closure; meanwhile the lips are spreading. For [du] the tongue gesture is similar, except that the final movement of the tongue is backward as well as downward; meanwhile the lips are becoming rounded. In both utterances the vocal cords begin to vibrate at or slightly before the release, and it is only then that the acoustic signal begins. If we compare this articulatory account of the production of [di] and [du] with the acoustic account given above, we note that two similar gestures of an articulator can have very different acoustic consequences; that by no means all the movement of an articulator is reflected in the acoustic signal; that the acoustic signal is the complex resultant of the independent but concurrent action of several articulators; and that the parallel transmission of information is achieved both by this concurrent articulation and by the transitional movement of an articulator from one position to the next (Cooper, 1966).

It appears, then, that we can describe speech more simply in articulatory than in acoustic terms. This suggests that the speech code is organized according to articulatory gestures or, more likely, the commands to the muscles that make these gestures (Liberman, *et al.*, 1967). In effect, the key that the listener has available to him is an articulatory model that relates the phonetic message to the signal. The listener's model need represent only those features of articulation that are crucial to speech perception, but must represent those features in a way nonspecific enough to allow for a wide range of speaker variation. Moreover, since every language rings its own variations on the universal phonetic code, the model must be able to adapt to the version used in the listener's native language. Evidently, the model must be a very general and versatile one.

Let us briefly recapitulate the basic characteristics of the speech code. A message, consisting of a string of phones, is linked with a stream of cues embedded in an acoustic signal. These cues do not correspond straightforwardly to the phones; they are an encoding of the message. The arbitrary appearance of the code can be rationalized by reference to an articulatory model. Such a model, therefore, must be available to the listener, though he is quite unaware of the code or of the model that rationalizes it.

Two or more cues can be transmitted simultaneously; by such parallel transmission, the limitations set by the temporal resolving power of the ear are circumvented. Yet, simultaneous cues do not necessarily represent one phone, nor, conversely, are the cues for one phone simultaneous; hence no simple pairing can be made between successive phones and successive segments of the acoustic signal. The acoustic shape of a cue varies extensively with the context; yet there are sharp restrictions on this variation. Acoustically similar events carrying cues for different phones are categorically perceived. Finally, cues are organized into larger units: syllables, stress patterns, breath-group patterns.

If language and speech are controlled by the same neurophysiological apparatus, we should expect to find resemblances between the speech code and the linguistic code, i.e., the grammar of language. We do, indeed, find such resemblances. In fact, we find them two or three times over, because grammar consists of a series of structural levels, each linked to the next by a recoding process. Each of these recodings invites comparison with the speech code. The resemblances are partly obscured by more striking differences, but the differences are natural consequences of the different functions of the various grammatical recodings; the resemblances, we suggest, stem from the basic nature of the apparatus.

We can make the comparison clearer if we make use of a specific theory of grammar. Though the comparison would hold good in most of its details for almost any serious grammatical theory, let us, for argument's sake, use the generative grammar proposed by Chomsky and his colleagues (Chomsky, 1957, 1965; Chomsky and Miller, 1963).[4] According to this theory there is a linguistic level, "deep" structure, at which a string of grammatical and lexical morphemes, the latter represented by classificatory distinctive-feature matrices, is developed from more complex units, the structure taking the form of a labeled tree. One or more of these strings is processed by the rules of the transformational component, which nest, rearrange, and delete the morphemes of deep structure to produce a new string of morphemes at the level of "surface" structure. A tree structure is assigned to this string also. The surface structure is processed, in turn, by the phonological component of the grammar, which applies morphophonemic rules and converts the columns of the classificatory feature matrices—the phonemes—into columns of phonetic features—the phones. Because of the cyclic character of the phonological rules, the phones are organized into more complex units corresponding to gradations of stress and other prosodic features.

In generative grammar, then, there are two conversions, one syntactic,

[4] Lamb (1966) argues that the successive recodings of his "stratificational grammar" can be regarded as a neuro-physiological model. His examples are quite consistent with the points we make here about linguistic recoding. But for him the relationship between "phonon" and sound is of interest only to the physiologist of the vocal tract, not to the linguist or to the neurophysiologist concerned with language.

one phonological, relating three streams of information. In neither conversion is the relationship between the two related streams trivial or straightforward; each can reasonably be called a code. For each code, on the other hand, a model of some kind can be suggested which rationalizes what would otherwise appear arbitrary and eccentric. Generative grammarians have occupied themselves in devising such grammatical models: sets of syntactic or phonological rules that try to account in some elegant and economical way for the correspondence between the two streams of information linked by the code. The rules of the transformational component, together with the branching rules which organize deep structure, are a syntactic model: they explain the relationship between deep structure and surface structure. The rules of the phonological component, similarly, explain the relationship between the surface structure and the phonetic level.

Whatever the virtues and defects of particular models, we presume that models of some kind are available to the speaker-hearer in terms of which he "knows" the grammatical rules that relate the higher levels of linguistic information, just as we concluded that an articulatory model rationalizes for him the interconversion of phone and sound. Yet, the linguistically naive speaker is no more consciously aware of his competence in generative grammar than he is of his competence in phonetics.

In both of these grammatical codes we observe parallel transmission of information. The values in any one column of the phonetic feature matrices that are the output of the phonology may depend on the values in more than one column of the input classificatory matrices that represent the lexical morphemes; this is done by the application of allophone rules like the one that de-aspirates voiceless stops after initial /s/.[5] Grammatical morphemes are often combined in the phonetic representation, notably in the inflectional languages. Thus in Latin the values of the number and case categories (singular, plural; nominative, genitive, dative . . . . .), which are distinct grammatical morphemes in the input to the phonology, are represented together at the phonetic level by a set of unanalyzable suffixes—the declensional endings—each having both a case value and a number value. Similarly, in English, a tense and root combine in certain irregular verbs, e.g., "sing, sang, sung."

In the syntax, also, parallel processing is the norm; indeed, it is reasonable to suppose that one of the main purposes of the transformational component is to speed up communication by transmitting several deep-structure strings at the same time. Suppose there were no transformational component. Discourse would consist of a series of simple deep structure sentences like:

[5] A symbol set off by slashes / / represents a phoneme, or column of classificatory distinctive feature values, at the input to the phonology: a symbol in brackets [ ] represents a phone, or column of phonetic feature values, at the output of the phonology. Thus, the phoneme /t/ is recoded at the phonetic level as aspirated [t^h] in many contexts (such as *top*), but after /s/ by unaspirated [t] (as in *stop*).

The man sings.
The boy is tall.
The dog chased the cat.
The girl is blond.

If the deep structure were assumed to have rules for sentence nesting, a more complex string could be generated in cases where referential identity occurs. Thus the series

The man sings.
The man is tall.
The man married the girl.
The girl is blond.

(where the three occurrences of *man* refer to the same person; likewise the two occurrences of *girl*) would yield:

The man (the man (the man is tall) sings)
married the girl (the girl is blond).

But this complex string would surely take at least as long to process as the first series, since there are still four sentences to be dealt with. By means of deletion and substitution rules that exploit the referential identities, the transformational component condenses the nest of four sentences to:

The tall man who sings married the blond girl.

While this string is still four sentences at the deep structure level, it is only one sentence at the acoustic, phonetic, and surface structure levels, and so can be produced and perceived much faster. For the first series no such condensation is possible, but it is the second series, not the first, that is typical of the deep structure strings underlying ordinary discourse. There is usually a great deal of referential identity, which permits an enormous amount of transformational condensation and hence the transmission of a number of underlying strings at once. In this way, the limitation on the rate at which sentences can be processed by the brain is circumvented, much as the limitation on the temporal resolving power of the ear is circumvented by parallel transmission of different cues.

Just as in the case of the speech code, the price of parallel processing in syntax and phonology is a lack of simple correspondence between higher- and lower-level elements. As Chomsky (1957, pp. 38–40) has shown, we can only account for the various forms of the English verb by a transformational rule—the "auxiliary" transformation—that transposes the order of certain elements of the string to which it applies. Thus, the perfect and progressive are represented in deep structure by 'have + past participle' and 'be + present participle', respectively. By the auxiliary transformation

have + pp. + be + pres. p. + sing

becomes

have + be + pp. + sing + pres. p.

i.e., 'have been singing'.

In the phonological component, similarly, voicing of a syllable-final stop in the input results, at the phonetic level, in the lengthening of the preceding vowel. By this allophone rule, /bæd/ becomes [bæ:d]. Moreover, a phonologically voiceless stop is phonetically voiced post-vocalically before a low-stressed vowel. When both these allophone rules apply, the essential information distinguishing what are phonologically a voiced stop and a voiceless stop is displaced at the phonetic level to the preceding vowel. Thus /lædɝ/, 'ladder', becomes [læ:dɝ] and /lætɝ/, 'latter', becomes [lædɝ].

The two grammatical codes share with the speech code the property that, depending on context, an item of information in one stream may have various representations in the other stream that the speaker-hearer does not distinguish. The deep structure terminal string 'John is a fool' may develop into 'Is John a fool?', 'What is John?', 'Foolish John . . . ', or into part of 'John and Tom are fools' or of 'I consider John a fool', and so on. The morpheme which signifies plural, "s", may become [s], [z], or [iz], or any of several irregular forms. One of the distinctive features in the matrix representing /t/ is '-voiced', and normally /t/ is phonetically as well as phonologically voiceless. By a rule we have already given, however, /t/ may become phonetically [d] in a certain context.

The speaker-hearer copes equally well with all developments of 'John is a fool', responding only to the difference in grammatical context. He fails to notice the variation in the regular plural at all, and accepts the irregular forms as exactly equivalent. He takes the two allophones of /t/ (and several others as well) in stride; in the context specified by the rule, he will never notice the occurrence of the [d] variant. But he will notice the "normal" [t] allophone of /t/ if it should occur in this context, and wonder about the speaker's dialect.

The grammatical rules sharply restrict the range of variation, however, and in each case the restriction is categorical. If the categorical character of speech perception, which we described earlier, is not yet widely appreciated, the categorical nature of language has long been accepted as one of its most obvious properties. Sentences are active or passive, not something in between; noun phrases are singular or plural. Changing a phonetic feature, except in accordance with an allophone rule, changes a distinctive feature to its opposite and yields an entirely different morpheme, or no morpheme at all, not a similar morpheme. Presented with a set of varied syntactic or phonological items, the speaker-hearer unerringly deals with them categorically, just as he immediately perceives the phones [b, d, g], in categorical fashion. Categoricalness is thus found to be an important design feature of linguistic perception, from the level of the acoustic signal to the level of deep structure. This is the more interesting, because categorical behavior of this kind is not commonly found in human beings apart from their use of language.

Finally, we note that just as the basic coding units of the acoustic signal are organized into larger and more complex units, the three grammatical levels are also highly structured. A string at the level of deep structure can be represented as a labeled tree with lexical and grammatical morphemes at the ends of the branches and *N, NP, VP,* and *S* at the nodes, from which develop nouns, noun phrases, verb phrases, and sentences, respectively. The surface structure is represented by another branching tree. And since certain phonological rules, notably those relating to stress, are applied cyclically to longer and longer phonemic strings, the phonetic level also has, in effect, a tree structure.

From the formal point of view, then, there is good reason to regard the acoustic signal as another linguistic level, and the conversion from the phonetic message to the acoustic signal as a process comparable, in an important sense, to the conversions at higher linguistic levels. If this is so, then it is likely that speech and the various levels of grammar are processed by similar physiological apparatus.

III. Speech and Linguistic Physiology

To this point we have tried to establish speech as part of language by exposing certain formal similarities between the two. But the case does not depend entirely on such resemblances. There are data that point to more direct and concrete links; these suggest that the physiology of speech perception is not merely auditory, but also linguistic.

Consider, first, the tendency to categorical perception we described earlier. As will be recalled, this is a kind of perception in which discrimination is no better than absolute identification. To perceive the consonants categorically means that the listener identifies the several phones—for example, [b, d, g]—but cannot hear differences among the physically different tokens of the same phonetic type. The most easily measurable consequence of this kind of perception is that discrimination of equal physical differences will be better at phone boundaries (between [b] and [d], for example) than within a phone. When we measure discrimination of continuous changes in the essential acoustic cue, we do, indeed, find high peaks in the function at each phone boundary.

There are at least two broadly different interpretations of categorical perception. One is, that what is perceived categorically is the acoustic cue, not speech. In that case we should expect that the listener would perceive the second-formant transitions categorically, whether, in a speech context, they cue [b, d, g] or whether, outside that context, they do not. If that were so, we should conclude that categorical perception is a consequence of the way our auditory physiology processes certain kinds of acoustic stimuli. The opposite possibility is that what is perceived categorically is speech, not the acoustic cue. We should expect in that case that perception

of the acoustic variable would be categorical only when it cues a phonetic distinction, and we should conclude then that the mechanism underlying categorical perception is not auditory but linguistic.

A recent experiment by Mattingly, *et al.* (in press), provides data that help us decide between those alternatives. These investigators compared the discrimination of various second-formant transitions in speech and nonspeech contexts. In the speech case, the various second-formant transitions were part of simple, two-formant patterns, and served as the only acoustic cues on the basis of which these synthetic patterns could be heard as [bæ], [dæ], or [gæ]. In the nonspeech case, the second-formant transitions were presented alone and were heard as glissandi or else as chirp-like sounds. In both cases, the second-formant transitions were the only acoustic differences among the stimuli.

The results were quite clear. In the case of the speech patterns there were, as usual, high peaks of discriminability at the phone boundaries. Discrimination of the nonspeech controls, on the other hand, was very different. There were, in general, no peaks at locations corresponding to the phone boundaries. Such peaks as did occur were in positions different from those obtained with the speech patterns, and they were, in general, a good deal lower. Discrimination of the nonspeech stimuli was also, in contrast with the speech, quite variable in level, both between and within subjects.

Since the second-formant transitions are perceived categorically only when they are heard as speech, we should conclude that such perception is not merely auditory, but is also an aspect of our capacity for language. The incoming speech signal must, of course, first undergo some processing by the auditory system. What the experiment on categorical perception suggests is simply that a significant part of phonetic perception is carried out in the linguistic mode, and that speech is part of language in that very physiological sense.

That speech is part of the physiology of language is also suggested by the results of other recent experiments which show that phonetic perception, like language in general, tends to be located more on one side of the brain than the other. The first step was the finding, by Kimura and others, that when competing spoken digits are presented to the two ears, most listeners hear better the signal in the right ear (Kimura, 1961; Bryden, 1963; Broadbent and Gregory, 1964). When the stimuli are simple melodies or sonar signals, the opposite result is obtained—that is, the stimulus to the left ear is heard better (Kimura, 1964; Chaney and Webster, 1965). These results are thought to be relevant to cerebral lateralization because the representation of the ears in the cerebral hemispheres is presumably stronger contralaterally than ipsilaterally. The finding has then been taken to mean that the spoken digits are more easily processed in the left cerebral hemisphere where, as has long been known, linguistic functions tend to be located.

Music and sonar signals (and, presumably, many other complex nonspeech sounds) are processed by most people in the right hemisphere.

The experiments with the spoken digits were not conclusive, from our point of view, because these signals are meaningful and therefore require something more than simple phonetic perception. The next step was taken by Shankweiler and Studdert-Kennedy (1967a, 1967b). Using dichotically presented nonsense syllables that differed only in the initial stop consonant (for example, [ba] to one ear and [da] to the other), they found a significant right-ear (hence, left hemisphere) effect. We know, then, that phonetic perception, even when separated from syntax and meaning, is cerebrally lateralized; it is, moreover, on the same side of the brain as language, while music, and presumably many other complex nonspeech sounds, are on the other side.

The fact that phonetic and nonphonetic perception take place on opposite sides of the brain reinforces the view that they are carried out by different processes. That phonetic perception is on the same side with the rest of language suggests, further, that the difference between the phonetic and nonphonetic processes is related to the difference between language and nonlanguage. We shall have illuminated this matter still more when we know the outcome of an experiment on the ear effect that is similar to the experiment on categorical perception we described earlier. In the case of categorical perception, it will be recalled, we asked what it is that is perceived categorically: is it the auditory event corresponding to the acoustic cue, or is it speech? With respect to the ear effect, we should ask, similarly, whether the left hemisphere deals on a purely auditory basis with the particular kinds of cues that underlie phonetic perception, or whether it processes such cues only when they are part of speech. To answer that question, we should compare the lateralization of synthetic stop consonants, for example, that are cued only by differences in the second-formant transition, with the lateralization of those same second-formant transitions when they are presented in isolation and do not sound at all like speech. The experiment is now being carried out (Shankweiler, *et al.*, in progress).

If speech, in our narrow sense, is as much a part of language as we think, then it ought, like language, to be found only in man. We should assert that if it is reasonable to suppose that animals do not talk because they have nothing to say, it is at least as reasonable to suppose that they have nothing to say because they do not talk. In a recent study of several species of primates, Lieberman concluded that these animals do not produce a repertoire of speech-like sounds (Lieberman, 1968; Lieberman, *et al.*, in press). Vocalizations consist of a single [ə] vowel, characteristic of the vocal tract in neutral position, i.e., when its shape approximates that of a uniform tube. The vowel formants move, but they do so in exact proportion, indicating that the length of the vocal-tract tube changes because

of lip rounding or shifts in the position of the larynx, but that the shape remains uniform because the tongue is inert. Unlike the prelinguistic child, the monkey does not babble. Moreover, laryngeal excitation is quite irregular; there is little sign of the precise timing of phonation and aspiration so characteristic of human speech and so important for distinctions of phonemic significance (Lisker and Abramson, 1964; Abramson and Lisker, in press). The monkey's apparent incapacity to produce many speech sounds is due not just to his having a vocal apparatus that is less flexible in certain respects than that of a human being, but also to an inability to program in speech-like fashion the gestures of the vocal apparatus available to him. Apparently, the monkey's lack of the neural apparatus for language renders him incapable of speech as well.

Unfortunately, we know as yet almost nothing about the way animals perceive speech. We are reasonably sure that they do not understand speech in the usual sense, but one might suppose that this is only because they lack the machinery that comprises the semantic and grammatical components. If our view is correct, however, we should expect that animals would not perceive speech as we do, even at the phonetic level. Lacking a speech-sound processor, they are presumably unable to discover the segments of the message, or to hear as the same segment a consonant that appears in different vowel contexts and has, as a consequence, very different acoustic shapes. To the extent that the animal's auditory system is like ours, it should hear speech much as we hear the essential acoustic cues when they are sounded outside a speech-pattern context. On being presented with the syllables [di] and [du] cued only by the second-formant transition, the animal should perceive, not the unanalyzable linguistic event we human beings call [d], but instead the glissandi or chirps we hear when we listen to isolated second-formant transitions. If the animal's auditory physiology is significantly different from ours, then it should hear not glissandi or chirps, but something else, or, perhaps, nothing at all. But if it lacks the speech processor, as we suspect all nonhuman animals do, then, no matter what its auditory physiology, it should not hear [di] and [du] as utterances that begin with the same first segment.

It is possible, at least in principle, to determine experimentally how animals hear various kinds of speech sounds. This has not been done yet, at least not in such a way as to indicate whether they decode the sounds of speech as we do. If experiment should prove that they do not, then we shall have found that the conversion from sound to phone is, in a very deep biological sense, one with the rest of language.

IV. The Primacy of Speech

To support the assertion that speech is an integral part of language, we have used arguments based on data provided by experiments. But we

should note, if only briefly, that there is relevant evidence of a nonexperimental kind. Speech is the only universal vehicle of language; in contrast, reading and writing are recent, rare, and comparatively difficult for most human beings. Furthermore, language is acquired by the congenitally blind, but not by the congenitally deaf. We know well enough why a deaf child should have trouble learning to speak, but why, if speech is merely a way of transmitting language, should he find it so very difficult to learn to read and write? As a sensory channel, the eye is at least as good as the ear; why, then, are the optical shapes of an alphabet not easily substituted for the acoustical shapes of speech?

It is both tempting and easy to attribute such facts about the primacy of speech to conditions that have nothing to do with language. One thinks, for example, of the possibility that speech is better responded to than print because the child has no earlids to shut out the sound, or because he can hear speech no matter how his head is turned. But we believe that such peripheral considerations have little importance. In our view speech is the primary vehicle for language because both speech and language belong to the same special system.

V. Summary

We have tried to show that the interconversion of phone and sound is an integral part of language and of its underlying physiology. For that purpose we have considered three kinds of evidence.

The relation between phone and sound is that of a complex and efficient code bearing formal resemblances to the codes we know as syntax and phonology. Each of these codes is characterized by parallel transmission of information and by a consequent lack of direct correspondence between the elements of the linguistic levels that are linked by the code. Because we human beings have ready access to models that rationalize these codes, we are normally unaware that they might appear arbitrary or eccentric, or indeed, that they even exist.

Data from several experiments indicate that phonetic perception is not carried out entirely on an auditory basis, but rather requires the participation of mechanisms that are part of the linguistic system. Certain acoustic variables that are perceived categorically when they cue a phonetic distinction are not perceived categorically outside a speech context. Experiments with dichotically presented signals have shown that phonetic perception tends to be lateralized, like the rest of language, on the left side of the brain, while complex nonspeech sounds are processed primarily on the right.

It is a matter of common observation that speech has a privileged relation to language. Centuries of experience with reading and with the

problems of deaf children make it clear that no other vehicle for language is so natural or easy as the sounds of speech.

We conclude, then, that a knowledge of the mechanisms underlying the encoded relation between phone and sound would throw light on more general linguistic processes. Because such mechanisms are more accessible than those that govern the higher levels of grammar, they should be of special interest to the physiologist who cares about language.

REFERENCES

Abramson, A. and Lisker, L. (in press). Discrimination along the voicing continuum: cross-language tests. Proc. 6th Int. Cong. Phonetic Sci., Prague, 1967.

Broadbent, D. E. and Gregory, M. (1964), Accuracy of recognition for speech presented to the right and left ears. Quart. J. Exp. Psychol., *16:* 359–360.

Bryden, M. P. (1963). Ear preference in auditory perception. J. Exp. Psychol., *65:* 103–105.

Chaney, R. B. and Webster, J. C. (1965). Information in certain multidimensional acoustic signals. Report #1339, U.S. Navy Electronics Laboratory Reports, San Diego.

Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MIT Press.

Chomsky, N. and Halle, M. (1968). The sound pattern of English. New York-Evanston-London: Harper and Row.

Chomsky, N. and Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush and E. Galanter (eds.), Handbook of mathematical psychology. New York: John Wiley, *2:* 269–321.

Coffey, J. L. (1963). The development and evaluation of the Batelle aural reading device. Proc. Int. Cong. Tech. and Blindness I. New York: American Foundation for the Blind, 343–360.

Cooper, F. S. (1950). Research on reading machines for the blind. In P. A. Zahl (ed.), Blindness: modern approaches to the unseen environment. Princeton: Princeton University Press, pp. 512–543.

Cooper, F. S. (1966). Describing the speech process in motor command terms (abstract). J. Acoust. Soc. Amer., *39:* 1221 (Text: Status Report on Speech Research SR-5/6, Haskins Laboratories, New York, 1966).

Delattre, P. C., Liberman, A. M. and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Amer., *27:* 769–773.

Freiberger, J. and Murphy, E. F. (1961). Reading machines for the blind. IRE Professional Group on Human Factors in Electronics, *HFE-2:* 8–19.

Kimura, D. (1961). Cerebral dominance and perception of verbal stimuli. Canad. J. Psychol., *15:* 166–171.

Kimura, D. (1964). Left-right differences in the perception of melodies. Quart. J. Exper. Psychol., *16:* 355–358.

Kozhevnikov, V. A. and Chistovich, L. A. (1965). Rech' Artikuliatsia i vospriiatie. Moscow-Leningrad. (Trans. as Speech: articulation and perception. Washington: Joint Publications Research Service, 1965).

Lamb, S. (1966). Linguistic structure and the production and decoding of discourse. In E. C. Carterette (ed.), Brain function III speech, language and communication. Berkeley-Los Angeles: U.C.L.A. Press, 173–199.

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

Liberman, A. M., Delattre, P. C., Cooper, F. S. and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr., *68* (8, Whole, No. 379).

Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol., *54:* 358–368.

Liberman, A. M., Harris, K. S., Kinney, J. S. and Lane, H. (1961). The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol., *61:* 379–388.

Lieberman, P. (1967). Intonation, perception, and language. Cambridge: MIT Press.

Lieberman, P. (1968). Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer., *44:* 1574–1584.

Lieberman, P., Klatt, D. L. and Wilson, W. A. (in press). Vocal-tract limitations of the vowel repertoires of rhesus monkeys and other non-human primates. Status Report on Speech Research SR-15, Haskins Laboratories, New York.

Lisker, L. and Abramson, A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. Word, *20:* 384–422.

Mattingly, I. G. (1968). Synthesis by rule of general American English. Supplement to Status Report on Speech Research, Haskins Laboratories, New York.

Mattingly, I. G., Liberman, A. M., Syrdal, A. K. and Halwes, T. (in press). Discrimination of F2 transitions in speech context and in isolation (abstract). J. Acoust. Soc. Amer.

Nye, P. W. (1965). An investigation of audio outputs for a reading machine. Autonomics Division, National Physical Laboratory, Teddington, England.

Nye, P. W. (1968). Research on reading aids for the blind—a dilemma. Med. Biol. Engineering, *6:* 43–51.

Ohman, S. E. G. (1966). Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Amer., *39:* 151–168.

Orr, D. B., Friedman, H. L. and Williams, J. C. C. (1965). Trainability of listening comprehension of speeded discourse. J. Educ. Psychol., *56:* 148–156.

Shankweiler, D. and Studdert-Kennedy, M. (1967a). An analysis of perceptual confusions in identification of dichotically presented CVC syllables (abstract). J. Acoust. Soc. Amer., *41:* 1581 (Text: Status Report on Speech Research SR–10, Haskins Laboratories, New York, 1967).

Shankweiler, D. and Studdert-Kennedy, M. (1967b). Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol., *19:* 59–63.

Shankweiler, D., Syrdal, A. Halwes, T. and Liberman, A. M. (in progress). Left-right ear effects in the perception of second-formant transitions in and out of speech context.

Stevens, K. N., Liberman, A. M., Ohman, S. E. G. and Studdert-Kennedy, M.

(in press). Cross-language study of vowel discrimination. Language and Speech.

Studdert-Kennedy, M. and Cooper, F. S. (1966). High-performance reading machines for the blind: psychological problems, technological problems and status. Proc. St. Dunstan's Int. Conf. on Sensory Devices for the Blind (London), 317–342.

Studdert-Kennedy, M. and Liberman, A. M. (1963). Psychological considerations in the design of auditory displays for reading machines. Proc. Int. Cong. Techn. and Blindness I. New York: American Foundation for the Blind, 289–304.

DISCUSSION

HABER: Presumably the neural inputs to the speech center in the left hemisphere have undergone some decision processing since their input to the ear. What puzzles me is that a subject cannot tell many differences between auditory sounds in a forced choice discrimination and yet, somewhere else in the nervous system he can tell a great deal about speech sounds.

LIBERMAN: There are two different decisions involved here and I think they are orthogonal. One decision is whether it is speech or not speech. There are numerous cues on the basis of which we can make that decision though it is not known which of these cues are most effective. We do know, however, that the perception of these cues must be independent of the cues by which we distinguish one speech sound from another. Somewhere we make a decision whether the signal is speech or not and then begin processing it differently depending on the decision. Of course, if there is something wrong with the auditory system, one will not perceive speech; but being able to hear does not mean one will perceive speech.

SCHMITT: I have three questions. Firstly, would you comment on communication by whispering; secondly, can cues be separated and discriminated ear to ear?—and thirdly, I am surprised that there is no intensity parameter in the coding.

LIBERMAN: Let us take these points in reverse order: Intensity, for all practical purposes, has very little cue value. There is only one place I know where it is of very great importance, that is in the voicing distinction of fricatives, for example, [sa-za], where the intensity of the noise relative to the vocal portion is a significant cue. With stress, e.g., súbject or subjéct, you may think it is intensity, but it turns out it is really duration or pitch that is the important cue.

LONGUET-HIGGINS: Is it duration or is it larynx frequency?

LIBERMAN: For stress it is larynx frequency or duration or both, but not intensity. You can change intensity and the worst that happens is that you change your approximation to realism.

Now, with regard to the second question, we know that a listener will put dichotically presented cues together when they are separated in the frequency domain. For example, the first formant of a vowel put into one ear and the second formant into the other will fuse to give the vowel. If we split the cues in the time domain, the result is a good deal more complicated.

Finally, with regard to whisper, you produce the same formant pattern, whether you whisper or vocalize normally. Let me illustrate this with a steady state vowel. If I had my vocal tract shaped to produce the vowel sound *a* there would be formants—that is, bands of relatively intense energy—at about 700 cycles and 1,200 cycles. These are the first two formants of the vowel, and they are quite sufficient to give you *a*. Now, if I substitute whisper, I keep the same vocal-tract configuration, hence the same formants, but I have a different sound source. Now there is noise rather than discrete harmonics. As a consequence, the formants are filled with noise rather than harmonics.

SCHMITT: This does not apply when you substitute hydrogen breathing.

LIBERMAN: No, because that changes the velocity and hence the resonances. For example, speech in a helium atmosphere sounds like Mickey Mouse. The atmosphere does not change the fundamental, but it does change the resonances of the vocal-tract cavities and hence the frequencies of the formants.

CLYNES: The customary use of Fourier analysis to look at sound and hearing is inappropriate with respect to what really happens in the nervous system and what one hears. I will mention two aspects. As Dr. Liberman has also pointed out in terms of speech, the sensitivities of the ear to frequency and amplitude are quite different. In certain of our experiments we found that a barely noticeable change in amplitude was about 15 per cent, whereas, it was only 0.2 per cent in frequency. We found that the rate of change of frequency was very important. When we measured the evoked responses we found that at the vertex of the head a large, "unspecific" response failed to appear after small amplitude steps were replaced by large glissandos or siren-like sounds. It turned out that even a small (sustained) frequency slide inhibited the vertex response to subsequent large changes in pitch in either direction. The effective inhibiting threshold of the slide parallels the threshold for perceiving it as an unsteady or sliding pitch. This reaction

of the brain, which we have called R-M function (Rest-Motion) occurs once with each word, and lasts about 0.3 seconds with a peak latency of O.2 seconds. The response occurs when a steady tone (or silence) is changed into a sliding tone, but not when a sliding tone is changed into a steady tone.

The second point I would like to make concerns the mixture of qualities one hears with simple physical stimuli. If one takes two single electrical pulses of, say, 0.5 msec. width separated by 3 msec., say, one can get a properly damped speaker system to reproduce reasonably accurately two such acoustic pulses. One can now vary the time between these two pulses, say between 2 msec. and 5 msec. What one then hears are two sensations: a "knock" that stays the same, and a pitch sensation which changes as the time between pulses is varied. If one increases the number of pulses, the knock sensation is relatively reduced and the pitch sensation increased. The knock sensation is apparently inhibited if there are enough pulses. This, too, illustrates the highly non-linear behavior in which several modalities of hearing interact. The idea that "timbre" is "overtones" of Fourier analysis is clearly false. But what is perhaps even more surprising is that there should be a pitch sensation at all for only two pulses: this appears difficult, if not impossible, to reconcile with Békésy's theory of hearing.

BOYNTON: How long can you delay two sound segments relative to each other either in one ear or dichotically, before it becomes non-speech?

LIBERMAN: I do not remember the figure, but, not too long. Timing is important.

LONGUET-HIGGINS: Do children who have been dumb from birth have any great difficulty in acquiring a speech recognition ability?

LIBERMAN: There is a claim by Eric Lenneberg in his book *The Biological Foundation of Language* that there are such children. He cites one case of a boy who presumably does not speak and presumably understands speech. Since this boy has normal hearing, I suspect that, like most infants, he used to babble. I do not recall whether there was any evidence about that.

AXELROD: Is it possible to put the consonants in one ear and the vocalic components in the other and get dichotic integration?

LIBERMAN: We have not done that, but we did something similar some time ago when we were not yet able, as we are now, to control the timing very carefully. We took a synthetic version of a fricative-vowel syllable like *sag*, cut it, and put

the noise portion corresponding to the *s* noise in one ear and the remaining segment in the other ear. There was no question that we heard *stag*, not *sag*. However, if you play the second segment by itself you get *dag* and if you just separate the two segments by, say, 50 msec. you also get *stag* even though you put both segments into the same ear. Though I cannot be too sure, in view of possible timing errors in our experiment, I believe that we do not fuse across the ears with respect to time, only with respect to frequency.