Reprinted from PERCEPTION AND ITS DISORDERS Res. Publ. A.R.N.M.D., Vol. XLVIII.

Printed in the United States of America

Chapter XVI

SOME CHARACTERISTICS OF PERCEPTION IN THE SPEECH MODE[1]

A. M. LIBERMAN

To speak of "perception in the speech mode" is to imply, of course, that speech and its perception are somehow special. That implication would be readily accepted, I am sure, if my reference were to speech in the broad sense, because it is generally understood that language is different from other forms of communication. But here I mean to discuss speech in the narrow sense. My concern is not with the abstract matters of meaning or syntax, but with the very concrete sounds of speech and their raw perception as the phonetic or phonemic segments that we all know as consonants and vowels.

All agree that phoneme-like segments are perceived, and that this activity is, therefore, somehow to be reckoned with. But few share the view that I want to present here, which is that such perception is a special process and a true part of language. Perhaps it has been too difficult to see what the special problem is, or, as I think more likely, too easy to assume that none exists. Most listeners can, without difficulty, hear speech as a string of phonemes. They suppose, then, that speech is as simple as it sounds, and so conclude that it is merely a substitution cipher or alphabet on the phonemic message that it conveys.

According to that conventional wisdom, speech is a set of sounds, much like any others, that arbitrarily signal the phoneme segments of the language, each phoneme being represented alphabetically by a unit sound. The listener is required only to connect each sound with the name of the appropriate phoneme. Of course, the sounds must be discriminably different, but that can be regarded, and usually is, as a problem in auditory psychophysics, having little relevance to broader areas of perception, and nothing whatsoever to do with language or any other special systems. And if those who hold this view begin to suppose, as sooner or later most do, that the connection between sound and phoneme

[1] This chapter is based on the research and ideas of my colleagues at Haskins Laboratories. Our work has been aided 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 research were aided by grants from the Carnegie Corporation of New York and the National Science Foundation.

is established in experience, they have only made the easy discovery that speech is in some sense a habit, and thus raised again the familiar questions about reinforcement and other conditions of learning.

If those simple assumptions were true, that is, if the sounds of speech were a simple alphabet on the phonemes, then speech perception would be no different from auditory perception in general, and there would be no reason to assume the existence of a speech mode. But the simple assumptions are wrong. Worse yet, they are truly misleading because they take us away from speech before we have fairly begun to look at it. A proper study of speech makes plain that its perception is not to be understood by combining auditory psychophysics with a few simple principles of discrimination learning. Indeed, speech perception cannot be understood, even in some more complex way, so long as we regard it as an ordinary auditory process that transmits language but otherwise stands apart from it. The principles that govern the perception of speech are different in interesting ways from those of nonspeech, and they lie deep in the biology of language.

To discover that there is a speech mode and to see something of its characteristics, we should turn to research on the perception of speech. And that is what I mean to do. But first I would say briefly why we ought to suspect, quite apart from such research, that the received and simple view of speech is wrong because for that will help, I think, to show what the interesting problems of speech perception might be.[2]

We begin with the fact that all languages consist at base of a limited number of utterly meaningless segments called phonemes. These empty vehicles are the shortest segments of language, and they lie always in the lowest phonological layer. As I have already indicated, nonlinguists know them in their somewhat less abstract form as consonants and vowels.

It should be of particular interest that the phonemes of language are universally transmitted by the sounds of speech. In some languages this is also done visually with the optical shapes of an alphabet. But reading and writing are comparatively rare; they are a recent invention. Even in societies that put a high premium on literacy, there are many who listen and speak but cannot read or write. This is only to emphasize what all of us know but do not always appreciate: the sounds of speech are not merely one set of vehicles for the transmission of phonemes; they are, rather, the only universal and natural set.

But some would explain the privileged status of speech by attributing

[2] Most aspects of the discussion in this part of the chapter are presented in greater detail, and with appropriate references, in a recent article (17) by several members of the Haskins group.

it to factors that have nothing to do with language or the existence of a speech mode. Thus, it is sometimes implied that speech works well because the ear is a first-rate channel, especially well equipped to handle phoneme communication on some very simple basis. To see that this is not so requires no special knowledge of language or of auditory perception, but only a passing acquaintance with the requirements of phoneme communication and the best known properties of the ear. We become aware of what is, perhaps, only the most obvious difficulty by noting that in listening to speech we can take in as many as 30 phonemes per second. But we know, given the temporal resolving power of the ear, that a listener could not even resolve 30 discrete acoustic events per second, let alone identify them. For this reason, if for no other, we should suppose that the ear is ill adapted to so simple a vehicle as an acoustic alphabet.

More commonly the primacy of speech is attributed simply to the many years that we have all spent practicing it. It is difficult to see how any amount of practice could overcome limitations on sound alphabets set by such low level physiological constraints as the temporal resolving power of the ear. I should emphasize, nevertheless, that there is a great deal of evidence that practice, even large amounts of it, does not produce efficient perception of acoustic alphabets. This is clear, not only in the example of the Morse code, but even more convincingly, perhaps, in the repeatedly unsuccessful attempts to find nonspeech sounds that will work well as part of a reading machine for the blind. Many sound alphabets have been given a thorough trial, but none has proved adequate. It must surely give us pause to know that, although sounds are the universal carriers of language, only one set of sounds—those of speech—serves well. We ought to suspect, before doing any research of our own, that the speech signal is special in some interesting way, and that man has special mechanisms for dealing with it.

Motivated by the arguments that I have just outlined, my colleagues at Haskins Laboratories and I began almost 20 years ago to study the perception of speech (6, 7). Our goal was to discover why the sounds of speech are perceived so well. The first task was obvious enough and not different from that which confronts the scientist who sets out to study the perception of anything. It was to find the cues, that is, the physical stimuli, that control the perception. We chose to do this experimentally by using the spectrographic display as a basis for controlling and, in the extreme case, synthesizing speech. The method proved flexible and convenient, enough so that 6 years and many thousands of experiments later we had found the major acoustic cues for the segmental phonemes.[3]

[3] For reviews of the work at Haskins and elsewhere, see (8, 15–17, 31).

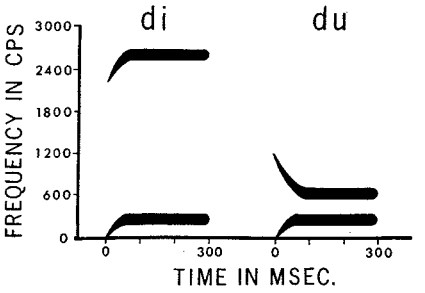

Fig. XVI.1. Spectrographic patterns sufficient to produce the syllables /di/ and /du/. (Reprinted from Psychol. Rev., *74:* 435, 1967.)

Much remains to be learned, even now, but we know enough to be able to synthesize speech from explicit rules (22, 26, 27) and to see the most important characteristics of the relation between the sounds of speech and the phonemic structure of language.

I try here, by way of an example, to describe what is special about the speech signal. In figure XVI.1 are drastically simplified spectrographic patterns that will, when converted to sound, produce the syllables /di/ and /du/. Each pattern consists of two bands of acoustic energy called "formants." At the left, or beginning, of each pattern the formants move rapidly through a range of frequencies. These rapid movements are called "transitions." When the transitions have run their course—a process that consumes about 50 msec.—the formants assume a steady state.

Consider, first, the steady-state vowels /i/ and /u/, although, as we see below, they are neither so special nor so interesting as the stop consonant /d/. The acoustic cues for these vowels are simply the positions of the formants on the frequency scale. As shown in the figure, /i/ is characterized by one formant at about 250 c.p.s. and another at 2500 c.p.s.; /u/ has the same first, or lower formant, but its second formant is at 700 c.p.s.

The relation between acoustic cue and steady-state vowel is quite simple. Although not apparent in the figure, the acoustic cue for the vowel will be the same in all phonemic contexts. Moreover, there is, between cue and phoneme, a perfect correspondence is segmentation. That is, we can isolate a piece of sound, in this case the steady state portion of the pattern, that is by itself sufficient to produce a single phoneme /i/ or /u/. These cues are, then, a simple alphabet or substitution cipher on the phonemes.

The cue for the stop consonant is more complicated. To isolate that cue, we should first notice, but then set aside, the transition of the lower

(first) formant. That transition, which rises from the bottom of the spectrogram and finds the appropriate steady state, is not specifically a cue for /d/; it rather tells the listener that the segment that he hears is some member of the class of voiced stops /bdg/. To produce /d/ in front of /i/ and /u/, we must add the transitions of the higher (second) formant. These, then, are the acoustic cues for /d/. But see how very different these transition cues are: one is an upgoing frequency modulation in the range 2200 to 2500 c.p.s., the other a downgoing modulation from 1200 to 700 c.p.s.

Another and perhaps greater complication in the relation between acoustic cue and stop consonant is that there is no segment in the sound stream that corresponds to the segment in the phonemic message. That is, there is no way to cut the sound pattern so as to obtain a piece that will produce the phoneme /d/ without also producing the next vowel or some reduced approximation to it.

Both characteristics of the stop consonant phoneme, the context-conditioned variation and the lack of correspondence in segmentation, are corollary aspects of a most important feature of speech: a single acoustic cue serves more than one segment. This is a kind of parallel transmission. In the case of our example we see in what way this occurs by observing that the second-formant transitions are, at every instant, carrying information simultaneously about two successive segments, the stop consonant and the vowel. It follows that the transition cue for the stop must be vastly different in different vowel contexts, and that there cannot be an acoustic segment corresponding only to the single phoneme /d/. Such cues are, quite exactly, a code on the phonemes, not a cipher or alphabet.

What we have seen here in the case of /d/ is characteristic of all of the consonants except, perhaps, the fricatives in slow articulation. The voiced and voiceless stops, the nasal consonants, liquids and semivowels are all complexly encoded in the sound stream (17). In most cases, indeed, the complications in the relation between sound and phoneme are far more severe than in our example of /d/.

The considerably simpler situation of the steady state vowels /i/ and /u/ is common to all of the vowels and fricatives in slow articulation. Those phonemes are enciphered, not encoded, in the sound stream, which is to say that the relation between phoneme and sound is one-to-one. (I should note parenthetically that at rapid rates of speech the vowels and fricatives may also become encoded or restructured to some degree.)

A salient and special characteristic of the speech code is, then, that information about successive phonemes is transmitted simultaneously

by the same acoustic cue. It is easy to see how this makes for perceptual efficiency, and, in particular, how it evades the severe limitations on rate of phoneme perception set by the temporal resolving power of the ear. The point is that parallel transmission reduces by a significant factor the number of discrete acoustic segments that must be perceived per unit time. Given that the phonemes are encoded acoustically into segments of syllabic size, the rate at which the sounds of speech merge into a buzz is set by the number of syllables per second, not by the number of phonemes.[4]

But this necessary advantage is bought at the cost of a very complex relation between the acoustic signals and the string of phonemes that they represent. In the case of restructured or encoded phonemes, such as the stop consonants, there are, as we saw, no commutable acoustic segments of phonemic size, and the cue for an encoded phoneme is vastly different in different contexts. If the complexly encoded sounds of speech are to be well perceived, indeed, if they are to be perceived at all, it can only be by means of a special decoder. We can barely guess at how that decoder might work, but we can say what it has got to do, and we can describe some characteristics of the mode of perception—the speech mode—that it produces.

Because speech can be perceived as an ordered string of phonemes, we must suppose that the speech decoder recovers the segmentation that was lost when the linguistic message was recoded as sound. To appreciate how hard it is to do that, one need only try to build an automatic speech recognizer or to read spectrographic representations of the speech signal (10, 18, 23, 24). The difficulty that has made it so far impossible to accomplish either of those tasks does not arise out of some obvious shortcomings of the engineer or of the spectrographic display; it is, rather, owing in significant measure to the fact that the message segments are thoroughly restructured in the sound stream and cannot be recovered in any trivially easy or obvious way.[5]

[4] It is possible that the encoding (and the parallel transmission) extends, in some cases and to some extent, across syllable boundaries. If so, there would, of course, be a further reduction in the number of acoustic segments required to transmit a given number of phonemes. At the present time, very little is known about this matter. For an experiment that shows the role of encoding in determining where syllable boundaries are perceived, see Malmberg (25).

[5] Able engineers have tried for many years to build machines that will perceive the phonetic content of speech, but none has succeeded in any important way. In contrast to their inability to cope with speech, engineers have done quite well with machines that read print. It is easier to build a print reader, in part, because print is a cipher and presents no problem of segmentation. There are other reasons, and these are also of some interest to us. With print one does not encounter the serious problems of context-

The speech-sound decoder must also take account of the context-conditioned variation that I spoke of earlier. That is, it must sort into the proper categories acoustic cues that vary greatly and even discontinuously as a function of context. Recall how different were the acoustic cues for /d/ in /di/ and /du/ and how different they looked. Yet the speech processor makes them sound the same.

The decoder must also make these very different cues sound like speech, that is, like /d/. Recall again the shape of cues: one was an upgoing frequency modulation rather high in the spectrum, the other a downgoing modulation considerably lower in the spectrum. If we isolate these acoustic cues and present them alone, that is, outside the speech context, they sound like auditory events, and we hear exactly what we should expect to hear: the one is an upgoing glissando on high pitches, the other a downgoing glissando on low pitches. But when we present and perceive exactly these same stimuli as speech, we cannot hear anything like a glissando, no matter how hard we try. What we hear when we listen to the encoded phonemes is speech, not sound; our perception does not directly mirror the external physical events that cause it, but rather yields at the lowest conscious level an irreducible linguistic segment called /d/. Thus, a characteristic of the speech mode is that the relevant auditory dimensions cannot be perceived. In that sense, perception is highly noniconic and abstract. But then noniconic perception of that kind may be characteristic of many physiological perceiving devices, for example, the bug detector of the frog, that respond to derived or patterned aspects of the stimulus environment.[6] If the frog could tell us what he experiences when a bug flies across his visual field, we might discover that he sees not motion but some unanalyzable event as remote from visual movement as /d/ is from an upgoing or downgoing variation in pitch.

conditioned variation that characterize so much of speech, and in the case of print the information-bearing elements are strong and clear, whereas in speech, as we will see below in this chapter, the essential cues are typically weak and poorly defined physically. It is surely of some biological and psychological interest that machines find it relatively easy to read print but impossible to perceive speech, whereas for us linguistically designed human beings the difficulty is exactly the other way around.

It is not possible to read spectrograms well for exactly the same reasons that it has not been possible to build a speech recognizer. Moreover, the difficulties that one meets in trying to read spectrograms do not appreciably diminish with experience or practice. The second formant transitions for /d/ look just as different to us now as they did when first we saw them 18 years ago. It would seem, then, that the speech decoder is linked only to an auditory input and cannot be made to work for the eye.

[6] It was Mr. Terry Halwes who suggested to me that categorical perception, which is discussed below in this chapter, might well be a consequence of processing by a feature detector. I have chosen to put that suggestion into a somewhat different framework.

I should remark here that conversion of the acoustic and auditory event into linguistic parameters produces an important, perhaps necessary, economy. To describe speech in acoustic terms is very expensive. For high fidelity, one needs about 70,000 bits per second, and for reasonable sentence intelligibility roughly half as much. We do not know what the rate becomes in auditory terms, but it is clearly quite high—up in the thousands—especially if we assume a frequency-volley mechanism through part of the speech range. Certainly, it is a great deal higher than the rate for the same message calculated in linguistic or phonetic terms, for that is never greater than about 50 bits per second. If, as we suspect, the signal is quickly recoded into phonemic parameters, that is, if the complex auditory signal is converted into /bæ/, /dæ/, /gæ/, for example, the saving in information rate is great and the load on nervous tissue is considerably reduced.

There is at least one other kind of operation that the speech processor may have to carry out, one that will not be at all apparent from the patterns in figure XVI.1. There, the information-bearing formants were shown as large and clear black lines on a white background. Real speech is not that clean. The acoustic energy is never concentrated so neatly in the linguistically important formants, but rather tends to be smeared untidily over the entire spectrum. In the case of the encoded phonemes, the essential acoustic cues are even more indeterminate, and for another reason. Thus, we know that the formant transitions, which are such important cues for the encoded stops, change frequency at a very rapid rate, scattering their energy in a prodigal way and creating a cue that is, in every physical sense, quite uncertain. These cues are more uncertain, surely, than those which underlie the perception of unencoded phonemes. Yet the encoded phonemes carry more than their share of the information load. We have, then, the paradox that what is most important linguistically is most indeterminate physically. If, in listening to speech, we are not much bothered by that circumstance, it is, perhaps, because our speech processor is somehow able to extract or attend to the important but indistinct parts of the signal that carry the linguistic information.

Now everything that I have so far described about the speech processor has been based on the results of just one set of research studies, those that uncovered the acoustic cues on which our perception of speech depends. There are other experimental data that point to the existence of a speech processor, and also tell us something about its properties. I would like to describe some data of that kind that are emerging from several related experiments now in progress.

The first experiment compares the perception of speech and nonspeech

and deals, more specifically, with a phenomenon that is typical of the speech mode, that is, of the encoded phonemes, a phenomenon that we have called categorical perception (17, 21, 32).

As you know, it is quite generally true that a listener discriminates many more stimuli than he can absolutely identify. We discriminate more than a thousand just noticeably different pitches, for example, but can ordinarily identify no more than half a dozen. Perception of the unencoded vowels is similar to that of pitch in that listeners discriminate many more vowel colors than they identify as phonemes. That is, they hear many differences within a phoneme class. The perception of encoded phonemes is very different, however: listeners discriminate very little better than they can identify absolutely, which is to say that they hear /bdg/ as categories and perceive almost nothing in the way of intraphonemic differences.

Let us look at an experiment on categorical perception being carried out by Ignatius Mattingly, Ann Syrdal, Terry Halwes and me (28). The stimuli are shown in figure XVI.2. At the left is a spectrographic representation of acoustic patterns adequate to synthesize the stop consonants /bdg/, and also a number of physically intermediate values on the acoustic continuum that connects them. The basic pattern is a syllable consisting of initial first- and second-formant transitions of about 40-msec.

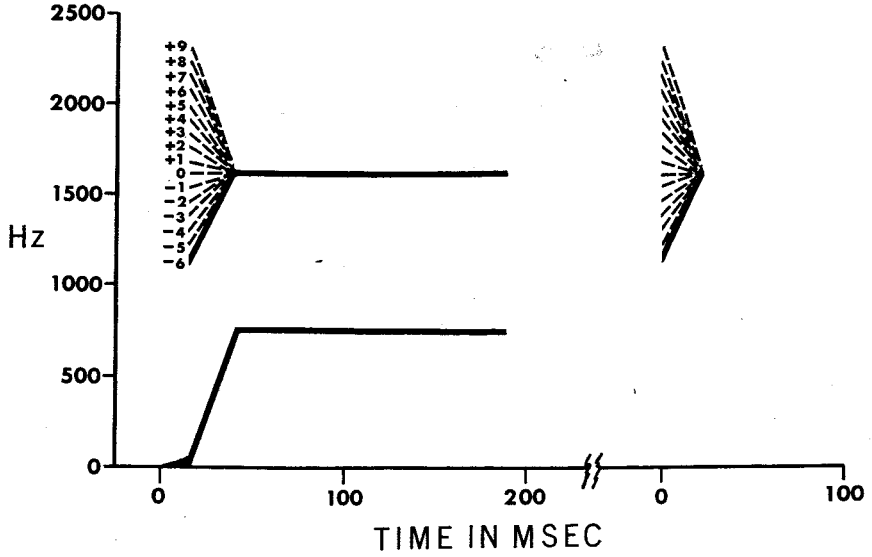

Fig. XVI.2. Schematic representation of patterns that will produce /bæ/, /dæ/ or /gæ/ by variations in the second formant transition, and, at the *right*, the second formant transitions alone.

duration, followed by steady-state formants that last 200 msec. Steady-state portions of the syllable are fixed at levels appropriate for the vowel /æ/. The first-formant transition is constant in all patterns. To see the experimental variable, one should look at the starting points of the second-formant transition. As shown in the figure, there are 16 such starting points, ranging from 1150 to 2310 cycles in steps of about 80 cycles. These starting points are labeled with reference to the one that corresponds in frequency to the steady-state level of the vowel; that one is called "zero" reference. When converted to sound by our computer-controlled formant synthesizer, the two-formant patterns are heard as /bæ/, /dæ/, /gæ/.

What it means to say, as I did, that these stimuli are perceived categorically, is most simply explained by telling you what a listener ordinarily hears when the stimuli are sounded in order from minus 6, the one whose second formant starts at the lowest frequency, to plus 9, the one whose second formant starts at the highest frequency. The first four or five stimuli are perceived as identical /bæ/'s; then the perception changes suddenly, that is, with the very next stimulus, to /dæ/. A similar discontinuity is encountered at the /dæ/-to-/gæ/ boundary. The point is that one's perception does not at all follow the step-by-step progression in the physical stimulus, but rather changes quantally from /bæ/ to /dæ/ to /gæ/.

As a measurable consequence of this tendency to categorical perception, we should expect that discrimination of equal physical differences would be better at phoneme boundaries than within a phoneme class. That is, when we measure discrimination along the continuum of second-formant transitions, we should find relatively high peaks in the function at each phoneme boundary. To measure discrimination of these stimuli, we use a simple forced choice technique. For every pair of stimuli to be compared, we arrange groups of three stimuli such that each group contains two tokens of the one stimulus and one of the other. The listener's task is to determine which of the three is different in any way from the other two. Of course, we use all six permutations of the two types taken three at a time; we do this for pairs of stimuli all along the continuum; and we present all of these stimulus triads many times in several randomized orders.

To find out how well one listener (H.C.) heard the various pairs of stimuli as different, we should look at the discrimination function at the top left of figure XVI.3. Paying attention to the solid line, and ignoring for the moment the dashed one, we see that the level of discrimination of equal physical differences is quite different for different parts of the continuum of second-formant transitions. In general, discrimination is

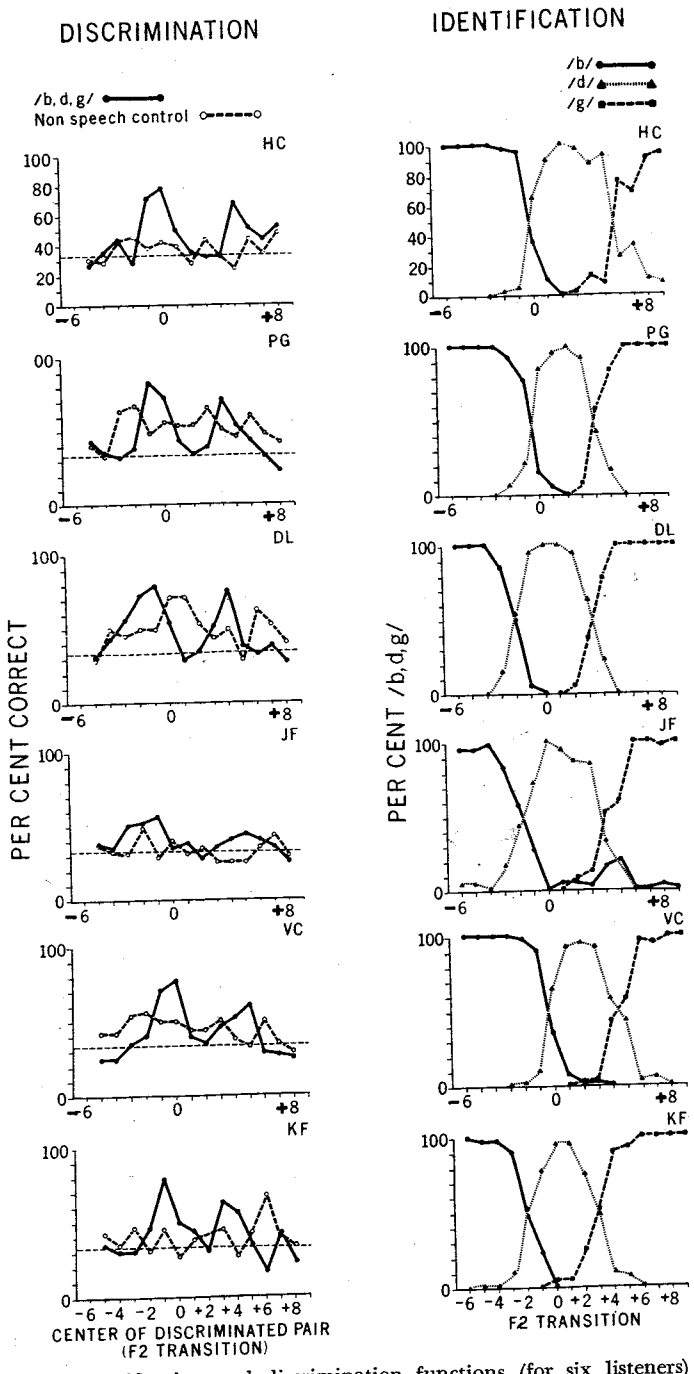

Fig. XVI.3. Identification and discrimination functions (for six listeners) obtained with synthetic /bæ/, /dæ, /gæ/ that are cued by variations in the second format transition, and discrimination functions obtained with the second formant transitions alone (as nonspeech controls).

relatively poor except at two points. That is, there are two rather high peaks separated by low troughs. Indeed, the troughs lie at or near chance, which is indicated by the dashed line running straight across the graph at 33 per cent.

At the top right of figure XVI.3 we see how this listener sorted the synthetic speech stimuli into phoneme classes. To obtain these data, we presented the 16 synthetic speech patterns in random order and asked the listener to identify each one as /b/, /d/or /g/. Comparing the discrimination function at the bottom with the phoneme identification functions at the top, we see that the peaks in discrimination occur at the crossovers of the phoneme identification curves. That is, the discrimination peaks occur at the phoneme boundaries.

What we have seen in the data of this one subject is an aspect of the categorical perception that we have found so many times in studies of the encoded stops (1, 2, 3, 19–21). It should be noted, parenthetically, that the unencoded vowels yield a different result, one more like that of nonspeech (9, 11, 32).

Consider, now, two broad interpretations of categorical perception. One is that what we perceive categorically is the acoustic cue, not speech. In that case, we should perceive the second-formant transitions the same way, whether they cue /b/, /d/, /g/, or whether, in some nonspeech context, they do not. If that were so, we should suppose that the speech processor is specialized, not for speech, but for the cues by which we perceive it. That processor would not discriminate between speech and nonspeech, and would most properly be regarded as an aspect or extension of the auditory system.

The other possibility is that perception of the relevant acoustic variation is quite different in a speech context from what it is in nonspeech. In that case, we should suppose that the speech processor belongs more properly to the mechanisms that underlie our use of language, and we should have to assume that the listener separates that which is linguistic from that which is not.

It will help us to decide between these alternatives to measure the discrimination of the essential acoustic cue—in this case the second-formant transition—when it is not in a speech context. Return, now, to figure XVI.2. On the left are the two-formant patterns that are heard categorically as /bæ/, /dæ/ or /gæ/. On the right are stimuli that we refer to as the nonspeech controls. As you see, they are exactly the same second-formant transitions that constitute the only acoustic cues in the speech patterns at the left. The single difference between the two sets—speech and nonspeech—is that in the nonspeech the constant steady state and the constant first formant are omitted. Thus, each set contains

all and only the differential cues found in the other. When the second formant transitions—the nonspeech controls at the right—are presented alone, they do not sound like speech at all, but rather like glissandos or the chirps of a bird.

Return to the graph at the top left of figure XVI.3. We looked at the discrimination function for the speech stimuli, which is represented by the solid line, and commented on the relatively high peaks at the phoneme boundaries. In the second condition of the experiment, discrimination of the nonspeech controls was measured by the same procedure that had been applied to the speech patterns. The result is represented by the dashed line. For the one listener (H.C.) whose speech data we examined earlier, we see that the nonspeech controls are discriminated at a low level throughout. There are no very high peaks, and certainly none in regions corresponding to the phoneme boundaries.

The results obtained with the other subjects so far run are shown in the several rows of figure XVI.3. We see that discrimination functions for the speech sounds quite consistently show peaks at the phoneme boundries. When the second-formant transitions are presented alone, very different results are obtained. There are no peaks in discrimination in positions corresponding to the phoneme boundaries. Such peaks as do occur are generally lower than those found with the speech stimuli; moreover, there is great variability among subjects in the height and position of these peaks.

These data are from an experiment that must be regarded as preliminary. We have been able to remove a source of random variation in the stimuli, and we are now, with these improved stimuli, more systematically investigating discrimination of the speech signals and the nonspeech controls, with particular attention to the parameters of syllable position and intensity. On the basis of the preliminary results just seen, however, certain conclusions relevant to this chapter can be drawn. When listeners discriminate tokens of /bdg/ that are cued entirely by transitions of the second formant, rather high peaks in discrimination appear at the phoneme boundaries. As I said above, this is a result that has been obtained in several previous experiments on perception of the encoded phonemes. What is added in this experiment is evidence that these peaks are not inherent in the perception of the second-formant transitions themselves. We should suppose, therefore, that the mechanism underlying these peaks works only on signals that are perceived as speech and not, more generally, on the corresponding auditory events.

So much, then, for categorical perception. I should like to mention briefly just one more set of findings because they point to a physiological basis for the assumption that there is a special speech mode. These

findings began in the work of Kimura and others (4, 5, 12, 14), who found that, when competing spoken digits were presented to the two ears, most listeners heard better the signal in the right ear. On the assumption that the representation of these signals in the brain is stronger contralaterally than ipsilaterally, this result has been taken to mean that the meaningful words of these experiments want to be processed in the left cerebral hemisphere. Kimura (13) then found that melodies give a left-ear effect; that is, they want to be processed in the right hemisphere. These results were, of course, of great interest to us and raised many questions in our minds. Thus, we wondered, first, whether this result could be obtained, not only with the meaningful words Kimura used, but with meaningless syllables that differ in only one phonemic segment. Shankweiler and Studdert-Kennedy (29, 30) did the appropriate experiments and found that the right-ear effect does, indeed, occur with such simple stimuli, showing thus that the effect is characteristic of the encoded consonants at the lowest level of phonetic perception. They also found that there is a much smaller right-ear effect for the unencoded vowels. Apparently, the encoded stops require the speech processor, which is, in most people, in the left hemisphere, but the unencoded vowels do not.

It is appropriate to ask about the right-ear effect essentially the same question that we raised in the experiment on categorical perception that I described above. There, as you will recall, we wondered what it is that is perceived categorically. Is it the acoustic cue as such, that is, the second formant transition, or is it encoded speech? In the case of the right-ear effect we wonder what it is that is processed by the speech decoder in the left hemisphere. Is it the stop consonants as a class of encoded speech sounds, or is it the acoustic cue—again the second formant transition—by which they are perceived? To answer this question, Shankweiler, Halwes, Ann Syrdal and I have just begun to do a dichotic experiment very similar to the discrimination experiment that I described above. In the one condition we present three two-formant patterns, selected from those of figure XVI.2, that are heard as /bæ/, /dæ/ and /gæ/. In the other condition, we present the second-formant transitions alone. As in the earlier experiment on categorical perception, these second-formant transitions are the only acoustic bases on which the listeners can distinguish the synthetic /bæ/, /dæ/, /gæ/ of the first condition. The outcome of this experiment will be of interest to us for obvious reasons. If, in the one case, we should discover that lateralization of the speech signal is different from that of the essential cue, we shall have found that what is processed in the left hemisphere is not the auditory event as such, but rather speech. This would be a still

further indication that the speech processor is part of the language system.

Perhaps I should summarize what I have meant to say. Having uncovered the acoustic cues, we are sure that the sounds of speech are nothing like an alphabet, but rather represent a complex restructuring of the phonemic message. This is a necessary part of an encoding, a grammer if you will, that matches speech to man and makes it a uniquely efficient vehicle. The chief characteristic of this grammar or code is parallel transmission: information about successive phonemes is carried simultaneously on the same acoustic cue. Such an arrangement makes for rapid communication by avoiding the limitations on rate of segment perception set by the temporal resolving power of the ear. But this important advantage is gained at the cost of a very complex relation between speech and the phonemic message. Speech perception therefore requires a special decoder. Knowing, as we do, the general characteristics of the speech code, we can say what it is that the decoder must do: it must recover the segmentation of the original message; it must compensate for a vast amount of context-conditioned variation; and it must somehow find and track those relatively indistinct parts of the acoustical signal that typically carry the most important linguistic information. In addition, the decoder apparently restructures the speech signal, with the result that the listener hears an unanalyzable linguistic event and cannot perceive the relevant auditory dimensions. In that sense, the decoder produces perception that is noniconic and abstract. There is evidence that the perceptual output of the decoder is also highly categorical. Research now in progress suggests that the decoder categorizes the acoustic cues only when they are heard as speech and not when they are presented and perceived as nonspeech. If that is so, then we should suppose that the decoder is not merely an extension of our auditory system, but is, more properly, an integral part of the mechanisms that underlie our use of language. Experiments on dichotic listening indicate that the encoded cues are processed in a part of the brain different from that used by the unencoded cues and by nonspeech, thus providing further and more directly physiological evidence for the existence of a device that is specialized to perceive the sounds of speech. On these bases we say of speech that it is, by comparison with other sounds, a special signal, processed by special devices, and perceived in a special mode.

REFERENCES

1. Abramson, A. and Lisker, L.: Discriminability along the voicing continuum: Cross-language tests. Proceedings of the VI International Congress of Phonetic Sciences, Prague, 1967.

2. Abramson, A. and Lisker, L.: Voice timing: Cross-language experiments in identification and discrimination. J. Acoustical Soc. Am., *44:* 377 (A), 1968.
3. Bastian, J., Eimas, P. D. and Liberman, A. M.: Identification and discrimination of a phonemic contrast induced by silent interval. J. Acoustical Soc. Am., *33:* 842, 1961.
4. Broadbent, D. E. and Gregory, M.: Accuracy of recognition for speech presented to the right and left ears. Quart. J. Exper. Psychol., *16:* 359, 1964.
5. Bryden, M. P.: Ear preference in auditory perception. J. Exper. Psychol., *65:* 103, 1963.
6. Cooper, F. S.: Spectrum analysis. J. Acoustical Soc. Am., *22:* 761, 1950.
7. Cooper, F. S., Liberman, A. M. and Borst, J. M.: The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc. Nat. Acad. Sc., *37:* 318, 1951.
8. Delattre, P. C.: Les indices acoustiques de la parole: Premier rapport. Phonetica, *2:* 108, 1958.
9. Eimas, P. D.: The relation between identification and discrimination along speech and nonspeech continua. Language and Speech, *6:* 206, 1963.
10. Fant, C. G. M.: Descriptive analysis of the acoustic aspects of speech. Logos, *5:* 3, 1962.
11. Fry, D. B., Abramson, A. S., Eimas, P. D. and Liberman, A. M.: The identification and discrimination of synthetic vowels. Language and Speech, *5:* 171, 1962.
12. Kimura, D.: Cerebral dominance and perception of verbal stimuli. Canad. J. Psychol., *15:* 166, 1961.
13. Kimura, D.: Left-right differences in the perception of melodies. Quart. J. Exper. Psychol., *16:* 355, 1964.
14. Kimura, D.: Functional asymmetry of the brain in dichotic listening. Cortex, *3:* 163, 1967.
15. Kozhevnikov, V. A. and Chistovich, L. A.: Rech'Artikuliastsia i vospriiatie, Moscow-Leningrad, 1965. (Speech: Articulation and Perception. Joint Public Research Service, Washington, *30:* 543.)
16. Liberman, A. M.: Some results of research on speech perception. J. Acoustical Soc. Am., *29:* 117, 1957.
17. Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M.: Perception of the speech code. Psychol. Rev., *74:* 431, 1967.
18. Liberman, A. M., Cooper, F. S. and Studdert-Kennedy, M.: Why are spectrograms hard to read? Ann. Deaf, *113:* 127, 1967.
19. Liberman, A. M., Harris, K. S., Eimas, P. D., Lisker, L. and Bastian, J.: An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. Language and Speech, *4:* 175, 1961.
20. Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C.: The discrimination of speech sounds within and across phoneme boundaries. J. Exper. Psychol., *54:* 358, 1957.
21. Liberman, A. M., Harris, K. S., Kinney, J. A. and Lane, H.: The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exper. Psychol., *61:* 379, 1961.
22. Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. C. and Cooper, F. S.: Minimal rules for synthesizing speech. J. Acoustical Soc. Am., *31:* 1490, 1959.
23. Lindgren, N.: Machine recognition of human language. II. Theoretical models of speech perception and language. IEEE Spectrum, *2:* 44, 1965.
24. Lindgren, N.: Machine recognition of human language. I. Automatic speech recognition. IEEE Spectrum, *2:* 114, 1968.

25. MALMBERG, B.: The phonetic basis for syllable division. Studia Linguistica, *2:* 80, 1955.
26. MATTINGLY, I. G.: Experimental methods for speech synthesis by rule. IEEE Tr. Audio & Electroacoustics, *16:* 198, 1968.
27. MATTINGLY, I. G.: Synthesis by rule of general American English. Supplement to Status Report on Speech Research. Haskins Laboratories, New York, 1968.
28. MATTINGLY, I. G., LIBERMAN, A. M., SYRDAL, A. K. AND HALWES, T.: Discrimination of F2 transitions in speech context and in isolation (abstract) *J. Acoustical Soc. Am., 45:* 314, 1969.
29. SHANKWEILER, D. P. AND STUDDERT-KENNEDY, M.: An analysis of perceptual confusions in identification of dichotically presented CVC syllables (abstract). J. Acoustical Soc. Am., *41:* 1581, 1967.
30. SHANKWEILER, D. P. AND STUDDERT-KENNEDY, M.: Identification of consonants and vowels presented to left and right ears. Quart. J. Exper. Psychol., *19:* 59, 1967.
31. STEVENS, K. N. AND HOUSE, A. S.: Speech perception. *In* Foundations of Modern Auditory Theory, edited by J. Tobias and E. Schubert. Academic Press, New York, (in press).
32. STEVENS, K. N., LIBERMAN, A. M., OHMAN, S. E. G. AND STUDDERT-KENNEDY, M.: Cross-language study of vowel discrimination. Language and Speech, *12:* 1, 1969.