Some Results of Research on Speech Perception*

ALVIN M. LIBERMAN
Haskins Laboratories, New York, New York and University of Connecticut, Storrs, Connecticut
(Received October 1, 1956)

Recent experiments with synthetic speech have succeeded in isolating some of the acoustic cues which underlie the perception of speech. This paper describes, and attempts to interpret, some of the research in that area.

I HOPE it is in keeping with the purpose of this Conference for me to talk mostly about the work of my colleagues at the Haskins Laboratories. I hope, too, that it is appropriate not to be concerned with specific experiments and detailed results, but rather to speak in very general terms about some of our findings, and then to discuss a few interpretations that are currently much on our minds.

As some of you know, we of the Haskins group have been trying to learn something about the perception of speech, and to that end we have spent a rather large amount of our time searching for the acoustic cues on which that perception depends. We have, we think, been quite catholic in our choice and use of methods. In the main, however, we have relied on techniques which enable us to make controlled changes in various aspects of the acoustic pattern, and then to evaluate the effects of those changes on the sound as heard. For that rather general, and certainly not uncommon purpose, we have depended largely on the use of spectrographic displays as a basis for synthesizing and modifying the sounds of speech. The essential point of this synthesizing technique is that we paint our own spectrograms, using a highly simplified form which omits many of the constant accompaniments—I should like to say stigmata—of speech. It is very easy, then, to introduce a wide range of experimental changes in what we suspect are important parts of the spectrographic pattern. When we listen to these spectrograms —having first, of course, converted them to sound—we find out which of these changes are important, and which are not. The conversion from spectrogram to sound is accomplished by a machine, called a pattern playback, which has been described and demonstrated at previous M.I.T. speech conferences.[1,2]

This particular way of synthesizing speech has given us easy access to many of the significant parts of the acoustic pattern, and has proved in general to be remarkably convenient and flexible. On occasion, however, we have found it desirable to synthesize speech by other means. In our investigations of the fricatives,[3] for example, a synthesizer we call "Octopus"[4] proved to have certain advantages over the playback.

For some purposes we have used experimental techniques other than those of synthetic speech. Thus, like many other investigators, we have modified speech by cutting and rearranging sections of magnetic tape, or by putting speech through filters. And, of course, before beginning any experimental work we have always wanted to examine the complex sounds of speech to see what we might profitably experiment with. For that purpose we have found spectrograms indispensable, though we have even made oscillograms, and, in occasional desperate moments, studied them.

By these means we have experimented with many aspects of the speech wave, and have succeeded, we think, in isolating some of the cues that carry the basic linguistic information. I will not attempt here to describe either the cues or their effects in detail. Much of this is to be found in papers that have already been published or that are about to be published, and we hope soon to undertake a comprehensive and detailed review of our findings. It will, however, be appropriate to the purposes of this paper to outline the types of acoustic cues—the acoustic stimulus dimensions, if you will—that we have so far found to be of some importance in the perception of the individual consonants of American English, and to offer a few examples of our results.

As a matter of convenience, I should like at the outset to divide the consonant cues into three classes, and to make this division according to where and how the sounds are produced. I am, of course, embarrassed to introduce a discussion of acoustic cues by classifying them on an articulatory basis. However, we find here, as we so often do, that it simplifies our data quite considerably to organize them by articulatory criteria. We certainly do not mean to imply by this that there are no acoustic differences among our classes, but only that it is hard to characterize these differences very simply in acoustic terms. For the purposes of this paper

* This paper was read, substantially as it is presented here, before the Conference on Speech Communication at the Massachusetts Institute of Technology on June 16, 1956. Apart from the particular form of the exposition—for which the author must bear sole responsibility—this paper should be regarded as a joint effort of the staff of Haskins Laboratories. The work of the Haskins Laboratories which is described in this paper has been supported in part by the Carnegie Corporation of New York and in part by the Department of Defense in connection with Contract DA49-170-sc-1642.

[1] F. S. Cooper, J. Acoust. Soc. Am. **22**, 761–762 (1950).

[2] Cooper, Delattre, Liberman, Borst, and Gerstman, J. Acoust. Soc. Am. **24**, 597–606 (1952).

[3] K. S. Harris, J. Acoust. Soc. Am. **28**, 160 (A) (1956).

[4] Meeks, Borst, and Cooper, J. Acoust. Soc. Am. **26**, 137 (A) (1954).

it would be footless to worry about this difficulty, so we will try, at least for now, simply to avoid the issue.

One class of consonant cues occurs in sounds that are produced at the consonant constriction. This class includes the frictions of the fricatives and affricates /f,θ,s,ʃ,tʃ,v,ð,z,ʒ,dʒ/ and the bursts of the stops /p,t,k,b,d,g/. It is a general characteristic of the constriction sounds that they are produced only during or just following the most nearly closed part of the consonant articulation. As the articulatory movement proceeds toward the more open position of the vowel, or, more generally, toward the next phone, the constriction sounds must quickly die away. Therefore, they would be expected to reflect little, if any, of the consonant movement, and in this respect the constriction sounds are to be contrasted with some other consonant cues that we will want to discuss shortly.

Among the constriction sounds we have found several types or dimensions of acoustic variation to be of importance for consonant perception. In one of our earliest experiments[5] we found that the frequency position of the burst enables listeners to distinguish among the voiceless stops /p,t,k/. More recently, it has been found that the frequency location of the friction noise, and in particular the lower frequency limit of this noise, is an overriding cue for distinguishing /s/ from /ʃ/, though this variable does not seem to contribute much to the perception of the other, less intense fricatives /f/ and /θ/.[6] In general it would appear that the frequencies of the constriction sounds provide the listener with significant information about the place of production of the stops and some of the fricatives.

Within this same group of constriction sounds, other dimensions of variation such as duration and the nature of the onset of the noise are proving to be of some importance, primarily as cues for various distinctions according to manner.[7] Even intensity appears to have some cue value in distinguishing /s,ʃ/ as a class from /f,θ/.[3] I say "even" intensity because we do not often find that intensity differences of any kind distinguish one consonant from another.

We should note, too, in any consideration of the constriction sounds as cues, that by their presence or absence they serve as important manner markers. Thus, a speech sound will not be heard as a member of the fricative class unless there is friction noise, or something that will pass for it: remove the friction from /ʃa/ and you will hear a perfectly satisfactory /ga/.

We come now to a second and very different class of consonant cues. These contrast with the constriction sounds in that the members of this second class result from, and can therefore provide information about, the movements of the articulators. It is characteristic of these sounds that they originate in the voice box, rather than at the point of consonant constriction, and that they must, therefore, travel through the entire vocal tract before issuing from the lips. Unlike the constriction sounds, then, these of the second class are affected by the articulatory movement that is made in going from the consonant to the next phone. Acoustically, they appear as the formant transitions, or frequency shifts, that we so commonly see in spectrograms. We know now that these transitions are not merely the incidental acoustic accompaniments of the movements that a speaker must make when he goes from "consonant" to "vowel." Rather, they are perceptual cues, and it is difficult to exaggerate their importance.

At an earlier Speech Conference here at M.I.T. we reported[2] the results of an experiment which indicated that the direction and extent of second-formant transitions is a potent cue for distinguishing within the classes of stop and nasal consonants. Since then we have found that this variable has a similar role with other consonant groups, such as /w,j,r,l/.[8] We also have evidence now that most of the distinctions that are cued by second-formant transitions are affected, though to a lesser extent, by transitions of the third formant. Our investigations into the effects of variations in direction and extent of first-formant transitions, which has so far been somewhat less systematic than our work on the second and third formants, indicates that these variations may help to distinguish among the classes: stops, nasals, liquids, and semivowels. These observations can, perhaps, be generalized by saying that variations in direction and extent of second- and third-formant transitions are cues for the perception of various consonants according to place of production, while comparable variations of the first formant are cues for manner.

We can generalize and simplify our description of these transition variations still further by assuming, as we have in an earlier publication,[9] that there are, for each consonant, characteristic frequency positions, or loci, at which the formant transitions begin, or to which they may be assumed to point. On this basis, the transitions may be regarded simply as movements of the formants from their respective loci to the frequency levels appropriate for the next phone, wherever those levels might be. The spectrographic patterns of Fig. 1, which produce /d/ before /i/, /a/, and /o/, show how this assumption suggests itself for the case of the second-formant transitions. We observe that all of these transitions seem to be pointing to a locus in the vicinity of 1800 cps. We should note, however, that the

[5] Liberman, Delattre, and Cooper, Am. J. Psychol. 65, 497–516 (1952).

[6] K. S. Harris, J. Acoust. Soc. Am. 26, 952 (A) (1954).

[7] The work referred to, most of which is as yet unpublished, has been carried out at Haskins Laboratories by P. Delattre, H. Truby, and L. Gerstman. For one aspect of this research see L. J. Gerstman, J. Acoust. Soc. Am. 28, 160 (A) (1956).

[8] O'Connor, Gerstman, Liberman, Delattre, and Cooper, "Acoustic cues for the perception of initial /w,j,r,l/ in English" (to be published).

[9] Delattre, Liberman, and Cooper, J. Acoust. Soc. Am. 27, 769–773 (1955).

transitions only "point" to the locus; they do not originate there. Indeed, we find in the case of all the stop consonants—and we believe that this will be true of the nasal consonants too—that they can be successfully synthesized only if we introduce a silent interval between the locus and the start of the transition. This does not hold for such other consonants as /w,j,r,l/; here, as we will have occasion to point out later, the loci are the explicit starting points of the transitions.

We should note, too, that the locus concept is complicated somewhat by several special problems. One has to do with the velar consonants /k,g,ŋ/. Here we find a high-frequency locus at about 3000 cps, and this works rather well when the following vowel is in the range /i/ through /a/. Between the vowels /a/ and /ɔ/, however, there is a real discontinuity in the transition cue—and in the locus—and beginning with /ɔ/ we find that the /g/ locus is now very low in frequency and so vague that we have difficulty in specifying its precise position.

The second complication is that the locus tends to move, at least slightly, with the frequency level of the following vowel. In a very significant study Stevens and House[10] have presented some calculations which indicate that this should happen with two of the stops. Our own techniques are such that we could not expect to detect these movements in the case of the stops and the nasal consonants. The Stevens and House results prompted us to look more closely at some other consonants, however, and with /w,j,r,l/ we have, indeed, found direct evidence for such movement.[8]

Thus, the locus is somewhat more complicated than it might ideally be. However, the movement of the locus is much less than the range of frequencies through which the formants of the following vowels move, and there are, even for /g/, fewer loci than there are different transitions to various vowels, so the locus would still appear, on balance, to have considerable utility for simplifying the data of transition direction and extent.

Besides the direction and extent of the transitions, we have found another type of transition variation to be important for consonant perception. This consists of a pair of more or less correlated variables. One is transition duration, which we have found to be reasonably sufficient for distinguishing stop consonants from semivowels.[11] The other is the presence or absence of a silent interval between the locus and the start of the transition. We have already seen that the stop consonants, and probably the nasal consonants too, cannot be synthesized without such a silent interval. In the synthesis of the semivowels, on the other hand, it helps greatly to start the transitions right at their loci. The liquids /r,l/ also require, even more than the semivowels, that there be no silent interval. Indeed, it is true of the liquids and the semivowels, but especially of the liquids, that they can be convincingly synthesized only if the transitions remain at their loci for 30 to 50 milliseconds.[8]

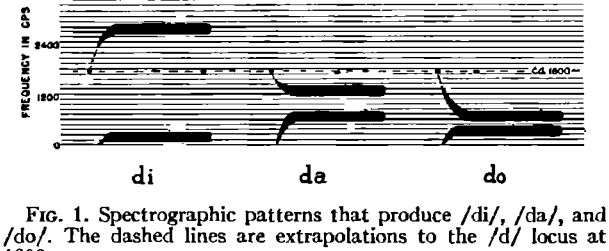

FIG. 1. Spectrographic patterns that produce /di/, /da/, and /do/. The dashed lines are extrapolations to the /d/ locus at 1800 cps.

[10] K. N. Stevens and A. S. House, J. Acoust. Soc. Am. 28, 578–585 (1956).

Thus, we have in regard to these two correlated transition cues at least two groups of sounds. There are in the one group the stop and, tentatively, the nasal consonants, of which we can say, first, that the total duration of the transition is quite short, and second, that there must be a silent interval between the locus and the start of the transition. In the second group are the semivowels and liquids, in which cases the total duration of transition is relatively long, and the transitions start at their loci.

We have so far covered two broad categories of consonant cues—constriction sounds and transitions—and between these two classes we have taken care of all but one of the cues that I want to include in this report. The remaining cue, which has got to be put into a separate class, results from the on-off action of a single fixed resonator, and, accordingly, this cue is either present or absent. The fixed resonator is in the nose, and the corresponding acoustic cue is an on-or-off nasal resonance that serves as an acoustic marker for the class of nasal consonants /m,n,ŋ/. This nasal cue does not, so far as we can tell, provide much of a basis for distinguishing the sounds within the class of nasal consonants. For that, the listener must rely on the transitions of the second and third formants.[12]

Having now come to the end of our discussion of these three types of cues, we ought to make some general observations concerning the number and variety of phones for which each type is important. We have already said that the constriction sounds occur in, and are important for, the fricatives, affricates, and stops, and that the nasal resonance is found only in the three nasal consonants /m,n,ŋ/, for which it serves in perception as a class marker. It has, perhaps, been only implicit in our discussion that all consonants produce transitions, and we have probably not made it sufficiently clear that for almost all the consonants the transitions have so far proved to be of considerable consequence as cues, either by themselves or together with the other cues we have been describing.

[11] Liberman, Delattre, Gerstman, and Cooper, J. Exptl. Psychol. 52, 127–137 (1956).

[12] Liberman, Delattre, Cooper, and Gerstman, Psychol. Monogr. 68, Whole No. 8, 1–13 (1954).

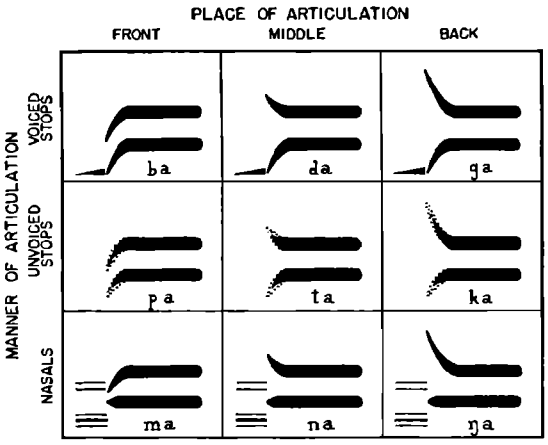

FIG. 2. Spectrographic patterns that illustrate the transition cues for the stop and nasal consonants in initial position with the vowel /a/. The dotted portions in the second row indicate the presence of noise (aspiration) in place of harmonics.

This is, perhaps, the point at which to say again that I have tried only to give examples of the kinds of cues we have found. Details and qualifications have been omitted virtually by the hundreds, and I have not gone out of my way to stake out the areas in which we are totally ignorant. It should be said, too, that the organization I have attempted to impose on our data is largely a matter of expository convenience. I hope nevertheless that this quick survey of our work has been sufficient at least to indicate the broader outlines of our results, and to provide a basis for the speculations about speech perception to which I should like now to turn.

In any search for the more general implications of what is known about the cues for speech perception, one can hardly avoid considering the fact that the data tend to arrange themselves in a very simple way. When we look, for example, at the transitions that are cues for the perception of the stop and nasal consonants, we find that they fall rather nicely into a three-by-three table such as we see in Fig. 2.

It is, of course, immediately obvious that this table parallels the well-known linguistic classification in terms of place and manner of articulation. And to the extent that there is a one-to-one correspondence between articulation and the acoustic result, the table seen in this figure is neither more nor less than we should expect. Though it is, perhaps, not surprising that the acoustic patterns should fall into place in this way, it is nonetheless marvelously convenient that they do. For this unearned increment of simplicity makes it possible for us to speak of an acoustic cue for an entire class of sounds. Thus, for a particular vowel, we can specify the second-formant transition that will produce an initial consonant having a bilabial place of production regardless of manner, and, similarly, we can describe the kind of first-formant transition that will serve as a manner marker for the class of voiced stops. Even more generally, of course, we can refer to the loci and avoid the necessity of specifying the associated vowel.

The table of Fig. 2 not only simplifies and generalizes some of our data, but, more importantly, it suggests a means by which distinctive stimuli are created. As we see from the figure, a limited number of cues on one dimension combine in all possible pairs with cues on each of several other dimensions to produce the sounds that we perceive as speech. Thus, we compound a rather large number of highly identifiable stimuli by freely combining a few values from each of several dimensions. As G. A. Miller has recently pointed out,[13] this possibility provides considerable solace to those of us psychologists who are currently much oppressed by the number seven. Lest the nonpsychologists in the audience think that we are now entering the field of numerology, I hasten to point out that, as Pollack[14,15] and others have found, seven is the upper limit on the number of simple stimuli —stimuli, that is, that vary on a single dimension— that a person can typically identify correctly. A basic problem in perception is to explain how, in the face of that limitation, we nevertheless identify as many stimuli as we do. The acoustic table of linguistic elements suggests that in the case of speech perception the problem has been solved by the use of stimuli that are simple mixtures of a relatively few stimulus values or cue elements from each of a number of different dimensions. In this way many distinctive stimuli are created without the perceiver being required to approach very close to the limit of seven. This general type of solution has for many years been at least implicit in the more familiar articulatory form of our table, and in the recent history of linguistic science it has become even more explicit in the work of Jakobson, Fant, and Halle.[16]

As a result of a very important experiment by Miller and Nicely,[17] we know now that the cue elements and dimensions are relatively independent of each other not only in the manner of their combination, but also in the way they are perceived. Some of the results of our own research point to essentially the same conclusion, but our evidence has been collected quite incidentally and is in general less systematic and less elegant than the Miller and Nicely data. I will not attempt to review any of this evidence here, but I would like to add a somewhat relevant observation that comes from experiments of a type made possible by our use of synthetic speech. Typically, as we have seen, we try to isolate the various cue elements, but often we put these elements together in various ways. We find in general that the individual cues retain their identities, so to

[13] G. A. Miller, Psychol. Rev. 63, 81–97 (1956).
[14] I. Pollack, J. Acoust. Soc. Am. 24, 745–749 (1952).
[15] I. Pollack, J. Acoust. Soc. Am. 25, 765–769 (1953).
[16] Jakobson, Fant, and Halle, Preliminaries to speech analysis (Acoustics Laboratory, Massachusetts Institute of Technology, Technical Report No. 13, Cambridge, Massachusetts, 1952).
[17] G. A. Miller and P. E. Nicely, J. Acoust. Soc. Am. 27, 338–352 (1955).

speak, no matter how they are combined. Thus, if we add together two acoustic features that have been found in isolation to be cues for the same phone, the sound that results is always heard as that phone. New or different qualities never emerge, and, in general, we find essentially no interaction. In this same connection, we sometimes use our synthetic speech techniques to combine acoustic elements that are, in isolation, cues for different phones. These combinations, which one of my colleagues refers to as "unspeakable," are always perceived as if they were simple mixtures of essentially unchangeable elements.

We have so far been concerned with the possibility that relatively free combinations of independently perceived elements might be the basis for perception at the phonemic level, where we are dealing with the so-called empty symbols of language. We might note here that such a system would appear, within limits, to apply at the higher linguistic levels too, where meaning, for example, enters to complicate the psychological picture. Thus, we all know how morphemic elements are entered into various combinations to create a variety of words, each element having a particular identity or meaning which it retains regardless of the combination in which it occurs. Indeed, we may suppose that combining independently variable stimulus elements is a workable basis for perception in the language area, and it is likely that this tells us something rather important about the human mind.

But whether we apply this perceptual scheme only to speech perception, or more broadly to language behavior in general, we will not be able to derive much satisfaction from our assumptions until two very pertinent questions have been answered. One of these asks whether any collection of dimensions and cue elements will do, or whether, alternatively, we must use only certain ones that have, perhaps, some special characteristics. It would take us far beyond the scope of this paper to discuss that issue. Besides, we have nothing to contribute at this point beyond an impression, based on our research experience, that some dimensions are going to work better than others. The second question takes into account the long experience that all listeners have had with the sounds of speech and asks then about the effects of learning on the distinctiveness of the cue elements and dimensions. We have some data and some speculations that bear on this point, and I will want to discuss them in a few minutes. But first I should like to return to the acoustic table of Fig. 2 and consider the possibility that proprioceptive, as well as acoustic, stimulus dimensions must be taken into account in any attempt to explain the perception of speech.

We said earlier that we should not be surprised to obtain this table of acoustic elements, given its familiar linguistic counterpart, and given, also, a one-to-one correspondence between articulation and sound. It cannot be too strongly emphasized that the correspondence between articulation and sound is not always one-to-one. Small differences in articulation sometimes cause very large differences at the acoustic level, and the converse is also true.

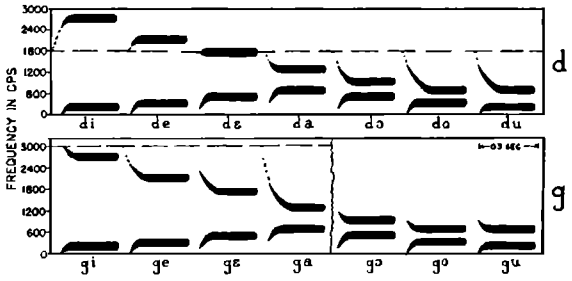

FIG. 3. Spectrographic patterns that produce /d/ and /g/ before various vowels. The dashed lines are extrapolations to the /d/ and /g/ loci.

The occasional complexity of the relation between articulation and the resulting sound wave is, for the most part, a nuisance, but it does provide us with a rare opportunity to ask this interesting question: when articulation and sound wave go their separate ways, which way does the perception go? The answer so far is clear. The perception always goes with articulation.

We have found several extreme and, we think, striking examples of this in our research, and we have discussed one of them in a published paper.[5] It may nevertheless be appropriate to consider another example here.

Figure 3 shows the various transitions of the second formant that are required before each of several vowels to synthesize /d/ and /g/. We note in regard to /d/ that the direction and extent of its second-formant transition is different for different vowels. We also note that in all these cases—that is, with these various transitions—the perception of the consonant is always the same. One hears /d/ throughout.

In the simplest case, we should expect to explain the unchanging perception of /d/ by finding some aspect of the acoustic stimulus that does not change. As shown in Fig. 3, we do, in fact, find such an acoustic invariance in that the /d/ transitions for the various vowels seem to be coming from the same frequency position—namely, the /d/ locus at 1800 cps.

The situation is very different for the consonant /g/, however, as we noted earlier in this discussion, and as we can see in the bottom row of the figure. Here, the consonant takes a progressively bigger transition from the vowel /i/ through the vowel /a/, and we have evidence from our locus research that these transitions are all coming from around 3000 cps. Between /a/ and the next vowel /ɔ/, however, there is a large and sudden change. The best /g/ transition is now very small, and the locus, if indeed there is one, has shifted its position radically.

In this sudden shift between /ga/ and /gɔ/, we have a real discontinuity at the acoustic level. We have been

able to find no acoustic invariant to correspond to—or, if you will, to explain—the unchanging perception of /g/ in this series, and there are reasonable grounds for supposing that none exists. The important thing, of course, is that this discontinuity at the acoustic level is not paralleled by any corresponding discontinuity in articulation or in perception. The /g/ articulation is essentially the same throughout, and so also is the perception.

If the perception depended most directly on the acoustic stimulus, then, in order to preserve the /g/ as perceived, we should have had to change the articulation radically so as to hold the acoustic pattern constant. The fact that we haven't done that, and that we nevertheless hear /g/ with all the vowels, would seem to argue that the perception is somehow more closely related to the articulation than to the acoustic stimulus.

All of this strongly suggests, as do other similar cases, that speech is perceived by reference to articulation—that is, that the articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception. In its extreme and old-fashioned form, this view says that we overtly mimic the incoming speech sounds and then respond to the proprioceptive and tactile stimuli that are produced by our own articulatory movements. For a variety of reasons such an extreme position is wholly untenable, and if we are to deal with perception in the adult, we must assume that the process is somehow short-circuited —that is, that the reference to articulatory movements and their sensory consequences must somehow occur in the brain without getting out into the periphery. I realize that this qualification tends to shield the theory from some of the revealing light of fact, but it does not render it wholly meaningless. The example we talked about a moment ago is still reasonably well explained by the assumption that the perception depends most directly on the sensory consequences of a mimicking articulation. On that basis we should expect that the very different acoustic stimuli for /g/ would come to sound exactly alike because they are produced by the same gross movement.

I should like to turn now to the results I spoke of a while back when I said that we had some data relating to the effects of learning on the distinctiveness of speech sounds. It is, I think, appropriate to have put off this discussion till now, because, as we will see, the results I want to describe appear to be somewhat easier to understand in the light of our assumptions concerning the importance of the sensory return from a mediating articulation.

The results have to do in general with the relation between a listener's phonemic identification of speech sounds and the extent to which he can discriminate these sounds as being different in any way. We find—and this will surely not surprise the linguist—that discrimination is better near the phoneme boundary than it is in the middle of the category.[18] For the particular experiments that we have so far carried out, we used a series of 14 synthetic speech patterns in which the extent of the second-formant transition was varied in small steps through a range sufficient to include /b/, /d/, and /g/. In one part of the experiment these stimuli were presented singly and in random order with instructions to identify them as /b/, /d/, or /g/, and to guess if necessary. As we had reason to anticipate on the basis of previous work with similar patterns, our listeners identified the stimuli in such a way as to divide the acoustic continuum up into three sharply defined phoneme categories, the shifts from one category to another being very abrupt. In another part of the experiment we arranged these same stimuli pairwise and determined by an ABX procedure how well these speech sounds could be discriminated as being different on any basis whatsoever. We found in general that discrimination was better near the phoneme boundaries than it was in the middle of the phoneme categories. This is to say that, with acoustic differences equal, our listeners more easily discriminated between sounds to which they habitually attach different phoneme labels than they did between sounds which they normally lump into the same phoneme class. Indeed, these effects were so great as to approximate rather closely to what we would expect to obtain on the basis of a most extreme assumption: namely, that the listener can discriminate these sounds only to the extent that he can identify them as different phonemes.

We will for the time being simply assume that we are here dealing with the effects of learning on discrimination—for we almost certainly are—and instead of asking whether this is learned, we will ask rather what is learned? It is tempting to speculate that what is learned is simply a connection between various acoustic stimuli and certain articulatory responses. Given that, and given the assumption that the articulation and its sensory consequences mediate between the acoustic stimulus and the perception, then the results we have been discussing follow.

One possibility, of course, is that in the raw these speech stimuli were all as highly discriminable as the most discriminable pair, and that the effect of our linguistic experience has been to dull our sensitivity to the differences within the phoneme category. This is called acquired similarity by one group of psychologists, and it can be made to fit our assumptions quite easily. In the case of that part of the acoustic continuum that runs from /b/ through /d/, for example, we may suppose that our naive American listener has only two possible responses available to him: the /b/ response with the lips and the /d/ response with the tongue. Intermediate articulations are not possible, and neither are intermediate perceptions. All the transition extents that get themselves attached to the /b/, or labial,

[18] Harris, Liberman, Hoffman, and Griffith, J. Acoust. Soc. Am. 28, 760 (A) (1956).

response become indistinguishable because their perception comes to depend most directly on the sensory consequences of a single articulatory gesture.

The contrary possibility is that our stimuli were originally as little discriminable as the least discriminable pair, and that our discriminations have been selectively sharpened as a result of our long experience with the language. This effect, if it occurs, would be similar to what has been called acquired distinctiveness. Such a phenomenon probably has disturbing implications for some people in that it suggests the wrong kind of entropy and calls up visions of Maxwell's demon stationed at some strategic spot in the brain. If this were a problem, and I doubt that it would be in any case, we would want to dismiss it, because we have some direct, if very preliminary, evidence that in the case of some speech cues there does, in fact, appear to be a rather large amount of acquired distinctiveness.[19]

The mechanism to account for acquired distinctiveness is very easy to imagine in terms of the assumptions we have been making. We should suppose that the perceived difference between two relatively similar external stimuli could be increased if we could attach to those stimuli two very different mediating responses and hence gain the added distinctiveness of their very different proprioceptive returns.

The phenomenon of acquired distinctiveness has been investigated in various psychological laboratories. In setting up the conditions under which this distinctiveness is to be acquired, however, the investigators I know about have omitted one arrangement which is peculiar to, and possibly important for, linguistic perception. For in the perception of speech, the mediating articulation not only produces distinctive proprioceptive stimuli, but also external sounds which can be matched against the sounds being perceived. Or, perhaps, we should say that these sounds would be produced if the movements were overt, as they conceivably are in our early years. Considering our very great ability to discriminate two stimuli—that is, to tell whether they are the same or different—and considering also that the possibility of mimicking reduces the first step in the perception process to this very easy discrimination, we might suppose that the articulatory mediation would be of some help initially in getting some of the acoustic stimuli attached to the appropriate articulatory responses. To see this a bit more clearly, let us leave the field of speech perception and consider an example involving the identification of unidimensionally varying lights.

We would expect, on a basis which has already been mentioned, that a subject will be able to identify only about seven different brightnesses of light. These are brightnesses which are presented singly and in random order for absolute judgment. We know, too, that practice doesn't seem to help very much. Suppose now that we provide the subject with a second or comparison light, the brightness of which he can control with a series of levers. If he is permitted to match this comparison light with the standard stimulus, the number of brightnesses he can identify will obviously be increased very greatly. Given a sufficient number of distinctively different lever responses, his ability to identify the brightnesses is now limited only by his ability to discriminate them, and we know that the latter ability is normally very great.

In the beginning it would, of course, take our subject a long time to make each match. It is quite reasonable to suppose that practice would reduce the amount of trial and error, and that our subject would ultimately be able to make the matches much more quickly. The obviously critical question is this: what would happen, after a great deal of practice, if we removed the comparison light and then the levers, thus taking a step which might be analogous in the area of speech perception to putting an end to overt mimicry. If we knew the answer to that question we would, I think, be much closer to an understanding of how learning affects the perception of speech.

[19] In an exploratory study at Haskins Laboratories Mr. Gerstman has dealt with classes of speech sounds (fricative, affricate, stop) that can be distinguished on the basis of duration. This makes it possible to compare discrimination data for durations of acoustic stimuli that are in the one case perceived as speech sounds and in another, only slightly different case, as sounds which are not speech. We should have evidence for acquired distinctiveness or acquired similarity if, with equal differences in duration, the speech stimuli should prove to be more (acquired distinctiveness) or less (acquired similarity) discriminable than the comparable nonspeech stimuli. The preliminary indications are that there is considerable acquired distinctiveness for the speech stimuli that lie near phoneme boundaries. There appears to be no acquired similarity for the stimuli in the middle of the phoneme categories.