This paper was presented at the Speech Communication Seminar (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm) in September, 1962; it will appear in the Proceedings of that seminar.

A MOTOR THEORY OF SPEECH PERCEPTION*

Alvin M. Liberman,** Franklin S. Cooper, Katherine S. Harris, and Peter F. MacNeilage

Haskins Laboratories, New York

The accuracy and speed with which speech is perceived must surely rank as one of its most important properties. To appreciate how remarkably good language is in this respect, one need only try to find or to fabricate a set of non-speech sounds that will serve a human being as well. Attempts to do this in connection with reading machines for the blind, for example, have not been notably successful, and we know that Morse code is relatively poor by comparison with speech, even after years of practice.

Because the sounds of speech are highly distinctive -- that is, absolutely and quickly identifiable -- they are efficient vehicles of information transmission. But the distinctiveness of speech sounds is more than a matter of mere convenience and utility; it is, in fact, necessary if language as we know it is to exist. An obvious consideration is that a phonemic system requires by its very nature that the sound elements be identifiable in absolute terms. The phoneme /d/, for example, must be perceived not merely as something which is more or less like /d/ than the last sound heard, but as /d/ itself. Nor does the grammar of our language permit that this or any other phoneme be transformed into a radically different perception according to the context in which it appears. Phoneme perception must be absolute, or very nearly so, if language is to be phonemic. We should note, too, that there is a sound psychological reason why languages are phonemic: a non-phonemic code would present difficulties for much the same reason, and to roughly the same extent, that a syllabic or word system of writing and reading is more difficult than one based on an alphabet.

The requirement that the phonemes be perceived quickly is, perhaps, less obvious, but it is none the less real. If one tries to understand language when it is read to him slowly, letter by painful letter, he sees that the phonemes must come

along at a rather high rate else the listener cannot organize them into morphemes, words, and sentences.

Although speech is highly distinctive in human perception, no machine has yet been designed which finds it so. Indeed, it may well be true -- and, if so, ironic -- that machines will have their greatest difficulty with those very phonemes (e.g., the stop and nasal consonants) which are for human beings most highly distinctive and which probably carry the heaviest load of information. There are, of course, various codes, all of them non-linguistic, which are well received by a variety of machines, but these work badly with human beings. That man and machine are so different in this respect is not very remarkable. It merely indicates that there is something psychologically interesting about the efficiency of speech perception, and that we shall increase our understanding of language, and man, by inquiring into the conditions of distinctiveness and the perceptual mechanisms which underlie it. It also emphasizes, if emphasis were needed, that those who would design a speech recognizer might somehow profit from the same kind of inquiry.

There is now available a large amount of data bearing in one way or another on speech perception. Where, among these data, do we find the conditions of distinctiveness?

One looks first, of course, at the acoustic cues themselves -- that is, at those aspects of the speech wave on which the identification of the phonemes depend -- and asks whether there is anything about them which might tend to make them inherently distinctive.[1] In this connection we ought first to offer the comment that the acoustic cues appear in and of themselves to be quite ordinary. This is, of course, no more than an opinion. Far more convincing are the observations that are made when, by using synthetic speech as a means of achieving stimulus control, we present an acoustic variable which cues a phonemic distinction and then the same variable in a non-speech context. For some phonemes the extremely distinctive difference one hears in the speech case is considerably less distinctive, if indeed the difference can be heard at all, when the variable is listened to in isolation or in a non-speech pattern which is most nearly equivalent. This strongly suggests that distinctiveness is not inherent in the acoustic signal, but is rather added as a consequence of linguistic experience. More important, perhaps, it indicates that even with a considerable background of linguistic experience on the part of the listener, the acoustic

signal is distinctive only when, being heard as a speech sound, it engages some kind of speech perception system.

We ask, then, how is distinctiveness increased when the incoming signals enter this speech perception system, and, further, what are the properties of the system? One answer to both questions can be developed out of research findings which indicate that some of the consonants are perceived categorically. The impressionistic data, in a typical case, are to this effect: when we listen to a series of synthetic speech sounds in which the second-formant transition is varied progressively in such a way as to produce in succession /b/, /d/, and /g/, we do not hear a gradual change corresponding to the gradually changing stimulus; rather, we hear the first three or four stimuli as identical /b/'s, then, very abruptly with the next stimulus, the perception is of /d/, where it remains essentially unchanged until, again abruptly, it shifts to /g/.

To assess this effect more precisely we measured discrimination along the acoustic continuum on which /b/, /d/, and /g/ lie, and found that, for equal physical differences, acuity was considerably greater across phoneme boundaries than within the phoneme classes[2]. Indeed, the peaks were so high as to create a situation in which the discriminability of the speech sounds was very nearly predictable from the frequency with which they had been identified by the listeners as one or another phoneme. To this extent, the listeners could discriminate the sounds only as well as they could identify them absolutely as phonemes.

These results contrast with those usually obtained in psychophysical studies of non-speech continua. Typically, one finds that discrimination along an acoustic continuum is monotonic -- without peaks or dips -- and, as is well known, people can ordinarily discriminate many more stimuli than they can identify. This is to say that in the perception of non-speech continua, perception is continuous, and stimuli are dealt with very much better on a relational basis than in absolute terms.

Although the stop consonants lie on an acoustic continuum, the perception is essentially discontinuous. Because of the discrimination peaks at the phoneme boundaries, the incoming sounds are heard categorically -- that is, in absolute terms, rather than in relation to other stimuli -- and they are, therefore, quickly and accurately sorted into the appropriate phoneme bins.

We wonder, then, whether these discrimination peaks (and the categorical perception they produce) are part of our innately given sensitivity to the acoustic variable, or whether they are, alternatively, a result of our long experience with the language. In several studies similar to those referred to earlier, we have measured the discriminability of an acoustic variable when, in an appropriate pattern, it cues a phonemic distinction and when, in the most nearly equivalent non-speech context, it does not.[3] Where discrimination peaks occurred in the speech-sound continua, none was found in the non-speech controls; we conclude from this that the peaks are a result of learning. By comparing the discrimination levels in the speech and non-speech cases, we collect evidence that bears on the direction the learning has taken: it appears to consist entirely of a sharpening of discrimination across phoneme boundaries; there was no evidence of a reduction of discrimination within the phoneme category.

Having reason to believe that the discrimination peaks are a result of learning, we ask next what is learned, and, more broadly, what kind of mechanism underlies the categorical perception of the consonants. The answer seems to us to lie in a theory about speech perception which other aspects of our research had led us previously to adopt, namely, that the perception of speech is tightly linked to the feedback from the speaker's own articulatory movements.[4] Looking at the discrimination peaks from that standpoint, we should say that what has been learned is a connection between speech sounds and their appropriate articulations. In time, these articulatory movements (or, more likely, the corresponding neurological processes) come to mediate between the incoming acoustic stimulus and its ultimate perception. If this is so, then indeed we should expect that the perception of many of the consonants would be categorical because the articulation is so obviously categorical. Consider the case of /b,d,g/. The appropriate acoustic cues occupy different positions along a single acoustic continuum (extent of second-formant transition), but they are produced by very different articulations. Thus /b/ is produced by a movement of the lips and /d/ by a movement of the tongue. Given the discontinuous articulation , we should expect, in spite of the continuous nature of the acoustic variations, that the perception would be discontinuous (i.e., categorical), and that a discrimination peak would appear at the phoneme boundary. There are no intermediate articulations between /b/ and /d/, and, as a consequence, no intermediate perceptions.

To provide a further, and more nearly precise, test of this view, we selected another phonemic contrast in which we might expect (1) that perception of a continuously varying acoustic stimulus would be categorical and (2) that we would be able, with fair accuracy, to measure the listeners' attempts to mimic the various settings of the stimulus. The point, of course, was to find, as a matter of fact, whether the mimicking responses are graded (i.e., changing as the stimulus changes), or whether they are, like the perception, essentially categorical. The phonemic contrast was that between /sl/ and /spl/ in the words slit and split, and the acoustic variable was simply the duration of the interval of silence between the "s" friction (noise) and the vocalic part of the syllable.[5] There was, in fact, a peak in discrimination at the phoneme boundary and, perception of the acoustic continuum (variations in the silent interval) was essentially categorical. To obtain the mimicry measures, we presented the stimuli several times over with instructions to the subjects to mimic each one as best they could. Acoustic and electromyographic recordings of the subject's responses were made. The acoustic records showed that the time intervals introduced in mimicry were, indeed, essentially categorical. We then examined the electromyographic records to determine whether there was, nevertheless, some tendency for the subjects to make tentative or partial lip gestures in the middle range of the continuum. The electromyographic results were also essentially categorical: either there was a normal burst of muscle potential at the lip (indicating a p-gesture) or there was not; there was no evidence of intermediate gestures, such as partial closure, in response to stimuli near the center of the acoustic continuum.

From an articulatory standpoint, the vowels are different from the stops and some of the other consonants in that the articulators can move continuously from one vowel phoneme to another. We might expect, then, that the perception of the vowels would prove to be quite different from the stops. This expectation has been confirmed by an experimental study of synthetic vowels which paralleled the earlier studies of the stops.[6] There was, with these vowels, no increase in discrimination at the phoneme boundaries; moreover, the obtained discrimination functions lay considerably above those that were derived on the assumption that the listener can only hear these sounds phonemically, which is to say that the listener heard many intra-phonemic variations. Thus, in contrast to the perception of the stops, which is very nearly categorical, the perception of the synthetic vowels is continuous. Results

similar to those with the vowels have been obtained with two other phoneme distinctions[7] (vowel length and tones in Thai) in which, as in the case of the vowels, the production can vary continuously from one phoneme to the other. In one of these studies (vowel length as the basis for a phonemic contrast in Thai) where the perception was found to be continuous, measures of the subjects' attempt to mimic the stimuli disclosed that the production of the sounds was also continuous.

We have said that the categorical perception of the stops (and some other consonants, too) is an important condition of their distinctiveness.[8] We must, then, indicate here that the vowels and phonemic distinctions based on vowel length should be considerably less distinctive. Having said earlier that the phonemes of a language must be distinctive, we should now revise that statement to say that not all the phonemes must be highly distinctive, but only some reasonable number. Lest this seem to be a meaningless hedge, we should go further and suggest that the degree of distinctiveness will determine, in part, the linguistic role of the phoneme in question. Thus, we should guess that the information load (as measured by the number of minimal pairings) would be greatest on the most distinctive phonemes. This can be tested, and we hope one day to have the relevant data.

In any event, there is evidence from perceptual studies that speech sounds are perceived by reference to the articulatory movements that produce them, and that this articulatory reference is important for the distinctiveness of speech as perceived. We should, perhaps, add here a point, made in an earlier publication,[9] that also has to do with the advantage to a listener of making a mimicking response, at least in the early stages of learning to perceive the language. It is well known, at least in the case of simple stimuli, that man's ability to discriminate (i.e., to determine that two stimuli are the same or different) is very good, but that his ability to identify in absolute terms (i.e., to tell which stimulus it is) is very poor. We can, for example, discriminate one to two hundred times as many pitches as we can identify absolutely. Now when a child hears a speech sound and undertakes, conceivably by trial and error, to mimic it, he is limited only by his differential sensitivity -- that is, by his ability to determine whether the two sounds are the same or different -- and that, as we know, is extremely acute. Given that the two acoustic stimuli are rather similar, but that the "matching" responses are made with different muscles, as,

for example, in the case of /b/ and /d/, the feedback stimulation would be more distinctive than the acoustic signals themselves. Thus, one can, by mimicking, use his keen differential sensitivity as a basis for making absolute identification, provided, of course, that the mimicking gestures themselves provide distinctive feedback stimulation.

In general, however, the articulatory reference theory does not rest primarily on calculations of what might have increased distinctiveness, but more empirically on the evidence which indicates that the relation between phoneme and articulation is more nearly one-to-one than is the relation between phoneme and acoustic signal. Thus, we have seen in this paper that perception of the phonemes is discontinuous (categorical) or continuous depending on the nature of the appropriate articulatory movements and not on the properties of the acoustic signal. Elsewhere we have reviewed evidence bearing on the same general point.[10] In that connection we pointed out that because of the characteristics of the articulators and the vocal tract, and because of the overlapping in time of the articulatory gestures, the relationship between articulation and the acoustic signal is often complex. For example, when stop consonants are articulated into different vowel cavities, as in the normal production of stop-vowel syllables, we find extreme cases in which large differences in articulation produce little or no difference in the "consonant" part of the acoustic signal; there is, also, the opposite case in which essentially the same consonant articulation produces (in different vowel contexts) very different acoustic results. We ask in these cases what happens to the perception and find always that it follows the articulation, not the sound.

Given the complex, highly encoded relation between phoneme and sound, and the more nearly one-to-one correspondence between phoneme and articulation, we have been led to assume that a reference to articulation might well be a stage in the perceptual process. We have seen why a connection to articulation might have been established as the child learns to perceive the language, and we find it easy to suppose that such a connection would remain in the adult, though surely in some short-circuited form, since it would help him to decode the complex acoustic signal.

We should now emphasize, because it is most important to our theory, that the simple correspondence with phoneme perception is not to be found in the conventional conceptions and descriptions of articulatory phonetics, which are concerned largely with the changing shapes of the vocal tract, but rather

in the motor commands that actuate the articulators.[11] Because of the interactions and constraints inherent in the mechanism of the vocal tract, the encoding of motor commands into shapes and movements is often a complex transformation. The motor commands operate ahead of these complications, and so escape this kind of recoding. Of all the speech events -- acoustic signal, articulatory shape, or neuro-motor commands -- about which we can now reasonably expect to collect information, the neural commands to the articulators will, in our view, provide the simplest relationship to phoneme perception.

On the theory that a reference to articulation helps the listener to decode the acoustic signal, we must suppose that he would use that information about articulation which corresponds most closely to the phoneme. We believe, as we had said, that a sufficiently direct correspondence exists at the level of the neuro-motor commands to the articulators. Just what kind of information about the motor commands would be available to the listener cannot now be known. Proprioceptive and tactile cues could, of course, indicate the nature of articulatory gestures, and thus provide information about the structure of the command system. In this connection we want to make very clear, however, that we ought not to rule out the possibility that the neural commands themselves, or rather their equivalents in the central nervous system, might be used to provide the reference system in terms of which the decoding is carried out.

On the basis of all these considerations, and in order to provide evidence relevant to a motor-command theory of speech perception, we are trying now to find out what we can about the neural signals that actuate the articulators. We have not undertaken to measure the neural command signals themselves, but rather various aspects of articulatory activity, induding in particular muscle action potentials, from which the neural commands can be rather directly inferred.

The search for the system of motor commands has only begun. The studies so far carried out do indicate, however, that the relevant measurements can be made, and give promise of significant data.

In early studies[12] of /p,b,m/, electromyographic recordings were made from the lips and from the velum, and these were supplemented by recordings of glottis action. The electromyographic activity at the lips was found to be essentially the same for /p/, /b/, and /m/. The stops /p,b/ contrast with the nasal /m/

in that there is for /p/ and /b/ an identical burst of activity which can be measured on either the superior or inferior palatal surface, indicating the closure of the velum, while for /m/ no electromyographic activity is observed anywhere in the palatal region. The contrast between /m,b/ on the one hand and /p/ on the other appears solely in the relative timing of glottis action and lip opening. Thus, it appears that one can expect to obtain measures of articulatory activity, even from the velum, which reflect the system of motor commands, and the results in this case reassure us as to the simplicity and independence of place, nasal manner, and voicing.

Another study[13] investigated the acoustical and articulatory characteristics of /f/ in absolute final position and when embedded in final clusters. The acoustical aspects were determined from spectrograms, while information about the articulation was obtained from recordings of the electromyographic activity of the appropriate lip muscles. The acoustic signal for /f/ is different depending on the phonetic context, the most obvious effect of context being a variation in the duration of the /f/ friction by as much as two to one. The electromyograms, on the other hand, were virtually identical regardless of phonetic context. Thus, it would appear that the /f/ command is the same for all cases, but that changes in the timing of the /f/ gesture relative to other gestures alters the acoustic signal from one context to another.

Other experiments now in progress show signs of yielding similar and similarly promising results. Where the acoustic differences between phonemes show rather complex interactions, electromyographic and other measures of articulating activity suggest that the underlying motor commands are independent and simple.

Thus, we have seen how an attempt to understand the perceived distinctiveness of speech sounds led us to the hypothesis that speech is pereeived by reference to articulation. This seemed a reasonable hypothesis on the basis of repeated findings that phoneme perception was in more nearly one-to-one relation to articulation than to the acoustic signal. Consideration of how the speech production system works, from motor commands to articulatory gesture and from gesture to sound, makes it clear why the acoustic signal should be rather complex and remote from the phoneme, and, further, why the simplest relation to the phoneme might be found in the motor commands that actuate

the articulators. Pilot research employing electromyographic and other appropriate measures of articulatory activity have yielded data from which inferences about the motor commands can be drawn. While far from definitive, the results indicate that the methods are feasible; they also suggest that the motor commands do stand in a very simple relation to the phonemes, and thus lend some further credence to the view that these commands provide a reference system in terms of which the complex acoustic signal is accurately and quickly identified.

Footnotes

* The work reported here is an outgrowth of a program of psychological and physical research on auditory patterns supported by grants from the Carnegie Corporation of New York. Some phases of the work are currently supported by a grant from the National Science Foundation.

** Also, the University of Connecticut.

1. For a description of some of these cues, see Liberman, A.M. Some results of research in speech perception. J. Acoust. Soc. Amer. 1957, 29 pp. 117-123.

2. 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. Psych. 1957, 54 pp. 358-368.

3. Liberman, A.M., Harris, K.S., Kinney, J.A., and Lane, H. The discrimination of relative onset-time of the components of certain speech andnon-speech patterns. J. Exper. Psych. 1961, 61 pp. 379-388; Liberman, A.M., Harris, K.S., Eimas, P., Lisker, L., and Bastian, J. An effect of learning in speech perception: the discrimination of durations of silence with and without phonemic signtficance. Language and Speech 1961, 4 pp. 175-195; Bastian, J., Eimas, P., and Liberman, A.M. Identification and discrimination of a phonemic contrast induced by silent interval. J. Acoust. Soc. Amer. 1961, 33 p. 842 (abstract).

4. For a fuller account see Cooper, F.S., Liberman, A.M., Harris, K.S., and Grubb, P.M. Some input-output relations observed in experiments in the perception of speech. Proceedings of the Second International Congress on Cybernetics 1958, Namur, Belgium; Lisker, L., Cooper, F.S., and Liberman, A.M. The uses of experiment in language description. Word (in press); Liberman, A.M., op. cit.

5. Harris, K.S., Bastian, J., and Liberman, A.M. Mimicry and the perception of a phonemic contrast induced by silent interval: electromyographic and acoustic measures. J. Acoust. Soc. Amer. 1961, 33 pp. 842 (abstract).

Footnotes continued.

6. Fry, D.B., Abramson, A.S., Eimas, P., and Liberman, A.M. The identification and discrimination of synthetic vowels. (in preparation).

7. Abramson, A.S. Identification and discrimination of phonemic tones. J. Acoust. Soc. Amer. 1961, 33 p. 842 (abstract). Bastian, J. and Abramson, A.S. Identification and discrimination of phonemic vowel duration. J. Acoust. Soc. Amer. 1962, 34 pp. 743-744 (abstract).

8. We do not suppose that this is the only condition. Other characteristics of the phonemes which may relate to distinctiveness -- the multiplicity of cues, for example, and the way they sum -- will be dealt with in a forthcoming paper.

9. Liberman, A.M. op. cit.

10. Cooper, Liberman, Harris, and Grubb, op. cit. Also, Lisker, Cooper, and Liberman op. cit.

11. See Cooper, Liberman, Harris, and Grubb op. cit. Also, Lisker, Cooper, and Liberman op. cit.

12. Lysaught, G.F., Harris, K.S., and Rosov, R. Electromyography as a speech research technique with an application to the labial stops. J. Acoust. Soc. Amer. 1961, 33 p. 842 (abstract Harris, K.S., Schvey, M.H., and Lysaught, G.F. Component gestures in the production of oral and nasal labial stops. J. Acoust. Soc. Amer. 1962, 34 p. 743 (abstract).

13. MacNeilage, P. Electromyographic and acoustical study of the production of certain final clusters. J. Acoust. Soc. Amer. 1962, 34 p. 743 (abstract).