Some Observations on a Model for Speech Perception[†]

A.M. Liberman,[*] F.S. Cooper, Katherine S. Harris, P.F. MacNeilage and M. Studdert-Kennedy[**]

Haskins Laboratories, New York City

Since this symposium is concerned with models, we should say at the outset that we do not have a model in the strict sense, though we are in search of one. What we do have is some notions about the general characteristics that a model of speech perception should have. We will try this morning to describe these notions and the facts that have led us to entertain them.

We should examine first certain facts that exist independently of any research on speech itself. Some of these facts emerge from an inquiry into the structure of language, others from a look at the limitations of the auditory system. Together, they define an important part of the job that a model of speech perception must do, and significantly constrain the assumptions it can use.

We begin, then, with some facts about linguistic structure. These have great generality in that they transcend any particular modality through which the message might come. They must be taken into account, that is, whether communication is by eye, by ear, or through the skin. One might say of these facts, therefore, that they are somehow cognitive, but for our purposes we need only say that, being most general, they have an obvious priority and we should consider them first.

It is the more appropriate that we should begin with such general considerations because this symposium is concerned with the perception of optical as well as acoustic signals, and we have in mind that language is perceived, in literate people at least, by eye as well as by ear. We will, in time, narrow our concern to the perception of spoken language. In keeping with the purpose of this symposium, however, and also in order to see our own problem of speech perception more clearly, we will first consider what it is that must happen in all linguistic perception, and then measure these requirements against the abilities and shortcomings of the eye and ear.

Perception and linguistic structure. In studying the perception of language, we begin with the advantage that there exists a reasonably good description of what is perceived. We know that language is composed of segments arranged in hierarchically ordered layers. We know, too, that the structure formed by these segments and layers is described by two grammars: one, the phonology, for the segments that are themselves empty of meaning; and the other, comprising morphology and syntax, for the segments to which meaning can be ascribed. Among the phonological segments there are, at base, some 30 to 50 phonemes -- themselves composed of a smaller number of subphonemic features -- from which we proceed to a larger number of syllables, and then to units that may be called phonological words. Beginning, in the other grammar, with some thousands of morphemes, we move through a larger vocabulary of words and a vast repertory of phrases to an infinity of sentences.

A consideration of this structure suggests several questions for a model of speech perception. Is the model to deal with perception in both grammars -- that is, on both sides of the meaning barrier? Given that we are on the one side or the other, is it, then, to explain the perception of all the segments in each layer? And will it account for the way the segments in one layer are organized for use at the next higher one, or how the segments at one layer constrain, or operate in parallel with, those of another?

We can ignore these questions, but only if we assert that the segments and layers are figments of a linguist's imagination or, perhaps, linguistic abstractions that have no reality in perception. And, indeed, we may at times be tempted to do just that. We will find, for example, that the acoustic correlate of a given phoneme changes considerably with different phonetic contexts, and we may accordingly be moved to infer that the syllable, rather than the phoneme, is the smallest segment of phonological perception. This would be unwise. We have the greatest respect for the imagination of our linguist colleagues, but we cannot believe that they invented linguistic structure; we hold, with them, that they only discovered it. No more can we believe that this structure is not psychologically real and important. There is evidence, both direct and indirect, that attests to the psychological reality and importance of each one of the segments and the subphonemic features, too. But there isn't time to consider this matter in detail. We will only say, then, that the model we seek must deal with linguistic structure for the simple, and we think unassailable, reason that linguistic structure is a description of what the listener perceives.

We would like, of course, to understand the whole of linguistic perception -- that is, to answer all the questions we raised a moment ago. Our ambition has realistic bounds, however, so we have chosen to be concerned primarily with perception within the phonological system. Still, we should be aware that other problems exist. We are surely well advised to have in mind that a model of phonological perception will one day be married to a model for the perception of morphology and syntax. We ought, therefore, to be concerned in advance about their likely compatibility. At the very least, we should be gratified if the two models grew up speaking the same metalanguage.

In this connection we will get ahead of our story just long enough to say about the model we seek that it contains a key assumption, namely, that the sounds of speech are somehow perceived by reference to the way we generate them. We have discussed this assumption and the reasons for making it in a number of earlier papers.[1] An explicit model that embodies essentially this point of view is described by K.N. Stevens and by Morris Halle and Stevens.[2] They would recognize speech on the basis of analysis by synthesis according to generative or articulatory rules. Now, as Halle and Stevens pointed out in an earlier paper, such a model is not too different from one that seems peculiarly appropriate for perception at the higher levels of language, too.[3] This is hardly a decisive consideration in the selection of a model for phonological perception, but it is relevant, we think, and even interesting.

Requirements for perception at the phoneme level. Let us, now, take the shortest segment at the phonological level -- the phoneme -- and consider the requirements that are imposed on its perception by the nature of linguistic structure. There are two. The first is that the phonemes must not lose their identity as they enter into combinations with other phonemes. If this requirement is not met, the system is not phonemic, in which case it is not linguistic. Perception of the phoneme must also be independent of context in a somewhat different and weaker sense: it is not enough to hear a particular phoneme, say /t/ as something more or less like /t/ than the last phoneme heard; rather, it must be heard as /t/ itself. This is to say that we must identify absolutely, not merely discriminate.

The second requirement is one of speed. Since language is phonemic we need only a few basic segments -- not more than about 40, in fact -- to produce an infinite number of utterances.

But we pay for this economy, other things being equal, by the requirement that we must transmit and receive a large number of these elements per unit time. If we did not, communication would be slow. Worse yet, it might even be impossible, for if one is to organize the phonemes into units at a higher level, they must come in at a reasonably high rate. Here we are limited by our span of apprehension or attention in time, and if anyone doubts this, he need only have someone read to him slowly, letter by painful letter.

In short, the language demands that the phonemes be absolutely and rapidly identified. We shall, for convenience, use the word "distinctive" to refer to stimuli that can be so identified. Let us inquire now how the visual and auditory systems might meet this requirement of distinctiveness.

Phoneme perception in relation to the properties of the eye and the ear. Keeping in mind that the problem is to get the phoneme strings from one person's head into another's, consider first how this might be done if we were dealing with computing machines rather than people. The messages would be organized into word-length segments and then transferred either in parallel or serially. Parallel transfers are usually much faster, for obvious reasons, and we would suppose that this would be true for humans, too.

The visual system appears to be organized on a parallel basis, so one should anticipate no great difficulty in getting the phonemes through the transmission link and into the recipient's head. We need only devise a set of distinctive shapes, one for each phoneme, and set them side by side. We could use one of the alphabets already available, each of them highly distinctive, or we could invent a new one if for any reason this seemed desirable. Because of the eye's ability to perceive in space, we can have a one-to-one relation between phoneme and external signal, as we do with the alphabetic scheme we just described, and still encounter no real difficulty in perceiving word-length sequences of thse phoneme segments.

In the auditory case the problem is different and considerably greater. To the extent that the input channel operates in the serial mode, it is inherently slower than vision. Then, too, it isn't all that easy to get 40 highly identifiable signals, unless one is willing to use up time in patterning the acoustic characters. And to do this is, of course, self-defeating if our purpose is to communicate at high speed. But let us assume that we have somehow overcome this difficulty and succeeded in finding 40 or so highly identifiable acoustic signals that are nevertheless of short duration -- not appreciably longer, that is, than the integration time of the ear. Consider now what happens when we present these one after the other. At rates of 200 to 300 words per minute -- and we can perceive speech produced by time-compressing machines at rates considerably above that -- we should have something between 10 and 20 of these signals per second. This is below the threshold for pitch, but it is still so rapid that the signals would be hard to discriminate, let alone identify and otherwise process.

Perhaps the best way to see the ear's difficulty is to consider what reading would be like if the eye were subject to limitations like those that beset the ear. The appropriate analogy would be with a reading situation in which the individual letters were flashed on and off, one at a time, in succession, and in the same place. But we labor the point and put ourselves in danger of being about to predict that people will never be able to perceive speech, or that our species will develop the ability to read and write before it learns to speak and listen.

At the very least, however, we should not be surprised to discover that speech is a uniquely distinctive set of acoustic signals. And we do discover precisely that. No other acoustic signals will work nearly so well. Some will object at this point that such a statement is true but trivial, having in mind, perhaps, that no one has ever really tried to develop an alternative acoustic system, or that it is, in any case, difficult to match the amount of practice we've all had in listening to speech. In general, such an objection is not well taken. Non-speech ciphers on the phonemic structure of language have been developed and thoroughly tested, not only in the familiar case of Morse code, but also in fifty years of research and development in the attempt to build reading machines for the blind. It is most instructive to review the results of that work, but we don't have time now.[5] We can only say that non-speech ciphers work very badly by comparison with speech.

We would like to be able to compare speech sounds more broadly with acoustic signals that are in no way tied to a linguistic system. Unfortunately, directly comparable data for the rate of processing non-speech acoustic signals are not available.[6] We should consider, however, that in perceiving rapid speech, a listener processes information at rates of up to 50 bits per second. From such knowledge as we do have of the information-carrying capacities of non-speech acoustic signals, such a rate appears to be very high indeed.

What, then, does the unique distinctiveness of speech sounds tell us about a model of speech perception? At the very least it suggests that there must be something special either about the properties of the signal or about the way we process it. Our own very strong tendency is to embrace both alternatives. Speech sounds have the interesting property that they are also produced by the person who perceives them.

In this respect they are not truly unique, however, since we sometimes clap our hands, snap our fingers, tap our feet, and sneeze. Speech sounds are wholly unique in that they alone are produced by neuro-muscular events that are at some point equivalent to the grammar of a language, or, more specifically in this case, to the constituent units of phonology. This, in turn, creates the possibility for a special kind of perceptual processing: namely, that the acoustic signal is somehow decoded by reference to the manner in which these segments are produced. We said only that the possibility is created. We do believe, however, that people take advantage of this possibility, and that if they did not, speech would be far less distinctive than it is.

So much, then, for the problems and facts that exist independently of research on speech perception itself. Let us turn now to that research.

Acoustic cues for phoneme perception: a complex code. A finding which turns up over and over again in a variety of different, yet related forms is that there is a complex relation between perceived language and the acoustic signal which conveys it. More specifically, the acoustic signal is quite commonly not invariant with respect to the phoneme as perceived. We should not be surprised by this since we have just seen what difficulties would plague any attempt to get language past the auditory bottleneck by any simple and obvious means. But we must examine the nature of the complexity and explore its implications.

In the structure of language the phonemes are discrete segments. In the acoustic stream of speech they are not. And, indeed, we could not expect them to be, given what we know about the temporal resolving power of the ear and the number of pulses per second we should hear if the phones were acoustically discrete. Of course, the acoustic cues for the phonemes are there. We know what they are and where they are, and we can, moreover, manipulate them so as to change the perception on a phoneme-by-phoneme basis. But the cues for the phonemes do not lie along the time axis like so many beads on a string. Rather, they are overlapped and, more generally, encoded into units of approximately syllabic size. This is why one cannot synthesize speech from pre-recorded segments of phonemic dimensions.[7] And in the reverse process of machine recognition, this is why the engineer finds it so difficult to segment speech into the phonemic constituents we humans perceive.

The encoding of phoneme strings into syllabic units reduces significantly the number of discrete acoustic segments (per unit time) the listener must hear. In this sense it helps to account for the rapid transmission of information through an auditory channel ill-equipped to pass the serially ordered segments of language. But if this encoding solves one problem, it poses another, for we must now explain how the phonemes are recovered from the encrypted signals. In our view, a suitable decoding mechanism would be specialized to use information about the manner in which the phonemes were encoded in the first place. How the encoding first occurred will be considered shortly.

The rather complex encoding of the phonemes shows in other ways, too. Thus, we have in earlier papers commented on examples of a situation in which phonemic cues for different consonants are acoustically the same before different vowels. We have also described the converse examples in which cues for the same consonant are very different before different vowels.[8] The cases that we have particularly emphasized in this connection are the by now fairly familiar ones that deal with the burst cue in the perception of /pi/, /ka/, /pu/ and the transition shift from /ga/ to /gɔ/.[9] It should be understood, however, that these are not isolated examples. Thus, with transitions that all begin at exactly th same point on the frequeacy scale, one can, before different vowels, produce all three stops /b,d,g/.[10] Conversely, one can, in the case of many consonants produce the same phoneme perception (before different vowels) with transitions that begin at frequencies that are, by any psychophysical standards, enormously different. If one chooses to define the acoustic cue differently -- for example, as the direction and extent of the transition rather than its starting frequency -- exactly comparable difficulties arise.

Indeed, the search for acoustic cues that would show a reasonable degree of invariance with respect to perception has had a long and largely unrewarding history, especially in the case of the consonants. The initial disappointments came in failing to find -- a few cases excepted -- component parts of the acoustic stream that were superimposable for occurrences of the same phoneme in different environments. These would, of course, have qualified as invariant cues in the simplest and strictest sense. Next best would have been such derived cues as the starting frequencies, or the direction and extent, of the formant transitions. Though these would not, by their nature, have met the strict test of superimposability, we could, nevertheless, with some satisfaction, have defined them as the acoustic cues. As we have seen, however, we do not find invariance even when we define the cue in this rather lax way. To unify, at least to some extent, the great variety of acoustic shapes that in different environments produce the same phoneme perception, we had, in the case of the stop and nasal consonants, to discover the "locus" -- the frequency to which the transitions point.[11] Although the locus is here a very convenient concept around which to organize many otherwise diverse acoustic facts, it is in no sense available to the listener as part of the acoustic signal. Nor can it be made an actual part of that signal without grossly upsetting and changing the perception.[12]

Now the loci are more nearly invariant with respect to the perception than are the actual acoustic signals. And in this connection it is of more than incidental interest that the locus incorporates an essentially articulatory transformation.[13] We would suppose that even better approximations to invariance with perception might be had by getting closer to the essential operations underlying articulation. We shall have more to say of these later.

The point to be made now is that for many of the important consonants there is no way to define the acoustic cues so as to have, except in a small number of phonetic contexts, an invariance between acoustic cue and phoneme perception. We cannot here list, or even classify, all the examples. We can only say that they are dramatic and numerous -- so numerous, indeed, as to make this lack of invariance seem almost the rule rather than the exception.

From phoneme to acoustic cue and back: encoding and decoding. The complex relation between acoustic cue and phoneme perception reflects, as we have already said, the encoding of acoustic signals into units of approximately syllabic length. In considering how these signals are perceived, we have at least two choices. As we have already implied, one is to reject the phoneme and begin with the syllable -- to assert, that is, that /ga/ differs from /ba/ only holistically and not just in the first segment. Similarly, we should have to assert that /ga/ differs from /gɔ/ holistically and not just in the second segment. But such assertions are, as we have already said, plainly contrary to fact. Besides, they lead us to assumptions that are uneconomical and inelegant: our model would have to perceive, at base, the thousands of syllables instead of the many fewer phonemes.

An alternative interpretation is that the listener manages somehow to recover the phoneme. To see how he might do this, we should consider how the signals became complexly encoded in the first place. We shall assume -- indeed, we think we must assume -- that somewhere in the speaker's central nervous system there exist signals which stand in a one-to-one relation to the phonemes of the languages. In the act of speaking, these signals, arranged of course in some temporal pattern, flow outward from the central nervous system and eventuate as commands to the articulatory muscles. At this level the relation to phonemic structure is conceivably still quite simple -- that is, we can, perhaps, find a close correspondence to the phoneme by inquiring which muscles are commanded to contract, when, and how forcibly. (Just how close or remote this correspondence is will have to be considered later.) The next steps, of course, are the transformation of these motor commands into a shape, or sequence of changing shapes, of the articulatory tract, and then into sounds. Now it is possible that in some instances the relation between the activation of a muscle and the resulting sound is simple, direct, and independent of neighboring movements, but this is surely rare. Given that the unit commands can and surely do overlap in time, and given all the interactions and constraints inherent in the anatomy and physiology of the vocal tract, we should expect to get the kind of scrambling that we do, in fact, find -- that is, a loss of segmentability and the frequent existence of a complex relation between acoustic cue and intended phoneme.[14]

Thus, the complexities we find in the acoustic signals were not always there. Accordingly, we can, perhaps, recover a simple relation to the language by getting back on the other side of the successive transformations by which the message was converted from neural signal to sound. We should remind ourselves here that the perceiver is also a speaker, and that he must, therefore, possess all the mechanisms for putting the segments through the successive recodings that result eventually in the acoustic signal. We wonder, then, whether it is necessary, or desirable, or even reasonable to endow him with an entirely different set of mechanisms for decoding that signal. We would prefer to assume that he has but one mechanism, one center, if you will, with some kind of link between sensory and motor areas.

At all events the kinds of data we mentioned a moment ago suggest that perception may be more closely related to articulation than to the acoustic signal. We should like to consider one more type of evidence for that generalization: this is an example which relates in a special way, we think, to the distinctiveness of speech, and which contributes, also, to the development of our thesis.

Categorical perception of some acoustic cues. We have found in speech perception a number of cases in which a continuously varied acoustic signal tends to be perceived categorically. In these cases the articulation is also categorical, which is, of course, the point. A specific example, and one of several we might choose, concerns the perceived distinction between the words "slit" and "split." Bastian *et al.* found first that a powerful and sufficient acoustic cue for this distinction is simply the duration of an interval of silence between the /s/ friction and the vocalic portion of the syllable.[15] They subsequently found, as others had found previously in work with other consonant cues, that perception of this cue is categorical in the sense that subjects tend to hear test stimuli, each with a different duration of silent interval, either as /slɪt/ or /splɪt/: given equal physical differences between successive acoustic signals, the listener discriminates very poorly within the range of a single phoneme, then experiences that amounts to a quantal jump in perception at the phoneme boundary.[16] In order to find out whether the articulation was also categorical, these investigators had the subjects attempt to mimic the experimental stimuli with their various durations of silent interval. Acoustic and electromyographic records of the subject's responses indicated that articulation was as categorical as the perception -- that is, the subject either closed his lips to articulate /splɪt/ or he kept them open to produce /slɪt/. There were no partial closures in response to stimuli near the center of the acoustic continuum.[17]

Here, then, is a situation in which the acoustic variation is continuous, articulation is categorical, and perception -- following articulation but not the acoustic signal -- is also categorical. There are other instances of this, in the perception of /b,d,g/,[18] /d,t/,[19] and /ræbId-ræpId/,[20] for example, though /slɪt-splɪt/ is the only case of categorical perception so far studied in which mimicry has been carefully measured and precise data on articulation have been obtained.

In this connection, we should say that in the perception of stress, which is, perhaps, not so categorical as the cases we just listed, Peter Ladefoged has found that the perception bears a simpler relation to the activity of respiratory muscles than to the properties of the acoustic signal. From this fact he concludes that in perceiving stress the listener is, in effect, perceiving the behavior that normally produces it.[21]

<u>The perception of speech and nonspeech</u>. If it is true that listeners hear speech in terms of the way they produce it, then speech signals and nonspeech signals must somehow be processed differently. We turn now to the evidence which suggests in one way or another that this may, in fact, be so. Consider again the case of /slɪt, splɪt/ and recall that discrimination of

equal physical differences in the acoustic cue was considerably better at the phoneme boundary than in the middle of the phoneme range. What happens when we ask subjects to discriminate essentially the same acoustic variable -- namely, the duration of a silent interval between a patch of noise and a complex tone -- but in a non-speech pattern? The answer is that there is then no peak in the discrimination curve but only the monotonic kind of function one most commonly finds in the perception of one-dimensional variations of a stimulus.[22] Essentially this same kind of experiment has been carried out for several other consonant distinctions, and the same kind of result has been obtained.[23]

Such results suggest that the discrimination peaks and the distinctiveness they provide are not inherent in the acoustic signal. Moreover, they would seem to indicate that experience with the acoustic variable, qua acoustic variable, is not enough. Learning to discriminate durations of silence of the magnitude that distinguish /slɪt/ from /splɪt/ did not generalize to the perception of that variable in a non-speech context; the discrimination peaks occurred only when those physical differences were in a signal that was itself heard as speech. One can interpret this, ad hoc, as pointing to some kind of perceptual interaction, or -- and this is our preference -- as indicating that the peak occurs only when the signal, being heard as a speech sound, somehow engages the special speech processing system. In our view, a prominent feature of that system is a reference to articulation.

These results are reinforced by other evidence which suggests that there is in perception no continuum between speech and non-speech. In one relevant experiment, House, Stevens, Sandel, and Arnold produced various degrees of acoustic approximations to speech.[24] Perceptually, there seemed to be no approximation. Their signals were heard as speech or they were not. If they were heard as speech, they were highly distinctive in the sense that they were easily learned in a paired-associate learning task. Those stimuli that were not heard as speech were not in this sense distinctive.

The many impressions one gets from everyday work with speech synthesis also support the notion that there is no gradual approximation to speech and, similarly, no gradual approximation to the distinctiveness of speech sounds. One seems either to be in the speech perception system or out of it.

It may also be relevant in this connection to consider the recent evidence that speech and non-speech are perceived on different sides of the brain.[25] These results don't tell us anything about the nature of the perceptual processes in either case, but the fact that speech and non-speech are so clearly dealt with in different places lends some support to the notion that they are dealt with in different ways.

Earlier we referred to two problems one might expect to encounter in trying to get the phoneme strings through the auditory channel. One was in connection with the temporal resolving power of the ear and the low ceiling set thereby on the rate at which discrete events can be heard. The other had to do with the difficulty of finding 40 acoustic signals that are highly identifiable though of short duration. We have already had occasion to remark how the encoding (and subsequent decoding) of the phonemes into the acoustic signal (and out of it) increases the rate at which strings of phonemes can be perceived. We should make explicit now how this encoding-decoding might improve identifiability. As has been implicit in the discussion so far, a code based on a reference to the articulatory system should enable the listener to take advantage of the relative independence of some parts of that system.

Given the existence of subsystems and the number of different muscle groups involved, we can suppose that the motor commands and feedback associated with their activation might be more highly discrete and identifiable than the corresponding acoustic cues. This should be especially so when different phonemes involve different muscles rather than related muscles in different degree.

Continuous perception of some acoustic cues. We have spoken of the complex relation between acoustic signal and perception. We should emphasize now that the relation is not always complex. Recall the case of /slɪt, splɪt/. There, acoustic variation was continuous, while articulation and perception were discontinuous. The articulatory situation for the vowels is quite different in that the speaker can presumably vary his articulation continuously. It is of some interest, then, that the perception of the vowels is different from that of the categorically articulated consonants in that vowel perception is more like that of most continuously varied signals: there are no quantal jumps in perception -- that is, no substantial increases in discrimination at the phoneme boundaries -- and the listener can discriminate many more stimuli than he can identify.[26] Here, then, is a situation in which acoustic signal, articulation, and perception tend to be in step: they all vary more or less continuously. There is, therefore, no basis on which we can say that a reference to articulation does or does not mediate the perception of these signals. We might conclude that we should not assume such mediation if we don't have to; alternatively, we might feel happier about keeping the whole speech perception system of one kind.

One model or two? It is, of course, possible that more than one mechanism is involved in speech perception. We have considered, for example, a duplex theory in which some phonemes would be perceived on an analysis-by-synthesis basis, but others would be processed straight through -- that is, in terms of their acoustic properties alone and without any reference to the means by which they are generated. Others, (for instance Chistovitch[27] and Ladefoged[28]), have entertained similar notions. If, then, duplexity -- I guess I should not say duplicity -- is a real possibility, we are interested in the kinds of experiments that might help us choose between the possibilities. More generally we are, of course, even more interested in how the difference between the two classes of phonemes can be characterized, whether we are driven to duplexity or not. Is it simply a question of categorical versus continuous articulation? Or is it, as K.N. Stevens has suggested to us, a matter of whether or not the important acoustic cue results from rapid movement of the articulators?[29] These questions can be answered by carrying out further research similar to that we have already described, but on a broader variety of phoneme classes.

However the question of mechanism is resolved, we will be left with the important observed differences between various classes of phonemes: the stops, for example, tend to be perceived categorically, while for the vowels and certain other phonemes, perception is very nearly continuous. We should guess that these differences would have important implications for distinctiveness -- that the categorically perceived stops would be more accurately and quickly identified than the continuously perceived vowels. We might also expect to find that phonemes with different degrees of distinctiveness would have different linguistic roles, or that they would, at least, carry different functional loads. There isn't much that can be said about these guesses at the present time except that it will surely be possible to collect the relevant facts.

Invariance and motor commands. Having just considered the vowels in which there is a direct relation between acoustic signal and phoneme perception we ought now to return to cases, which we made so much of earlier, in which the relation is not so direct. The particular point about these latter cases, as you will perhaps recall, is that an invariance is presumed to exist at some articulatory level. Our point, in effect, was that the speech message is processed through successive recodings and, indeed, circulates in a closed loop as we monitor our own acoustic output. At some point in this loop, there is a recoding which corresponds in discreteness of units and regularity of structure to our perception of the message, or more precisely, to the phonological constituents of the message. The argument, then, was that this correspondence almost certainly occurs at the level of the events in the central nervous system that initiate the process of which the speech sounds are an outcome. It is surely possible, perhaps likely, that the correspondence might still be very well preserved right down to the level of the motor commands that actuate the articulatory organs.

In any case, that possibility has the virtue of being empirically testable, and by taking electromyographic records we have for some time been trying to test it. Of course, the electromyogram actually measures the muscle potential that accompanies contraction. Such potentials are closely related to the neural commands that actuate the muscle, however, so we are able to make inferences about the motor commands. Unfortunately, the methods we use for the purpose of finding the EMG correlates of speech are not nearly so flexible or convenient as the methods now available for uncovering the acoustic correlates. Therefore we cannot, after only two or three years of research, make statemen:s that are firm and general. We have some data, however, and they warrant the following tentative conclusions.[30]

When we deal with temporally overlapping phonemes that are articulated by different groups of muscles -- such as /f/ and /t/, for example -- we can find an invariant EMG tracing for each phoneme, regardless of the context in which the phoneme occurs.[31] We should keep in mind that such invariance is most commonly not to be found in the acoustic signal.

For temporally overlapping articulatory gestures involving more-or-less adjacent muscles that control the same structure -- such as /t/ and /i/, for example -- it is, for obvious reasons, more difficult to discover what is going on. In several cases we think we have found evidence that some parts of the gesture may be reorganized when the phoneme appears in different contexts. Even in these cases, however, there appears to be a common core of EMG activity that remains invariant.

We have several times said that while the language is segmented at the phoneme level, the acoustic signal is not. This is to say that there are no marks in the acoustic signal by which one can determine where one phoneme ends and another begins. In a strict sense, this holds also for the EMG signals, but with a difference: the separate pen traces from different muscles give information about independent dimensions of the articulation to a far greater extent than do the component parts (e.g., the formants) of a sound spectrogram. Hence, the onsets and offsets of EMG activity on the several traces represent a kind of segmentation, dimension by dimension, of the articulatory event. We see, then, that in this important sense, too, the motor commands bear a simpler relation to the perceived phonemes than does the acoustic signal.

We say again that all of this is necessarily very tentative. As things stand, however, we are encouraged to believe that the

EMG correlates of the phoneme will prove to be invariant in some significant sense. At all events, we very much hope that this will be so, not in the interests of a motor theory broadly conceived, since it is always possible to push the assumed invariance farther upstream, but simply because the motor commands are about as far upstream as we are likely to go experimentally. If we find invariance there, we shall have the solid basis for a simple description of the phonological structure of the language. Surely we can all agree that such invariance, or lack of it, is an essential consideration in the development of a model of phonological production. We think it is essential, too, for a model of phonological perception, for, as we have said, our strong inclination is to assume that phonemes are perceived by reference to the way they are produced. In this view, production and perception are two aspects of the same process. If that general concept is also useful for a perceptual model at the level of morphology and syntax, then at some future time it may be easier to produce a unified model that explains speech in the most general case.

Footnotes

Paper given at the AFCRL Symposium on Models for the Perception of Speech in Visible Form, Boston, Mass., 11-14 November 1964.

The research described in this paper was supported in part by the National Science Foundation under grant G23633, the National Institute of Dental Research under grant DE-01774 and the Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration under contract V1005M-12353. This paper was written while one of the authors, A.M. Liberman, was a Fellow at the Center for Advanced Study in the Behavioral Sciences and also a Research Fellow, National Institutes of Health.

Also, University of Connecticut.

Also, Barnard College, Columbia University.

See: A.M. Liberman, P.C. Delattre and F.S. Cooper. The role of selected stimulus variables in the perception of the unvoiced stop consonants. Amer. J. Psych. 65, 497-516 (1952); A.M. Liberman. Some results of research in speech perception. J. Acoust. Soc. Amer. 29, 117-123 (1957); F.S. Cooper, A.M. Liberman, K.S. Harris and P.M. Grubb. Some input-output relations observed in experiments on the perception of speech. 2nd Internat'l Congress of Cybernetics 930-941 (1958) Namur, Belgium; A.M. Liberman, F.S. Cooper, K.S. Harris and P.F. MacNeilage. A motor theory of speech perception. Proc. Speech Communication Seminar (1962).

Royal Institute of Technology, Stockholm.

2. K.N. Stevens. Toward a model for speech recognition. J. Acoust. Soc. Am. 32, 47-55 (1960); M. Halle and K.N. Stevens. Speech recognition: a model and a program for research. IRE Transactions on Information Theory IT-8, No. 2 155-159 (1962).

3. M.Halle and K.N. Stevens. Analysis by synthesis. Proc. of the Seminar on Speech Compression and Processing II, AFCRC - TR - 59 - 198 (1959). Edited by W. Wathen-Dunn and L.E. Woods.

4. It was on the basis of such oversimplified assumptions about the relation of sound segment to phoneme that R.H. Stetson was led to conclude that the phoneme did not exist as a psychological unit. See: R.H. Stetson, Motor Phonetics, 2nd edition, North-Holland Publishing Co., Amsterdam, The Netherlands (1951).

5. See: F.S. Cooper. "Research on Reading Machines for the Blind" in P.A. Zahl (ed.) Blindness: Modern Approaches to the Unseen Environment, Princeton University Press, Princeton (1950) 512-543; H. Freiberger and E.F. Murphy. Reading machines for the blind. IRE Professional Group on Human Factors in Electronics, 8-19 (March,1961); J.L. Coffey. The development and evaluation of the Battelle Aural Reading Device. Proc. Internat'l Congress on Technology and Blindness. American Foundation for the Blind, New York (1963) Vol. 1, 343-360.

6. But see: I. Pollack. The information of elementary auditory displays. J. Acoust. Soc. Am. 24, 745-749 (1952); I. Pollack and L. Ficks. Information of elementary multidimensional auditory displays. J. Acoust. Soc. Am., 26, 155-158 (1954); M. Studdert-Kennedy, A.M. Liberman and R.J. Rosov. Rate of information transmission for one-

two- and three-dimensional acoustic stimuli. J. Acoust. Soc. Am. 35, 808(A) (1963).

7. See: C.M. Harris. A study of the building blocks of speech J. Acoust. Soc. Am. 25, 962-969 (1953); G. Peterson, W. Wang and E. Sivertsen. Segmentation techniques in speech synthesis. J. Acoust. Soc. Am. 30, 739-742 (1958); and A.M. Liberman, F. Ingemann, L. Lisker, P. Delattre and F.S. Cooper. Minimal rules for speech synthesis. J. Acoust. Soc. Am. 31, 1490-1499 (1959).

8. In these cases we should distinguish two kinds of invariance lack. The one kind -- and this is the kind we are particularly interested in here -- includes all cases in which very different acoustic shapes lead to essentially indistinguishable perceptions. In the other kind, different acoustic shapes lead to discriminably different perceptions which are nevertheless classed as members of the same phoneme. An example of the latter kind is found with the vowels, where intra-phonemic variations are easily heard. (See: D. Fry, A. Abramson, P. Eimas and A.M. Liberman. The identification and discrimination of synthetic vowels. Language and Speech 5, 171-189 (1962).)

9. See: A.M. Liberman, P. Delattre and F.S. Cooper. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. Am. J. of Psych. 65, 497-516 (1952); A.M. Liberman, P. Delattre, F.S. Cooper and L. Gerstman. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs 68, No. 8 1-13 (1954); P.C. Delattre, A.M. Liberman and F.S. Cooper. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773 (1955); and A.M. Liberman. Some results of research on speech perception. J. Acoust. Soc. Am. 29, 117-123 (1957).

10. See: P.C. Delattre, A.M. Liberman and F.S. Cooper. *Op. cit.*

11. Ibid.

12. If, with synthetic speech, one carries the transition all the way to the /d/ locus, for example, thus converting an imaginary extrapolation of the acoustic cue into a true part of it, a listener will hear /d/ in only a very few vowel contexts; in all others he will hear /b/ or /g/, as in the example given earlier in the text and referenced in Fn. 10.

13. See: K.N. Stevens and A.S. House. Studies of formant transitions using a vocal tract analog. *J. Acoust. Soc. Am.* *28*, 578-585 (1956)

14. See: F.S. Cooper, A.M. Liberman, K.S. Harris and P.M. Grubb. *Op. cit.*; L. Lisker, F.S. Cooper and A.M. Liberman. The uses of experiment in language description. *Word* *18*, 82-106 (1962)

15. J.Bastian, P. Eimas and A.M. Liberman. Identification and discrimination of a phonemic contrast induced by silent interval. *J. Acoust. Soc. Am.* *33*, 842 (A) (1961)

16. Ibid.

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

18. See: A.M. Liberman, K.S. Harris, H. Hoffman and B. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *J. Exptl. Psych.* *54*, 358-368 (1957); B. Griffith. A study of the relation between phoneme labeling and discriminability in the perception of synthetic stop consonants. Unpublished Ph.D. dissertation, U. of Connecticut, 1958; and P. Eimas. The relation between identification and discrimination along speech and non-speech continua. *Language and Speech* *6*, 206-217 (1963).

19. See: A.M. Liberman, K.S. Harris, J. Kinney and H. Lane. The discrimination of relative onset-time of the components of certain speech and non-speech patterns. J. Exptl. Psych. 61, 379-388 (1961).

20. See:A.M. Liberman, K.S. Harris, P. Eimas, L. Lisker and J. Bastian. An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. Language and Speech 4, 175-195 (1961).

21. P. Ladefoged, M. Draper and D. Whitteridge. Syllables and stress. Phonetica 3, 1-14 (1958).

22. J. Bastian, P. Eimas and A.M. Liberman. Op. cit.

23. See: A.M. Liberman, K.S. Harris, J. Kinney and H. Lane. Op. cit.; A.M. Liberman, K.S. Harris, P. Eimas, L. Lisker and J. Bastian. Op. cit.

24. A. House, K.N. Stevens, T. Sandel and J. Arnold. On the learning of speech-like vocabularies. J. Verbal Learn. and Verbal Behav. 1, 133-143 (1962).

25. D. Kimura. Cerebral dominance and the perception of verbal stimuli. Canad. J. Psych. 15, 166-171 (1961); idem. Some effects of temporal-lobe damage on auditory perception. Canad. J. Psych. 15, 156-165 (1961); idem. Left-right differences in the perception of melodies. Quart. J. exp. Psych. 16, 355-358 (1964); D.E. Broadbent and M. Gregory. Accuracy of recognition for speech presented to the right and left ears. Quart. J. exp. Psych. 16, 359-360 (1964); B. Milner. "Laterality effects in audition" in V.B. Mountcastle (ed.) Interhemispheric Relations and Cerebral Dominance, Johns Hopkins Press, Baltimore (1962).

26. See: D. Fry, A.S. Abramson, P. Eimas and A.M. Liberman. The identification and discrimination of synthetic vowels. *Language and Speech* 5, 171-189 (1962).

27. L.A. Chistovich. Continuous recognition of speech by man. *Machine Transl. and Applied Linguistics* *7*, 3-44 (1962).Mosc.

28. P.Ladefoged. "The perception of speech" in Mechanization of Thought Processes, H.M. Stationery Office, London (1959).

29. Personal communication.

30. K.S. Harris, M.M. Schvey and G.F. Lysaught. Component gestures in the production of oral and nasal labial stops. *J. Acoust. Soc. Am.* *34*, 743 (A) (1962); P.F. MacNeilage, Electromyographic and acoustic study of the production of certain final clusters. *J. Acoust. Soc. Am.* *35*, 461-463 (1963); K.S. Harris. Behavior of the tongue in the production of some alveolar consonants. *J. Acoust. Soc. Am.* *35*, 784 (A) (1963); and P.F. MacNeilage. Electromyographic study of the coarticulation of vowels and consonants. *J. Acoust. Soc. Am.* *36*, 1989 (A) (1964).

31. It is not, of course, to be expected that all the muscles that are active during the articulation of a given phoneme will show a close and invariant correspondence between the EMG signal and that phoneme, nor, in fact, do they; the strongest assumption would be that some particular muscle (or closely related set of muscles) would show such highly correlated behavior. The experimental problem is, then, to search the plausible locations for EMG signals that are diagnostic in this sense. We have found several instances of such signals,all in situations of temporal but nonspatial phoneme overlap. (See the references cited in the preceding footnote.)