# Some input-output relations observed in experiments on the perception of speech [1]

by Franklin S. COOPER, Alvin M. LIBERMAN [2], Katherine S. HARRIS, and Patti Murray GRUBB (U. S. A), *Haskins Laboratories.*

---

It may be useful, at the start of this paper, to explain briefly what it is about and how it relates to the general topic of cybernetics. The experiments on which my remarks [3] will be based have been concerned with the perception of speech, with particular emphasis on the acoustic entities that we use to recognize the sounds of our language. Much of this work has been done with synthetic speech, which has the unique advantage that one can manipulate the stimulus sounds without the usual limitations imposed by the human articulatory apparatus. The results of these studies have appeared in a number of publications by my colleagues at the Haskins Laboratories [4].

It is not, however, speech *per se* that I wish to consider but rather some hypotheses concerning the perceptual mechanisms that operate when we listen and when we speak. In particular, I wish to speak in rather general terms about some relations between inputs and outputs that we think are involved in the perception of speech, and that we have found interesting in their own right and provocative of additional experiments. I shall say a little about the sounds of speech, but only as a basis for these more general remarks.

---

1. This research was supported by the Carnegie Corporation of New York and by the Department of Defense in connection with Contract DA49-170-sc-2564.

2. Also University of Connecticut.

3. The paper was presented at the Congress by Franklin S. COOPER.

4. A review of the research and references to the literature are given in *Some results of research on speech perception* by Alvin M. LIBERMAN, J. Acoust. Soc. Am., 29, 117-123, 1957.

A very useful technique in studying speech is to convert the extremely complex sounds into a visible display, so that one may enlist vision as an aid in dealing with a problem that is primarily auditory. In recent years, the sound spectrograph has become a standard tool for the study of the acoustic correlates of speech. The spectrogram has been called — very aptly — " visible speech " for, indeed, the same underlying pattern can be seen in spectrograms of a given word when it is spoken by various speakers or in various contexts. The experimental work that I mentioned a moment ago has been a direct extension of the spectrographic approach : we have studied the underlying patterns, then redrawn them in a much simplified form, and finally converted them back into sound so that a listener could tell us what he heard. This procedure has proved to be a powerful research tool ; its essential virtue is that one can deal with auditory phenomena *as if* they were visual patterns.

Now it may be laboring the obvious to ask *why* this is so but, with your permission, I shall proceed, since the observed fact that audible and visible patterns can be interconverted points to an important relationship between inputs in different sensory modalities. This relationship, when made explicit, has interesting implications for perceptual mechanisms and for the possible existence of as yet-undiscovered stimulus transformations. Perhaps the most general implication is that there may exist an important similarity in the perceptual processes of vision and audition, presumably at a rather high level in the nervous system — a functional similarity so close that there must also exist, in principle at least, a recoding procedure external to the organism that will preserve inherent structure in the stimulus data. Such a recoding procedure, or interconvertibility relationship, would then preserve patterning in the sense that stimulus changes which do not impair the perceptual identification of visible patterns would not, when recoded, impair identification of the audible patterns either. In short, stimuli that look alike can, by appropriate recoding, be made to sound alike.

Let us return to the sound spectrogram as an example of this recoding procedure (see Figure 1) [1]. Here, the fleeting sound of the spoken word has been transformed, dimension by dimension, into a static picture. Frequency and time of the sound have become the vertical and horizontal coordinates of the picture, with intensity of sound retained as intensity of blackening in the picture. But why should it be a *picture* ? — unless it be that this particular recoding is the appropriate intersensory transform, or a first approximation

1. The figures are reproduced by courtesy of the *Journal of the Audio Engineering Society*.

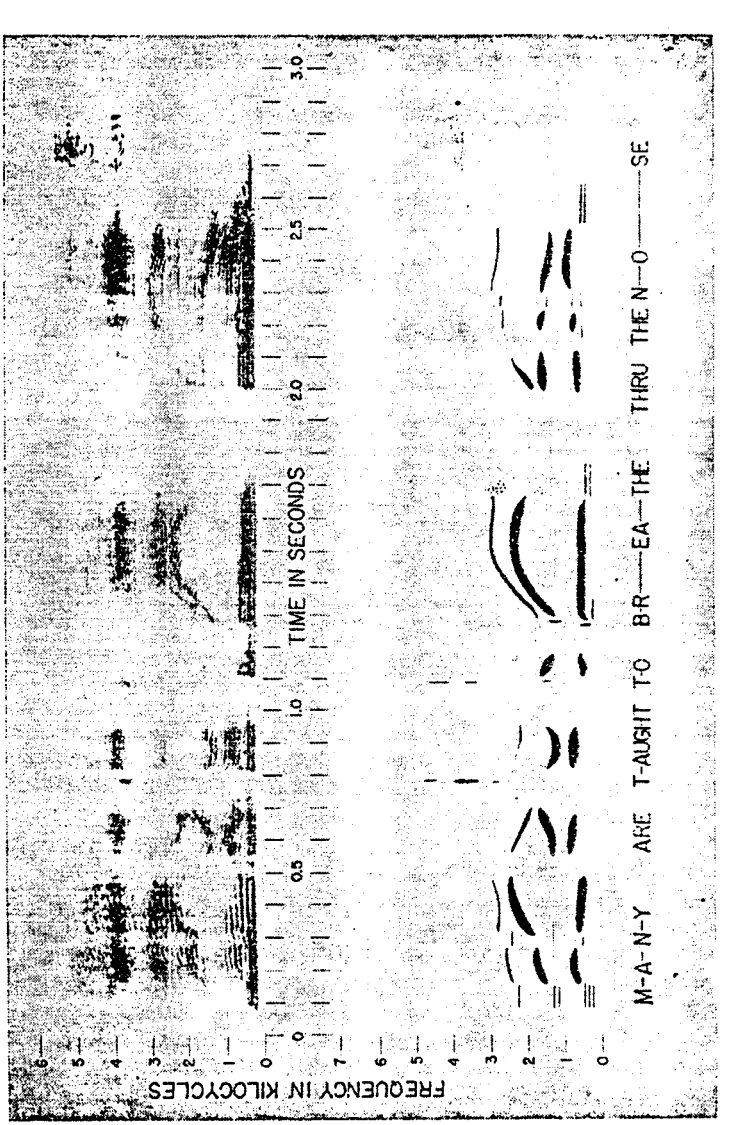
FREQUENCY IN KILOCYCLES
TIME IN SECONDS
M-A-N-Y ARE T-AUGHT T-O B-R——EA—THE THRU THE N—O———SE

Fig. 1.

to it. Certainly it is a better approximation than the oscillogram, at least for speech, and it has a long history of usefulness in music, in the slightly modified form of the musical score. Even though this transform is less than perfect — witness the inadequate portrayal of rhythm and harmony — it does suggest some interesting experiments on pattern perception. Thus, the visual patterns of such familiar geometric figures as circles, ellipses, triangles, and squares can be converted into sounds and one can then ask whether these sounds fall into disjunct auditory categories, and how resistant each category will be to variations in size, position, and deformation of the visual figure. Or again, looking back into the organism, the bisensory approach should be a powerful tool for investigating how stimuli are organized in perception and what kinds of inherent structure in the stimuli will be accepted as perceptual entities by the organism. Such perceptual laws will have gained generality by the very fact that they are operative for diverse modalities.

Speculation need not end here : one can search, with reasonable hope of success, for stimulus transformations among sensory modalities other than vision and audition. Indeed, I should like to indulge in some of this speculation by returning, a little later, to the possibility of a transform between articulatory gesture and vision. This is an interesting possibility because — as I shall now try to show — it appears that articulation is intimately involved in the perception of speech.

The relationship between articulation and perception can best be explained by describing some of the experimental work on the perception of speech sounds. I have mentioned already the technique by which we manipulate speech in terms of its visual patterns and then judge the resulting synthetic speech by ear. Most of you know already that the patterns of real speech can, in this way, be simplified drastically (as they are in the lower part of Figure 1) and yet retain a high degree of intelligibility. This finding led to a lively search for the acoustic invariants that characterize the significant sounds, or phonemes, of the language. Some workers attempted to find or construct a set of acoustic " building blocks " — sounds that could be assembled in any order like moveable type ; these efforts have not been very successful. The acoustic signals for the perceptually disjunct sounds of the language seem, rather, to overlap and to be intimately intermixed. To put it in another way, speech seems to be an encoding rather than an encipherment of the intended message.

I do not wish to imply that there are no clear acoustic signs for the individual sounds. The researches conducted in several laboratories, including our own, have now provided a fairly complete description of these acoustic " cues " for at least one language and partial descriptions for several others.

What are some of these cues ? You would expect, of course, that the short, explosive bursts in sounds like *t* and the longer stretches of noise as in *s* would contain cues for the stop and fricative consonants. The vowels are characterized by comparatively steady-state formants — the frequency regions where there is resonant reinforcement by the mouth cavities. It is characteristic of spectro-

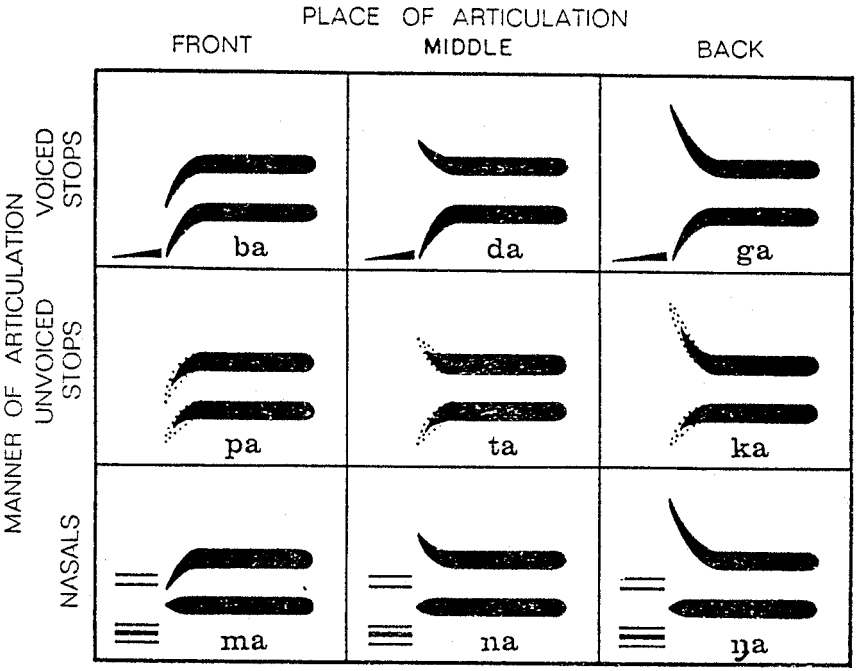


FIG. 2.

graphic patterns that the formants are usually shifting about in frequency, thereby reflecting the changing shapes of the oral cavities during speech.

Experiments with synthetic speech have made it clear that these same shifts, or transitions, are themselves important cues for the perception of many consonant phonemes. In Figure 2 are some highly simplified spectrograms — patterns that were painted by hand to resemble spectrograms of spoken syllables, but with only the most essential features retained. We know that essential features have been retained since, when these patterns are converted back

into sound, they are easily understood, even by an uninitiated listener [1]. You can see, in the top row, that the major differences among *b*, *d*, and *g* lie in the transitions of the second-formant (counting up from the bottom of the pattern). These three transitions are different in direction and extent. And, if you will scan across the middle and bottom rows, you will see that the same transitions serve to distinguish among the sounds in those rows, too. You will have noticed also that each row is distinguished by some characteristic held in common : in the top row, the first-formant transition extends all the way from zero frequency to the steady state of the vowel and is preceded by a low frequency " voice bar " ; in the middle row, the first-formant does not start so low in frequency and both formants start with aspiration ; and in the bottom row there is a brief, steady-state resonance common to all three sounds.

Thus, having found by trial and error the appropriate patterns for these nine sounds, we observed certain resemblances and found it possible to group the patterns — as they are grouped in the rows and columns of the figure — on the basis of acoustic similarities. An interesting thing about this arrangement is that the same grouping of the sounds that correspond to these acoustic patterns, fits precisely the usual phonetic classification of the sounds by place and manner of articulation ; there is no need to redraw the figure — we have only to add to the rows and columns a set of headings drawn from conventional articulatory phonetics.

What is, then, the real basis of classification — acoustic pattern or articulation ? These data cannot tell us, since they fit equally well within either frame. Indeed, we should expect them to do so, since the phonetic classifications are basically in terms of the shape of the mouth cavities and since articulatory shape determines uniquely the acoustic " shape " of the spoken sounds from instant to instant. But let us examine some additional data, in particular, the patterns for some of the same consonants with a variety of vowels. Figure 3 shows the voiced stop consonants with each of seven vowels that sample the normal range from front to back. You will notice, of course, that the formant positions are different for the different vowels and that all the first-formant transitions start at zero frequency. Again, it is the second-formant transitions that carry the differential information about the three consonants, though now it is not easy to characterize these second-formant transitions on a row-by-row, or consonant-by-consonant, basis — not easy, that is, if you will ignore the dotted lines and look only at the patterns them-

1. These are not, however, the best nor the most realistic sounds that one can generate synthetically. This was not the objective ; rather, it was to strip down the spectrogram to its essential pattern.

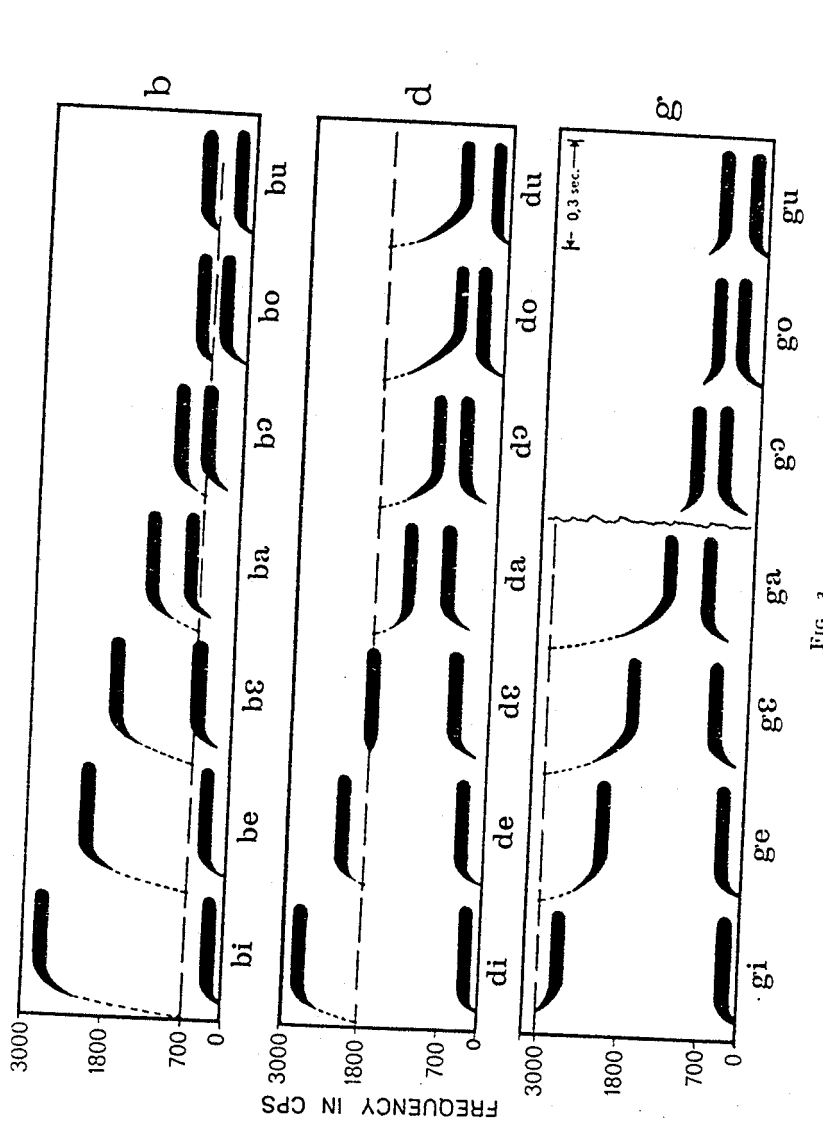
b
d
g
FREQUENCY IN CPS
3000
1800
700
0
bi
be
bɛ
ba
bɔ
bo
bu
di
de
dɛ
da
dɔ
do
du
gi
ge
gɛ
ga
gɔ
go
gu
0,3 sec.

FIG. 3.

selves. The middle row is particularly troublesome ; the patterns for *d* seem to sample the total available range of second-formant transitions. On the other hand, if we make a rather simple-minded asumption about the articulation of syllables that begin with *d* — and the acoustic consequences of that articulation — we would suppose that the relatively invariant starting position for the articulatory movement will be mirrored in a fairly constant starting frequency, or *locus*, for the transitions. This may be an implicit locus rather than an explicit one if the onset of the sound comes a little late. We have, on this basis, a rather simple way to characterize the patterns for *d* : the second-formant " points back " to a starting frequency of about 1800 cps, as indicated by the dotted line. There is still no firm basis for choice between articulation and acoustic pattern as a frame on which to hang the data in orderly array. There is, though, a plausible model in the articulatory case and only an arbitrary rule for dealing with the data as acoustic patterns.

We have attributed this lack of a basis for choice to the fact that the acoustic pattern follows so directly from the articulatory shape. There are, however, a few cases in which comparatively small changes in articulation result in rather large changes in acoustic pattern. These are isolated cases, but crucial for our purpose, since now the perception itself must " choose " between an articulation that changes but little, and an acoustic pattern that changes much and abruptly. Will the identity of the sound change abruptly when the pattern changes, or will it, like the articulation, remain essentially the same ?

A case in point is the series of second-formant transitions for *g* with various vowels. The apparent anomaly in the lower right-hand corner of Figure 3 is shown again in the top row of Figure 4. Here, the large and abrupt change in acoustic pattern contrasts with the comparatively small and continuous shifts in articulation, shown diagramatically in the bottom row. The identity of the perceived consonant remains the same throughout, and thus parallels the invariance of the articulation.

This strongly suggests, as do other cases we have studied, that speech is perceived by reference to articulation. This is not, of course, to say that we must overtly mimic the incoming speech sounds as a necessary preliminary to recognition ; for the adult, with his tremendous experience in both talking and listening, we can and must assume that the neural operations involved in explicit mimicry are somehow short-circuited and that we are dealing instead with processes of implicit mimicry and of response *as if* to proprioceptive return. The important point is that the perceptual operation, however it may be mechanized, is organized in articulatory terms.

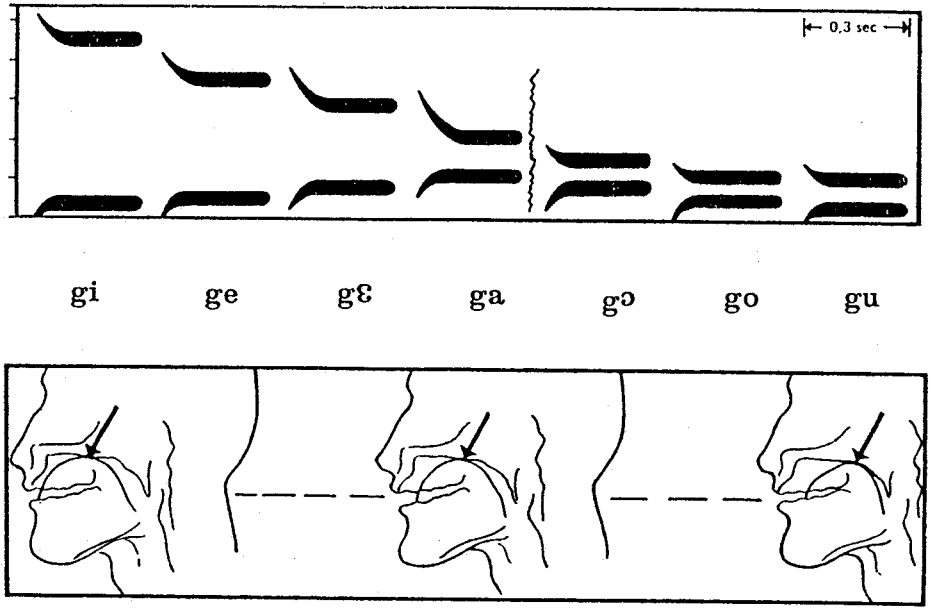


FIG. 4.

Such an interpretation serves to shift attention from reception by the listener to production by the speaker. For the latter, at least when he is generating his own output, we have no direct knowledge of the input to his speech system. We can — perhaps we must — assume that this input is, at some level, an intended sequence of phonemes that will shortly be observable in his output speech. Let us now try to work backward from the observed stream of speech sounds toward this presumed input. One of the intermediate outputs available to us is the changing shape of the articulatory apparatus. We can observe this output by looking at the speaker's lips or by looking through his head with x-rays or by accepting the reports he gives us of his gestures. These descriptions are the content of conventional articulatory phonetics. The relationship of this output to the acoustic output is, of course, quite close ; it is not necessarily a simple relationship but it is computable at any instant of time and so the signal conversion from shape to sound would be classed as a rather involved substitution cipher.

Let us now move back another step and examine the articulatory signals at the neuro-muscular junctions. Here, the commands arriving by way of the motor nerves give rise to contractions of the muscles that actuate the articulatory apparatus. The resulting gesture can be specified in terms of the identities of the muscles that are activated, the intensity of this activity, and the relative timing of activation in different muscle groups. If we can tolerate gross simplifications at this level, as we were able to do at the level of the acoustic pattern, we may hope to describe speech events in terms of a rather limited number of muscle groups responding at only a few levels, perhaps on an all-or-none basis, and with relative timings that need not be specified very precisely — in short, a rather simple description in terms of what actuators are used, when, and how forcibly. This will be a multidimensional description, since we shall need one dimension per major muscle group, but rather coarse-grained in its time and intensity quantizations. Such an " *action pattern* ", if I may refer to it in this way, could well be a simpler description than we can hope to find at the acoustic level, in the sense that the components of the action pattern may correspond more immediately to the intended phonemes. All of this is speculation, to be sure, but there is some basis for optimism : we have taken a long step toward the level at which simplicity is assumed to exist and, in particular, we have bypassed one of the encoding procedures that could hardly fail to introduce complexity into its output. I refer, of course, to the dependence of the *shape* of the articulators on the *manner* and *timing* of their actuation. The dependence is a complex one and it can no longer be computed, even in principle, without taking account of temporal stretches of the order of syllabic length.

Let us suppose then that the reasons for studying the action pattern are as attractive as I have suggested. We should consider two points : what instrumental measures can we take of the action pattern, and what conceptual tools would be most helpful in planning experiments and interpreting results ?

The conceptual tool that we have come to depend upon in working at the acoustic level, derives its utility — so we think — from the existence of an intersensory transform between vision and audition, plus knowledge of an approximate realization of that transform, namely, the sound spectrogram. My colleagues and I, having worked with the spectrographic transform, have no doubt whatever of its value or of the potential value of an appropriate visual representation of the action pattern. Such a transform would be of enormous help as a conceptual tool, and possibly later as an experimental one. I must admit that I find it none too easy to think of the anatomical picture — the musculature of the articulators and the timing and intensity of their actions — as a simple array ; nor do I feel at ease with a binary or trinary representation in, say, eleven dimensions. The desired transform should be much simpler than this. I wish that I could exhibit it to you but, unhappily, I cannot ; however, in the transform hypothesis, I do find reason to believe that such a transform exists.

The instrumental methods for obtaining action patterns are, however, at hand. We do not even need to tap the neural messages on the motor pathways since they regularly herald their arrival by a burst of electrical activity from the muscle itself. These muscle action potentials can be recorded from electrodes placed on the surface of the face and tongue or from needle electrodes in the appropriate muscles. There are no special problems in determining timing and intensity of muscle activity ; there may be some difficulty in separating the signals from adjacent muscles, though this may not be too serious if one is concerned primarily with muscle groups rather than with individual muscles — that is, if one is looking for the coarse-grained picture rather than the fine details.

It would be pointless to attempt a description of our preliminary results in studying action patterns. We have just begun to work in this area and much of the effort has necessarily gone into the development of techniques such, for example, as putting electrodes on the tongue by using small sunction cups. We do have even now some findings that parallel our observations on spectrograms and our expectations from articulatory phonetics, but we do not yet have an adequate body of data to serve as the basis for a systematic description of speech.

In summary, I have attempted to review for you some ten years

of work on the perception of speech sounds, not in terms of the mass of experimental details or even the generalizations about speech that have emerged, but rather in terms of such perceptual relationships as the intersensory transform between the auditory patterns of words and their spectrographic pictures, and the very close relationship that seems to exist between the perception of speech sounds and the articulatory gestures involved in their production. I have tried to show how these considerations provide the rationale for a further and perhaps even more penetrating experimental approach to the problem of characterizing the phonemes of our language in measurable terms — and, happily, the simplest possible terms.

Finally, by studying the structure of messages at various levels of the speech process, we may hope also to gain some insights into perception, the end effect of communication in man and perhaps — who can guess how soon ? — in machines, as well.