Why Are Spectrograms Hard to Read?[†]

Alvin M. Liberman,[*] Franklin S. Cooper, and Michael Studdert-Kennedy[**]
Haskins Laboratories, New York City

Some of what we know about speech bears on the practical problem that concerns us all today: how to provide the deaf with a useful nonacoustic representation of the speech signal. At the risk of having to say what you already know very well, I would like to review some of our knowledge about speech and consider its implications for our practical problems.

I think we can all agree that visual and tactile representations of speech are hard to read or feel -- that is, the speech signal makes much better sense to the ear than it does to the eye or the skin. My purpose is to ask why this is so. Why are the sounds of speech perceived so well by ear and so poorly by any other organ? For convenience, I will talk only about perception of speech by eye, having in mind that what I say may apply to other non-auditory pathways as well. We arrive, then, at the more specific question: why are spectrograms (and other visual representations of the speech signal) so hard to read?

By way of answer I want to say most generally that the reasons are not to be found at the surface. They do not lie in the characteristics of the peripheral sensory apparatus; they do not derive primarily from superficial inadequacies in the sensory transform; and they are not merely matters of training or experience. The reasons we seek are subtle, and they lie deep.

In the simplest of all speech worlds, the phonemes would be transmitted by an acoustic alphabet. That is, there would be an

acoustic shape -- a unit sound -- for each phoneme. We should have to suppose, of course, that the sounds were discriminably different, but this would be a low-level problem in auditory psychophysics. And if we were to assume that the connection between sound and phoneme had somehow to be established, we should only have made the easy discovery that speech is in some sense a habit and raised again the familiar questions about reinforcement and other conditions of learning.

If speech were like that -- if it were a simple alphabet on the phonemes -- then we might expect that some simple conversion from sound to sight would produce patterns that would be not only visible, but also readable. The point of my talk is that visible speech is hard to read largely because speech is not a simple alphabet. Speech is, rather, a complex code in the sense that the phonemic message is quite drastically restructured at the level of sound. As a consequence, the acoustic signal corresponding to a particular phoneme is typically different in different phonemic environments. Worse yet, from the standpoint of one who is trying to read spectrograms, definable segments of sound do not correspond to segments at the phoneme level. All this is part of a complex coding -- a grammar if you will -- that makes the sounds of speech highly efficient as vehicles for the transmission of phonemic information. But these encoded sounds can be efficient vehicles only if there is a decoder -- a special device that processes the complex signal so as to recover the string of phonemes. There is such a device in human beings, but, unfortunately for those of us who would send speech through the eye, that decoder can be made to work only from an auditory input. It does not process speech signals (in spectrographic form, for

example) that come in through the eye, and it cannot be made to do so by training. But let me start from the beginning and try to say more clearly, if still quite briefly, what I mean.

To see our problem, it helps to ask, first, whether speech could, in any event, be alphabetic. That is, could language be efficiently communicated by sounds that stand to the phonemes in essentially one-to-one fashion? To answer that question with an emphatic and unqualified "No" requires no special knowledge about speech perception, but only a passing acquaintance with the most obvious requirements of linguistic communication and the equally obvious properties of the ear.

Consider, first, that we can perceive artifically speeded up speech at rates that require us to take in as many as 30 phonemes per second. If each phoneme were represented by a discrete sound, we would, in listening to 30 per second, surely hear an unanalyzable buzz. It would not help that normally we can perceive speech without correctly hearing each and every phoneme. We must, in any circumstance, hear and identify some reasonable number of phonemes, but at thirty discrete sounds per second it seems most unlikely that we could sufficiently resolve the signal so as to hear any at all.

Thus, simply resolving the units would surely be a problem if phonemes were to be transmitted serially, as in an acoustic alphabet, but it would not be the only problem. We might also expect to have difficulty finding as many identifiable acoustic signals as we would need, though it is not especially relevant to our purposes to consider this difficulty here.

At this point you might be supposing that the difficulties I have spoken of are somehow illusory or at least exaggerated.

To show that they are real, and that they do loom large, one need only look at the results of work with nonspeech ciphers on the language, including in particular the results of more than fifty years of research in the attempt to build reading machines for the blind. The difficulty there has not been to transform print into sound, but to find a set of nonspeech sounds that will work. Many people have practiced for months, years, even decades, and with a variety of sound signals, yet top speeds are less than a tenth of the rate we can attain with speech.

Since our concern today is with delivering a representation of the speech signal by eye, we should pause here to note that the particular limitations on auditory perception just discussed do not seem to apply to vision. Since the eye perceives in space, there ought, in principle, to be no special difficulty presented by the need to take in strings of discrete phonemes; nor should we expect to have difficulty finding as many optical shapes as we need. We should suppose, then, that an optical alphabet would work, and we do, indeed, find that an alphabet is the most efficient way to communicate language by eye. We also know, however, that reading this simple cipher is a good deal less easy and natural, psychologically, than hearing the complex speech code. At all events, perceiving language by ear and by eye are bound to be, in certain very important respects, quite different processes.

Returning now to speech perception by ear, we are not surprised to discover that the sounds of speech are not an alphabet on the phonemic structure of language, but, as I said before, a special code. And given that speech is in general well perceived, we are not surprised to find that the code is well designed to evade the special limitations of the ear. But what

is the nature of the speech code, and what is the evidence for it? Research on the acoustic basis for speech perception shows that in many cases a single acoustic cue carries information in parallel about successive phonemic segments. It is this parallel delivery of information that makes it possible for us to evade the limitations imposed by the temporal resolving power of the ear and to hear speech as fast as we do. But the advantage is dearly bought, since in the case of speech parallel delivery of information can be accomplished only at the cost of a very complex relation between acoustic cue and perceived phoneme: the cue for a particular phoneme is very different according to context, and there are, in these cases, no commutable acoustic segments of phonemic size. Consider, for example, the case of a simple syllable consisting of a stop consonant followed by a vowel, surely one of the most common and important kinds of syllables in the language. The most important acoustic cue for place of production is the transition of the second formant. We know that the physical specifications of this transition -- in particular, the frequencies at which it starts and stops -- will be quite different for the same consonant segment followed by different vowels. We also know about such syllables that there are not two acoustic segments corresponding to the two phonemes, but only one. That is, there is no way to cut the acoustic signal so as to recover a segment that will by itself be perceived as only the voiced stop. This is because the consonant and the vowel are overlapped in production and because the gestures appropriate to each are influencing the same important part of the acoustic signal, the second-formant transition.

The kind of complex relation just described is very common. It is found for almost all the cues for almost all the consonants.

The exceptions among consonants are the relatively long-duration noise portions of the fricatives (in slow articulation). And, of course, vowels are also an exception. But the complications caused by the parallel delivery of information characterize the acoustic signals for most of the phonemes, including in particular those that bear the heaviest information load.

It would be well, perhaps, to recapitulate what I have so far tried to say. First: given the properties of the ear, it seems clear that phonemes could not be rapidly communicated by means of an acoustic alphabet. Second: it is now obvious from attempts to do precisely that, as in the work on reading machines for the blind, that, in fact, it cannot be done well. Third: when one uncovers the acoustic cues for phoneme perception, he finds in general that they are not an alphabet, but a special and complex code. A salient feature of that code is that, at every instant, the same, single aspect of the acoustic signal (e.g., the second-formant transition) provides information about more than one phoneme.

By encoding the phonemes so as to deliver information about successive phonemes in parallel, we reduce the number of discrete acoustic segments that must be heard per unit time, and so avoid the limitations imposed by the temporal resolving power of the ear. In this respect the code is well designed and ought to work well, provided there is, in perception, a device to decode the complex acoustic signal and recover the phonemes. Because there is such a device available to the auditory system, all is well when speech is presented to the ear. Assume, now, that no such decoder is available to the eye and consider, then, how difficult it must be to read spectrograms. Because of the nature

of the speech code, the pattern for a particular phoneme looks quite different in different contexts. Moreover, there are, in general, no acoustic segments corresponding to the phoneme segments. As a consequence, looking at a spectrogram does not readily reveal how a stretch of speech might be divided into segments corresponding to phonemes, or even how many phonemes it might contain. The difficulty here is that the spectrogram presents speech in its undecoded form. The audible signal has been made visible, but it has not been decoded. As a consequence, it is not highly readable.

There is yet another difficulty that is, in some sense, independent of the fact that speech is the kind of code we have so far described. I refer to the ironic circumstance that some of the most important acoustic cues are among the least prominent aspects of the acoustic landscape. The formant transitions, for example, which are surely among the most important cues for the consonants, are often hard to see; in particular, one is often hard put to decide where the transition begins, though it is precisely this that one needs to know if he is to discover the identity of the phone it signals. It is as if one were trying to read printed alphabetic shapes in which some of the identifying characteristics presented very little contrast to a background containing a great many wholly irrelevant lines.

One is tempted to blame part of the difficulty of reading spectrograms on inadequacies of the sensory transform or on lack of training. And, indeed, it is reasonable to suppose that various kinds of changes in the spectrogram would make the patterns more nearly readable and hence more useful as a way of providing information for the deaf. Surely the spectrogram would be better

for this purpose if the features that carry the important linguistic information were more prominently displayed. We should expect, for example, that cartoonized spectrograms of the kind we use in speech synthesis would probably be easier to read than the real and natural kind. But no such emphasis on the linguistically important features, and no change in the transform as such, would by itself suffice to make spectrograms readable, since these improvements would not obviate the need to decode. The pattern for the same phoneme would still be different in different contexts, and it would still be difficult to see how many phoneme segments there are.

What about training? Can people learn to read spectrograms? What do we know and what ought we to expect? We have suggested that a major difficulty in reading spectrograms is that one must decode the encoded signal. When one <u>listens</u> to speech, this decoding is carried out by a special mechanism, and the process occurs below the level of conscious awareness. Thus, the very different acoustic cues for a particular phoneme in different contexts sound the same in immediate and raw perception. One hears /d/ in the syllables /di/ and /du/ even though the acoustic cue responsible for /d/ is very different in the two cases. Moreover one hears this identity immediately. No reflection or conscious calculation is required. So far as the listener knows, there is, indeed, no code and no problem. But when that same listener looks at a spectrographic representation of /di/ and /du/, he sees different patterns and must then make various conscious and deliberate calculations in order to discover what the message is. Now surely experience will produce some improvement in the speed and accuracy with which he can make those calculations. But will

training in reading spectrograms ever cause the different patterns to look alike? Those of us who have been studying spectrograms these many years would testify, I think, that training will not produce that happy result. The patterns (for the same phoneme) that looked very different to me when first I saw a spectrogram twenty years ago look just as different today. We believe that no amount of training will cause an appropriate speech decoder to develop for a visual input. The speech decoder is, we suspect, biologically linked to an auditory input and cannot be transferred or redeveloped for any other modality. If that is so, then we are faced with an inherent and permanent limitation on our ability to read spectrograms fluently, and we cannot expect that the undecoded speech signal will be highly intelligible to the eye.

What, then, are the possibilities of providing usable feedback to the deaf by presenting the visual signal in decoded form? One thinks immediately of several ways. Conceivably, we might process the speech signal so as to present to the reader, not a spectrogram or any other straightforward transform of the acoustic signal, but rather a phonemic or phonetic transcription. In practice, however, this is merely to trade one difficulty for another, for we should have first to build a speech recognizer, and that, as we know, is a task that has for twenty years frustrated the best efforts of some of the world's best engineers. We should say in passing that it is difficult to build a speech recognizer for the same reasons that it is difficult to read spectrograms: a speech recognizer needs a decoder just as badly as a reader of spectrograms does.

Even if we had such a speech recognizer, it would not do for the deaf all that we want done. It might be of some use to deaf adults who had learned to read, since they could then perceive

the speech of other people. But the deaf child who is trying to learn to speak would, we should think, profit little from a device that merely prints out the name of the nearest phoneme without providing him with any more detailed information about the nature of his articulatory error. Besides, reading is so parasitic on the spoken language that it is, apparently, rather difficult to teach a deaf child to read before he has learned to speak.

How else might we present the speech signal in its decoded form? That is, how else might we present a visual pattern that would bear a simple relation to the linguistic message and also offer some hope of aiding the deaf? To answer that question it is relevant to ask where the encoding occurs, for if we can tap in ahead of that point, we may have a very useful signal. Let us assume, as I think we must, that somewhere in the speaker's central nervous system there exist signals that stand in a one-to-one relation to the phonemes of the language. As one speaks, these signals flow outward from the central nervous system and eventuate as commands to the articulatory muscles. We believe that these commands bear still a fairly simple relation to the sub-phonemic structure of the language, that is, that we can find a fairly close correspondence to the subphonemes (and hence the phonemes) by inquiring which muscles are commanded to contract and when. 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. It is in these latter steps, we think, that most of the complex encoding occurs. Given that the unit commands overlap in time, as they surely do, and given the complex anatomy and physiology of the vocal tract, we should expect to get the kind of encoding that we do, in fact, find -- that is, a loss

of segmentability and the frequent existence of a complex relation between acoustic signal and intended phoneme. If this is so, then the complexities we find at the acoustic level were not always there, and we might expect to 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.

But is this so? Fortunately, the question can be answered by appropriate research -- specifically, for example, by measuring the muscle potential that accompanies contraction. Such measurements are being made for articulatory gestures, in several laboratories, but it is too early to try to draw firm conclusions. It appears, however, that the electromyographic record <u>does</u> bear a simpler relation to the phoneme than does the sound.

We are led to wonder, then, to what extent the deaf might be helped if they were given information, by electromyography and by other means, about the articulatory gestures that they and others make. Such information would, as we have said, bear a simpler relation to the language than do the sounds of speech. It does not follow, of course, that the information will therefore be easily assimilated and used, but it is surely possible that it might be.

I should recapitulate. The speech signal is a complex code on the phonemic structure of the language and needs to be decoded if it is to be correctly perceived. When the speech signal comes in through the ear, it finds a readily available processor that decodes it and recovers the string of phonemes it represents. There is no such decoder available to the eye. As a consequence, visual representations of speech are hard to read, no matter how good the transform or how long the training. One might want

therefore, to give special consideration to the possibility of presenting speech information to the eye in decoded form. Since much of the encoding -- perhaps most of it -- occurs at the conversion of muscle contraction to vocal tract shape, information about articulatory muscle contraction might hold some promise as a useful way of presenting the speech signal to the deaf.

Having said all this, I should emphasize that nothing in this talk was meant to imply that spectrographic or other direct visual representations of the speech signal are worthless to the deaf. My purpose was only to try to explain why such representations are hard to interpret. I would hope that a better understanding of the source of the difficulty would help us to improve these displays and, also, to develop very different ways of providing the deaf with truly useful information about the speech signal.

† This paper is substantially as it was when read to the Conference on Speech Analyzing Aids for the Deaf at Gallaudet College in June, 1967. More detailed treatment of the points made here, together with references to the relevant literature and acknowledgments of support for the research, are to be found in: Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. Perception of the Speech Code, Psychological Review (to appear in November, 1967).

* Also, the University of Connecticut.

** Also, the University of Pennsylvania.