Editor's Note: *This is the sixty-fifth in a series of review and tutorial papers on various aspects of acoustics.*

Acoustics in human communication: Evolving ideas about the nature of speech

Franklin S. Cooper

Haskins Laboratories, New Haven, Connecticut 06510
(Received 8 January 1980; accepted for publication 19 March 1980)

This paper discusses changes in attitude toward the nature of speech during the past half century. After reviewing early views on the subject, it considers the role of speech spectrograms, speech articulation, speech perception, messages and computers, and the nature of fluent speech.

PACS numbers: 43.10.Ln, 43.70. – h

When one thinks about the role of acoustics in human communication, the first thing that comes to mind is speech—and speech is all that this talk will consider. There are other roles, notably in music, architecture, and applications of speech acoustics to communications technology. These will be dealt with in a following talk by Manfred Schroeder.

Even speech is a large subject. In tomorrow's session on Fifty Years of Research in Speech Communication, the seven speakers with three hours at their disposal may be able to summarize the subject in a fairly adequate way. My assignment, for these twenty minutes, must be something else. What I hope to do is not so much to present new information as to explain for those of you who are *not* in speech research why those of us who are in the field find it important and interesting, even exciting.

Why important? Primarily because people like to talk, and because speech is an efficient way to communicate. The latter claim might seem surprising, since we all take speech for granted, and so tend to belittle it. An experimental test, reported by Chapanis[1] only a few years ago, showed how the time required for a cooperative task depended on the mode of communication that was used, or the combination of modes. His finding was that those modes that included speech were at least twice as fast as any of those without speech. Some ways of interpreting the same data gave even higher ratios, up to ten to one in favor of speech.

Many things about speech are interesting but, since we must choose, let us concentrate on *ideas* about speech: how concepts about the inner nature of this odd phenomenon have changed over the past fifty years. First, though, try to recapture the feel of that long ago time. This Society was born near the end of a decade of explosive growth in the technology of communication, due largely to the vacuum tube amplifier. Remember vacuum tubes? Things we now take for granted were brand new in the twenties: radio broadcasting, talking movies, radiotelephony across the Atlantic, the rebirth of the phonograph, even experimental—very experimental—television. Truly, the ability to amplify signals, even those as weak as speech, had consequences. Concepts have consequences, too, as we shall see.

I. EARLY VIEWS OF SPEECH

A view that was commonly held in the 1920's was that speech is a kind of "acoustic stuff," annoyingly complex as to detail but essentially homogeneous on average. We might refer to this, by lunar analogy, as the "green cheese theory." Dr. Crandall[2] of the Bell Telephone Laboratories wrote, in 1917, about speech

> "as a continuous flow of distributed energy, analogous to total radiation from an optical source. This idea of speech is a convenient approximation, useful in the study of speech reproduction by mechanical means."

And it was useful. Basic problems for the telephone engineer were to find out how much of the acoustic stuff the telephone must provide in order to satisfy the listener. What range of frequencies would just suffice? What range of intensities? What signal-to-noise ratios? and so on.

Yet this way of looking at speech had flaws. Even Crandall found it "interesting" that the vowel sounds have most of the energy, whereas the consonants carry most of the information.

However, ideas about speech evolved as new tools become available. By the late twenties, a new high-frequency oscillograph led Dr. Fletcher[3] to devote some twenty pages of his book on speech and hearing to the *waveforms* of various words. By the early thirties, emphasis had shifted from waveform to spectrum, with some attention to the internal structure of the sound stuff. For example, in a 1930 paper, Collard[4] put it this way:

> "a speech sound consists of a large number of components... and it is by noting at what frequencies these prominent components occur that the brain is able to distinguish one sound from another... they are called the characteristic bands... some sounds have only one chacteristic band while others have as many as five."

A different way of thinking about speech was pro-

 0001-4966/80/070018-04$00.80 © 1980 Acoustical Society of America

posed by Homer Dudley[5] at the end of the thirties. In a classic paper on "The Carrier Nature of Speech" he explained speech to his engineering colleagues by drawing an analogy with radio waves, which are not themselves the message, but only its carrier. So with speech: the message is the subaudible articulatory *gestures* that are made by the speaker; the sound stuff is only an acoustic carrier modulated by those gestures.

These ideas, novel except to phoneticians, were embodied in a communications device called the vocoder, which first analyzed the incoming speech and then recreated it at the distant terminal. The vocoder was modeled after human speech, with either a buzz (like the voice) or a hissy sound as the acoustic carrier. The gestures that comprised the message were represented by a dynamically changing spectrum—a necessary engineering compromise, but one which tended to obscure Dudley's main point about the gestural nature of speech and to reemphasize the acoustic spectrum.

II. SPECTROGRAMS AND THEIR CONSEQUENCES

Indeed, the view of speech as a dynamically changing spectrum had been growing in popularity even before the vocoder was invented. Although sound spectrograms were not to have their full effect on speech research until the latter half of the forties, Steinberg[6] had published in 1934 what is, in retrospect, the first spectrogram. It showed how energy is distributed in frequency and time for the sentence, "Joe took father's shoebench out." Since this one spectrogram required several hundred hours of hand measurement and computation, we can understand why this way of representing speech remained a curiosity for so long. In fact, it was not until 1946 that sound spectrograms—and a machine that could make them in minutes—emerged from the war-time research of Bell Laboratories. But spectrograms had a profound effect on speech research. They provided, literally, a new way to look at speech, as well as new ways to think about it. One way, of course, was the familiar description in spectral terms, but with a new richness of detail.[7] Now one could hope to be precise about those "characteristic bands" that distinguish the consonants.

A second way of thinking about speech was to view the spectrogram as a road map to the articulation: thus, the formant bars on the spectrogram told one how the vocal cavities had changed in size and shape. Gunnar Fant[8] and Stevens and House[9] did much to clarify and quantify these relationships for us.

A third way of thinking about speech was to view spectrograms simply as *patterns*. The richness of detail is now just a nuisance, since it obscures the underlying, simpler pattern.[10]

The dynamic character of speech, so evident in spectrograms, led to the development of a research instrument called the Pattern Playback. With it, spectrograms could be turned back into sound in much the way that a player piano turns a perforated musical score back into music. My colleagues Pierre Delattre and Alvin Liberman used the Playback in a long and fruitful search for what they called the "acoustic cues" for speech perception, that is, a search for those crucial parts of the spectral pattern that told the ear what sounds had been spoken.[11]

III. ARTICULATORY NATURE OF THE ACOUSTIC CUES

One might have expected the search for the cues to have a simple, happy ending: namely, the finding of one-to-one correspondences between unvarying parts of the pattern and the minimal units of the speech. I should note that linguists were not in agreement about how to characterize the minimal units of speech—whether as phonemes, (that is, short successive segments) or as distinctive features (that is, as co-occurring attributes of longer duration). The early work sought to relate acoustic cues to phonemes. Its outcome raised issues that are unresolved to this day, and set research on two seemingly divergent paths. Let us follow one of them to the present, then return to the 1950's to pick up the other.

The acoustic cues, when they were found, proved to be neither unvarying nor simple. For a given phoneme, the cues would change, sometimes markedly, whenever the neighboring phonemes were changed. Further, the ways in which a phoneme changed were not readily rationalized in acoustic terms, though they made good sense in terms of the articulatory gestures. These findings, and much else, led to a production-oriented view of the nature of speech. The main points were, first, that the speaker's underlying phonemic message emerges as a complexly *encoded* sound stream because of the several conversions it must undergo in the process of being articulated. This being so, perception of the speech by a listener necessarily involves a decoding operation, and probably a special speech device for that purpose—a built-in option available only on the *homo sapiens*. But what kind of mechanism, or special decoder, might that be? One possibility is a neural linkage between the auditory analyzer of the incoming speech and the motor controller of articulation, and thence upstream to the message in linguistic form—in short, perception achieved by reference to production.

This view of speech as an encoding operation[12] has had consequences for both experimental and conceptual aspects of speech research. On the experimental side, it motivated studies of how the articulators move when one is talking, of what the muscles do to make them move and, of course, how the sounds change with articulation.

On the conceptual side, interest has focused on how gestures relate back to linguistic units. The simplest relations—for example, correspondence between a particular phoneme and the contraction of a particular muscle, or between a phoneme and a target shape of the vocal tract—these simplest relations were found to be too simple to account for the data. They share that fate in varying degree with other, less simple, relationships proposed as alternatives. Indeed, the na-

ture of the relationship is a central question in speech research today: How is the motor control of speech organized? How do linguistic units give shape to gestures?

IV. PERCEPTION BY AUDITORY ANALYSIS

We must now go back to the 1950's, having traced the view that speech is articulatory in its very nature. There is an alternate view that stresses the role of the listener. It asks: Are there not, in the acoustic signal and its spectrogram, objective entities that correspond to the speech sounds that one hears so clearly? If this does not hold for the relationship of cues to phonemes, might it hold for distinctive features instead? The answer, despite persistent effort, has turned out to be that it is no easier to find invariant relations between features and acoustic spectrum than it is between phonemes and spectrum. I should say here that there are respected colleagues who do not share this assessment and who feel that, since the ear *must* do an initial analysis of speech, it is more than reasonable to suppose that the auditory system carries that analysis all the way to the linguistic units. The question, in their view, is not *whether* the ear does that analysis, but only *how* it does it.

One line of thinking has been that invariant relationships might be found in the signal after it has been transformed by the ear. A related approach has combined articulatory and auditory considerations by looking for quantal states for which variability of the gesture has only trivial effect on the sound. These stable sounds can then serve as auditory cues.[13] The principal mechanisms proposed for interpreting auditory cues is a set of property detectors turned to quite a variety of acoustic aspects of the signal. Most recently, interest has focused on the neural codings and recordings which the speech signal undergoes on its way up the auditory pathways.[14] This is exciting research, and there are those who hope that tracking speech to its engram in the cortex will clarify the relationship between acoustic units and linguistic units. This is surely the central point, if we are to understand speech perception.

V. THE NEED FOR A MODEL

So, in tracing ideas about the nature of speech from the 1950's through the 1970's, we have found unresolved questions about the choice of minimal units, and also about whether speech "belongs," in some important sense, to the mouth or to the ear. In one sense, of course, it belongs equally to each of them since the same waveform is both output and input. For a speaker, it is *both* at the same time. But what are the mechanisms?

We are concerned at two levels. We need, of course, to learn about the physiological mechanisms of production and perception, and we are making good progress. We need also to understand these processes at an underlying functional level—at the level of meaningful models. Do we need separate models for production and perception? It could be that each process has its own way of relating linguistic message and speech signal. In that case, we do have two models and we explain the ambivalent nature of the acoustic cues as the compromise made long ago by mouth and ear in arriving at a set of signals for spoken language. But parsimony, and a substantial body of data, argue persuasively for one model instead of two, that is to say, for a close functional linkage between production and perception that will explain how *both* relate to the message that speech conveys.

VI. MESSAGES AND COMPUTERS

What is the message? What is the nature of the information contained in speech? This question is by no means new, but for some of us in speech research, it has acquired a new meaning as our field has begun to reach out from "laboratory speech," i.e., nonsense syllables, words, and the simplest of sentences, to everyday fluent speech. The question has emerged in sharpest form in work on speech understanding by computers.[15]

This would justify a digression, if only time permitted, about human communication with machines. We are, I believe, on the leading edge of a new wave of technology as computers learn how to use spoken language. Never mind that the technology is still complex, expensive, and severely limited in what it can do. Remember the telephone: It was invented forty years too late, when telegraphy was already in use around the world. But how many of you have telegraphs in your homes and offices today?

A proper account would have to trace the early, and generally successful, efforts to synthesize speech automatically, and the parallel efforts, largely frustrated, on machine recognition; also, the related work on analysis–synthesis telephony and on waveform coding for better and cheaper communications between humans by the use of machines.

In the area of *ideas* about speech, two things emerge from the research on how to converse with computers. The first is that fluent speech is the real problem, and a different and harder one than dealing with careful, word-by-word "laboratory speech." Indeed, the direct phonetic analysis of fluent speech may not even be possible. The second thing is that computers *can* understand such messages, under suitably constrained conditions, when the partial information available from phonetic analysis is supplemented by syntactic, semantic, and pragmatic knowledge.

VII. THE NATURE OF FLUENT SPEECH

What is so different about fluent speech? Essentially, it is that the strategy is different, and that the acoustic signal is both less than, and more than, it was for laboratory speech: *less*, through depletion of phonetic detail: *more*, by accretion of acoustic cues direct to the syntax, the semantics, and the pragmatics. This can hardly be called an impoverished signal, but it is a different kind of signal, requiring new decoding techniques and new research on nonphonetic types of cues.

A case in point is the current interest in prosody, where some of these cues are to be found.

This view of fluent speech gives new status to the acoustic signal, for now it has become the carrier, not merely of phonetic elements, but of language entities at all levels. This seems a heavy load—all of language—for so frail a carrier. Perhaps we should re-examine some long-held assumptions about the nature of spoken messages: namely, that the speaker's message—all of it—is carefully packaged for aerial transport, then carried through the air to the ear and brain of a listener, where it is unpacked. But must all of it go through the air, or only those parts that are not already present in both heads, as information theory might suggest? To put it a little differently, speech could still perform its function if it carried no more than the *recipe* for making a message, just as a dandelion seed carries only the genetic code from one flowering to the next.

VIII. SUMMARY

We have seen that the principal role of acoustics in human communication is to let us talk with each other and, eventually, with our computers. For face-to-face communication, acoustics is quite adequate, but there is a large and growing technology in which acoustics, per se, plays a rather minor role. Even there, however, the way we conceptualize speech has important consequences.

We have come a long way in understanding the nature of spoken language. We have left behind the idea of speech as merely sound stuff, or even spectrum stuff that has certain characteristic frequencies. Spectrograms revealed the dynamic character of speech and hinted strongly at a dual nature: that speech is both a signal that is shaped by production and one that is tailored for perception. We have come to see speech as a carrier, at first mainly of phonetic messages, then as a carrier of cues to all levels of languages. We do not yet know how much, or how little, of the total message must actually be transported in fluent speech.

Finally, there are abiding questions that motivate much of the current research on speech:

(1) What are the units?—A persistent question at all levels of language, though we mentioned only phonemes and distinctive features as *minimal* units.

(2) What is the mechanism? Is speech truly anchored to production? or to auditory perception? Or, will this turn out to be a nonquestion, when we have finally arrived at a model that relates production and perception to each other, and to the message?

(3) What is the message? How is it carried by speech? Indeed, is the total message carried at all, from head to head? or is it created anew in the head of the listener from a recipe provided by the speech signal?

I hope you now see why some of us find research on speech so challenging, and I ask your indulgence for my biases, some of which must be showing.

[1]A. Chapanis, "Interactive Human Communication," Sci. Am. 232, 36–42 (1975).

[2]I. B. Crandall, "The Composition of Speech," Phys. Rev. 10, 74–76 (1917).

[3]H. Fletcher, *Speech and Hearing* (Van Nostrand, New York, 1929).

[4]J. Collard, "Calculation of the Articulation of a Telephone Circuit from the Circuit Constants," Electron. Commun. 8, 141–163 (1930).

[5]H. Dudley, "The Carrier Nature of Speech," Bell System Tech. J. 19, 495–515 (1940).

[6]J. C. Steinberg, "Application of Sound Measuring Instruments to the Study of Phonetic Sounds," J. Acoust. Soc. Am. 6, 16–24 (1934).

[7]R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech* (Van Nostrand, New York, 1947).

[8]C. G. M. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960).

[9]K. N. Stevens and A. S. House, "The Development of a Quantitative Description of Vowel Articulation," J. Acoust. Soc. Am. 27, 484–493 (1955); "Studies of Formant Transitions Using a Vocal Tract Analog," J. Acoust. Soc. Am. 28, 578–585 (1956).

[10]F. S. Cooper, A. M. Liberman, and J. M. Borst, "The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech," Proc. Natl. Acad. Sci. 37, 318–328 (1951).

[11]An informal account of these early experiments and their interpretation is given by A. M. Liberman and F. S. Cooper, "In Search of the Acoustic Cues," in *Mélange à la Mémoire de Pierre Delattre*, edited by A. Valdman (Mouton, The Hague, 1972), pp. 9–26.

[12]For a review of the experimental data and a mid-course interpretation, see A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," Psych. Rev. 74, 431–461 (1967); for an updated interpretation, A. M. Liberman and M. Studdert-Kennedy, "Phonetic Perception," in *Handbook of Sensory Physiology*, Vol. VIII: Perception, edited by R. Held, H. Leibowitz, and H-L. Teuber (Springer Verlag, Heidelberg, 1978), pp. 143–178.

[13]K. N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data," in *Human Communication, a Unified View*, edited by P. B. Denes and E. E. David (McGraw-Hill, New York, 1972), pp. 51–66; S. E. Blumstein and K. N. Stevens, "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants," J. Acoust. Soc. Am. 66, 1001–1017 (1979); K. N. Stevens and S. E. Blumstein, "The Search for Invariant Acoustic Correlates of Phonetic Features," in *Perspectives on the Study of Speech*, edited by P. D. Eimas and J. Miller (Erlbaum Assoc., New Jersey) (in press).

[14]These ideas are discussed at this 50th Anniversary Meeting in the invited papers of Session NN by N. Y-S. Kiang, K. N. Stevens, B. Delgutte, and M. B. Sachs and E. D. Young, J. Acoust. Soc. Am. Suppl. 1 65, S (1979).

[15]D. H. Klatt, "Review of the ARPA Speech Understanding Project," J. Acoust. Soc. Am. 62, 1345–1366 (1977).