Some Instrumental Aids to Research on Speech

FRANKLIN S. COOPER, *Haskins Laboratories*

Speech research shares with other areas of scientific investigation a history of dependence on the instrumental aids available to it. Very often, instruments and techniques have set the directions, and the limits, of investigation. One could trace this activating role of the tools of research throughout the history of studies on speech, noting, for example, the part played by kymographic techniques in the development of experimental phonetics, the use of vacuum tube amplifiers in telephony, and the application of more elaborate electronic techniques in present-day efforts to develop speech typewriters and other bandwidth compression devices which depend on a knowledge of the "invariants" of speech. The aim of this paper is, however, more specific and less ambitious: a description of two instruments in use at Haskins Laboratories and a brief account of the kinds of research which they have made possible.

One of these is a sound spectrograph, a device now so well known that it is mentioned only because of its role in the research and because of the special design characteristics

FIGURE 1. Diagram of the pattern playback. The tone wheel has 50 circular sound tracks which yield pure tones at 120 cps intervals from 120 to 6000 cps when the tone wheel is rotated at 1800 rpm by a synchronous motor. Light (from the lamp at the extreme left) is modulated by the tone wheel and directed onto the spectrogram in such a way that the spectrographic pattern will transmit (or reflect) to a photocell just those portions of the light which carry the frequencies corresponding to the pattern. (Reproduced by courtesy of the American Journal of Psychology.)

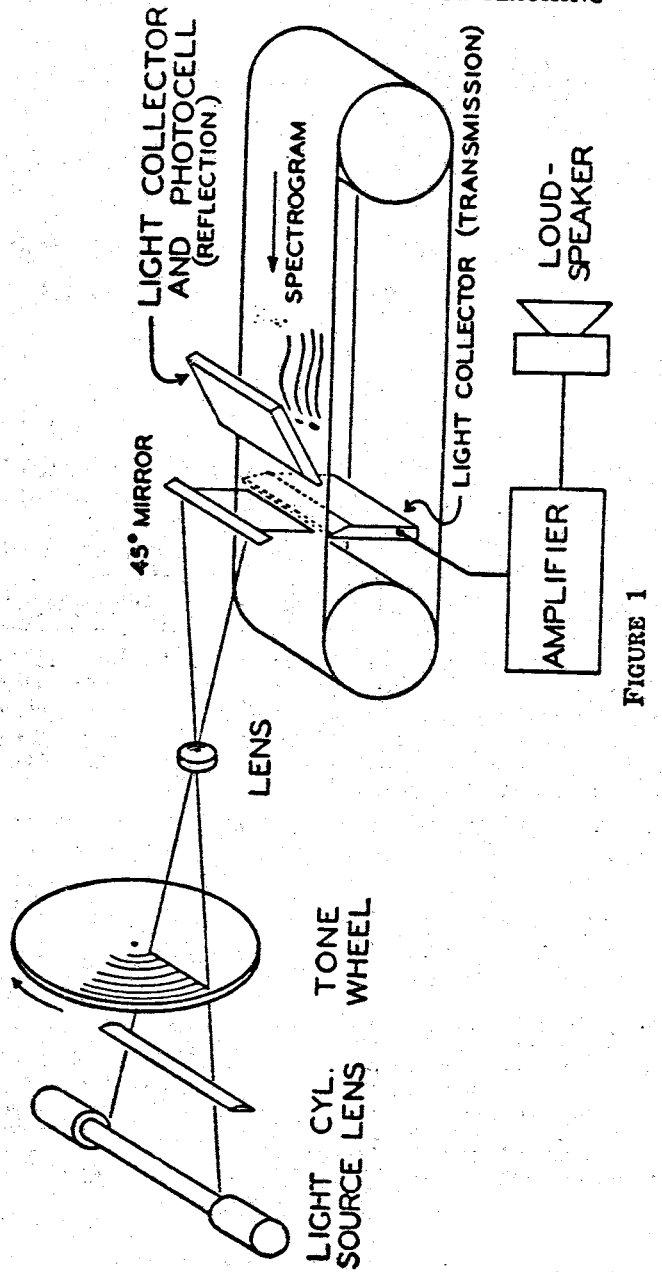

FIGURE 1

which make it suitable for use with the pattern playback. In operating principle, this special-purpose spectrograph is based on the recording spectrograph developed by the Bell Telephone Laboratories and currently manufactured by the Kay Electric Company. It differs functionally in two important respects: (1) the spectrogram is recorded as a photographic transparency in order that it may be used as an optical modulator in controlling the performance of the pattern playback; and (2) the spectrogram has a long dynamic range in order that it may reproduce both the weak and the strong components of speech in their correct relative intensities. The photographic density of the film varies directly with the intensity level of the sound over a range of approximately 40 decibels. Also, for convenience in experimentation, the spectrograms are fairly large (7 inches wide by 7 feet long) and each spectrogram accommodates a comparatively long speech sample (approximately 12 seconds).

The spectrograms provide visual patterns which correspond to the auditory patterns of speech. Because these visual patterns are stationary in time and because they contain suggestions about the particular acoustic events which may serve as cues for the perception of the several speech sounds, it is convenient to deal with speech as if it were primarily a visual display. One can, then, inquire about the importance of a particular portion of the spectrographic pattern, or how a change or omission of some aspects of the pattern would affect the sound as heard.

Such questions can be answered experimentally by reconverting the modified spectrogram into sound with the aid of an instrument called a pattern playback. The essence of the experimental method described here is, in brief, the manipulation of spectrograms as visual patterns, combined with evaluation by ear of the resulting changes in the audible pattern.

The playback, shown schematically in Fig. 1, scans a spectrogram from left to right along the time axis, using for this purpose a line of light modulated by a tone wheel at some 50 different frequencies which match approximately the frequency scale of the spectrogram. Certain portions of the modulated light are selected by the spectrographic pattern,

FIGURE 2. The pattern playback. The light collector (F) is positioned for use with reflection spectrograms. The component parts of the playback are identified in the text; they are arranged about as in Fig. 1.

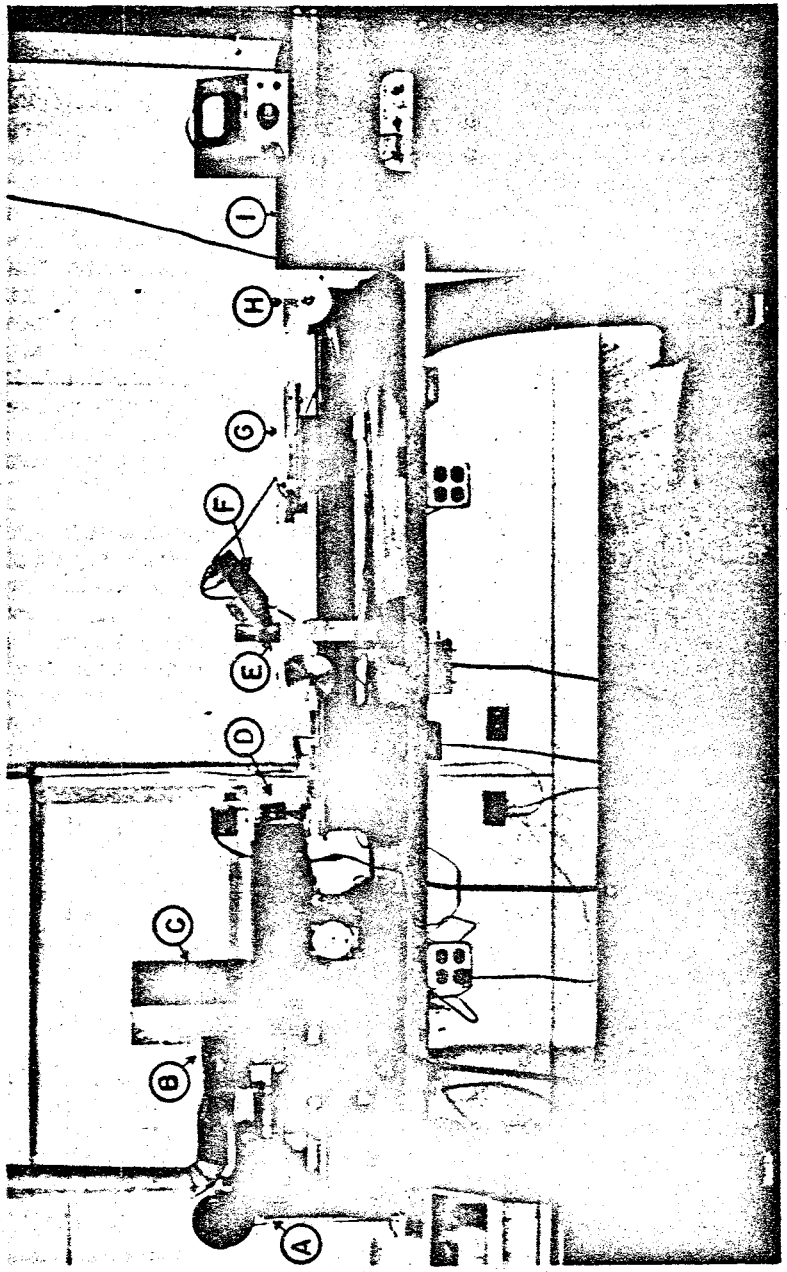

FIGURE 2

either by transmission through the photographic film of an actual spectrogram or by reflection from a painted version. In either case, the light is led by an optical system to a single phototube. Thus the photocurrent, amplified and fed into a loud speaker, produces sounds which have, at every instant, approximately the frequency components shown on the spectrogram. The instrument, shown in Fig. 2, has its component parts arranged about as in the diagram of Fig. 1. The tone wheel, a sheet of photographic film about 20 inches in diameter on which are recorded 50 sound tracks of the harmonics of 120 cycles, is located inside a square housing (C) and is driven at 1800 rpm by a synchronous motor (A) through a mechanical filter. The light source is a high-intensity mercury arc in a ventilated housing (B) immediately behind the tone wheel. The lens (D) and 45° mirror (E) direct the modulated light onto the surface of a cellulose acetate tape on which the spectrographic patterns are painted in white. The light collector and photocell (F) convert the light reflected from the pattern into a pulsating electric current which is led, after amplification (I), to a loudspeaker or headphones. The spectrogram is driven past the scanning point (E) by two rotating drums (one at H). A drawing table (G) permits convenient changes to be made in the patterns. Power controls and electrical filters for the arc lamp are not shown in the figure.

The method and instrument described above provide a very convenient basis for experimenting with the perception of speech—that is, for making a great variety of changes in the acoustic stimuli and then determining the effects of these changes on the sound that is heard. Early experimentation was directed to the question of the intelligibility of hand-painted simplifications based on spectrograms of spoken sentences. It was found that two, or at most three, vowel formants and somewhat stylized representations of the consonantal features could yield patterns which appear much simpler to the eye than the corresponding spectrograms of actual speech

FIGURE 3. Two versions of a sentence employing principally stop and resonant consonants. The lower version is a first draft which was painted directly from the typewritten text in accordance with the rules derived from our experiments. Revisions by ear (including the use of some third-formant transitions) resulted in the upper version. Both were highly intelligible when converted into sound by the playback. (Reproduced by courtesy of the Journal of the Acoustical Society of America.)

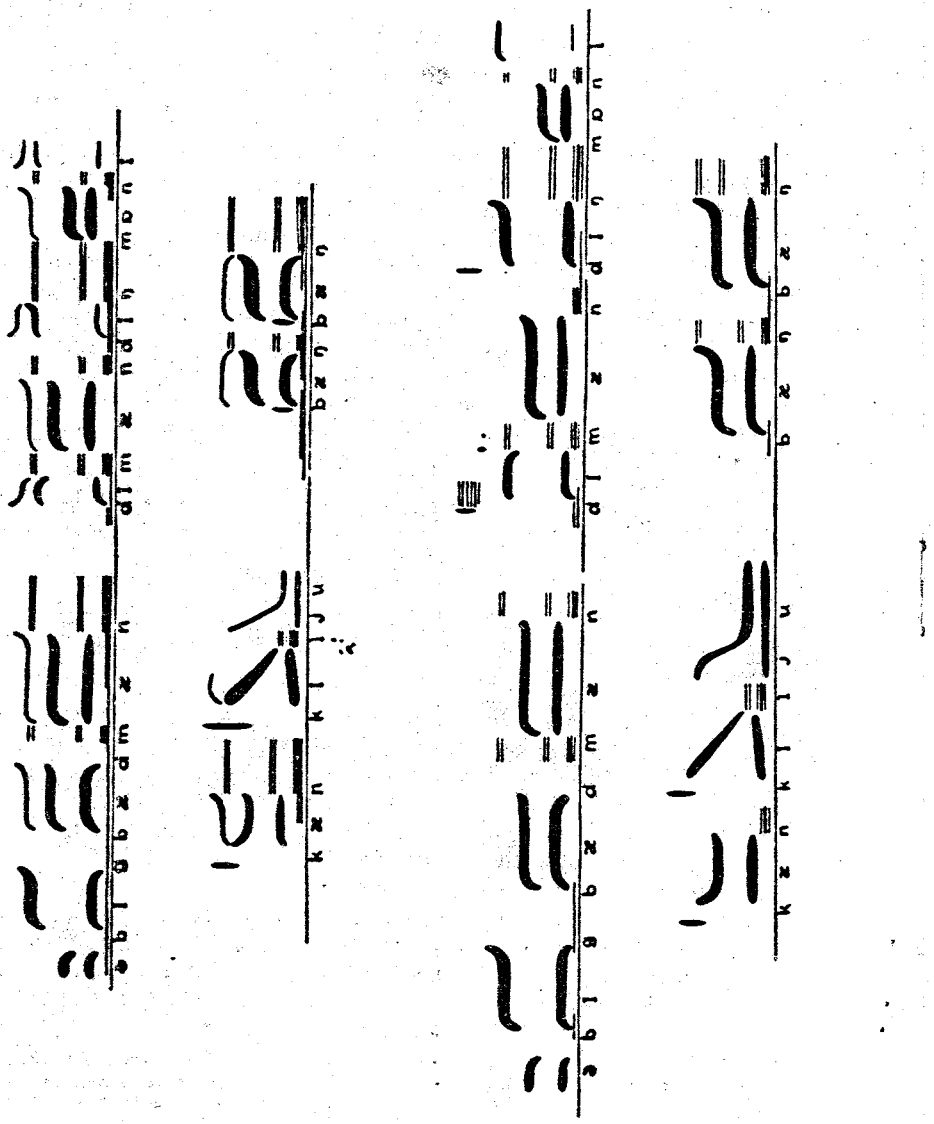

FIGURE 3

and which, as heard with the playback, were as intelligible as the complete spectrographic pattern.[1]

Further simplification of connected text was not attempted; rather, the course of the experimentation was directed to a study of isolated acoustic cues for single phones of American English. One study on the stop consonants dealt with bursts of noise as the acoustic cues. Consonant-vowel syllables comprising all possible combinations of 12 brief bursts of noise (centered at different frequencies) with seven two-formant steady-state vowels were presented to naive subjects who selected from the 84 combinations those which most clearly began with *p*, *t*, or *k*. The agreement among the listeners in identifying these sounds was impressive. There was a very clear indication that bursts and formant portions of the pattern were not judged separately as consonant and vowel but rather in combination, suggesting that the acoustic unit for perceptual purposes is of syllabic dimensions.

Another study, also on the stop consonants, dealt with the role of initial transitions (frequency variations) of the vowel formants as cues for the voiced stops. It is evident from this study that these transitions are not merely trivial consequences of articulatory movements but rather are very important determinants of the perceived phones; also, that there is, for transitions as for bursts, a substantial overlap of the acoustic cues for consonant and vowel. The existence of interdependent and overlapping acoustic cues suggests that the acoustic stream represents an encoding, as contrasted with an encipherment (in the cryptographic sense), of the message, assuming the message to be represented most simply by a succession of phonemes. If this view is correct, the search for the "invariants" of speech should be conducted with some caution.

One might well ask whether the sounds produced by the pattern playback are indeed speech sounds, or whether they might be so artificial and so little related to actual speech that the results of studies like the above might have no relevance for spoken communications. This seems unlikely since the painted patterns do resemble rather closely the spectrographic patterns of actual speech, and since naive listeners

[1] See P. Delattre, F. S. Cooper, and A. Lieberman, "Some Suggestions for Teaching Methods Arising from Research on the Acoustic Analysis and Synthesis of Speech," *Third Annual Round Table Meeting*, 1952, 31-45, and especially Figure 1.

agree rather well on the identifications of the synthetic patterns. In addition, a direct test by Carol Schatz has shown that stop consonants reassembled from recordings of human speech show the same behaviour as the synthetic combinations of bursts of noise and two-formant vowels.

The investigation of isolated cues is by no means complete. However, the cues for enough sounds have been studied to permit an attempt at the total synthesis of connected text which is limited to a somewhat restricted range of phonemes. An example is shown in Fig. 3, in which the third and fourth lines represent a synthesis *by rule*, that is, without reference to spectrograms of spoken sounds. The acoustic cues used in this synthesis are formant positions, transitions, and frequency positions of bursts of noise. The sentence as first painted was quite intelligible although the tempo and emphasis were somewhat strange for American English. The first and second lines of the figures show a revised version as edited by ear. The inclusion of third formant transitions seemed to improve the intelligibility. Synthesis by rule is not, of course, to be considered an end in itself, but rather a check on the validity of the experimental findings about the acoustic cues for various phones.

In conclusion, the method of studying speech sounds by manipulating spectrographic patterns and reconverting them to sound for judgment by ear has proven to be a useful tool. Its application has required the cooperative efforts of a diversified research team. Of the many areas which remain to be explored, mention was made of the isolation of additional acoustic cues for English; comparable studies of the sounds of other languages; the relative importance of the various cues for speech intelligibility in quiet and in noise; the role of pitch changes (which will, however, require the completion of a new playback); and the possible contributions which studies in acoustic phonetics can make to the art of teaching languages.