Reprinted from the Proceedings of the National Academy of Sciences,
Vol. 37, No. 5, pp. 318–325. May, 1951

THE INTERCONVERSION OF AUDIBLE AND VISIBLE PATTERNS AS A BASIS FOR RESEARCH IN THE PERCEPTION OF SPEECH*

By Franklin S. Cooper, Alvin M. Liberman† and John M. Borst

Haskins Laboratories, New York

Read before the Academy, October 10, 1950

In investigating the acoustic aspects of speech it has long been the practice to convert these extremely complex sounds into a visible display, and so to enlist vision as an aid in dealing with a problem which lies largely in the area of auditory perception. Of the various displays which have been used, perhaps the most effective is provided by the sound spectrograph, which has come to be recognized as a valuable research tool for the study of the acoustic correlates of perceived speech.[1] By examining numerous spectrograms of the same sounds, spoken by many persons and in a variety of contexts, an investigator can arrive at a description of the acoustic features common to all of the samples, and in this way make progress toward defining the so-called invariants of speech, that is, the essential information-bearing sound elements on which the listener's identifications critically depend. The investigator can also take account of the variations among spectrograms, and by correlating these with the observed variations in pronunciation, he can begin to sort out the several acoustic features in relation to the several aspects of the perception.

There are, however, many questions about the relation between acoustic stimulus and auditory perception which cannot be answered merely by an inspection of spectrograms, no matter how numerous and varied these may be. For any given unit characteristic of the auditory perception, such as the simple identification of a phoneme, the spectrogram will very often exhibit several features which are distinctive to the eye, and the information which can be obtained from the spectrogram is, accordingly, ambiguous. Even when only one feature or pattern is strikingly evident, one cannot be certain about its auditory significance, unless he assumes that those aspects of the spectrogram which appear most prominently on visual examination are, in fact, of greatest importance to the ear. That assumption, as we shall try to point out later in this paper, is itself extremely interesting, but it has not been directly tested, nor, indeed, has it always been made fully explicit.

To validate conclusions drawn from visual examination of spectrograms, or, more generally, to determine the stimulus correlates of perceived speech, it will often be necessary to make controlled modifications in the spectrogram, and then to evaluate the effects of those modifications on the sound as heard. For these purposes, we have constructed an instrument, called

a pattern playback, which reconverts spectrograms into sound, either in their original form or after modification.

The basic operating principle[2] is quite simple (Fig. 1). The playback scans a spectrogram from left to right along the time axis, using a line of light modulated by a tone wheel at some fifty harmonically-related frequencies which match approximately the frequency scale of the spectrogram. Those portions of the modulated light which are selected by the spectrogram—by transmission through a film transparency or by diffuse reflection from a painted design—are collected by an optical system and led to a phototube. Thus the photocurrent, amplified and fed into a loudspeaker, produces sounds which have, at every instant, approximately the frequency components shown on the spectrogram.

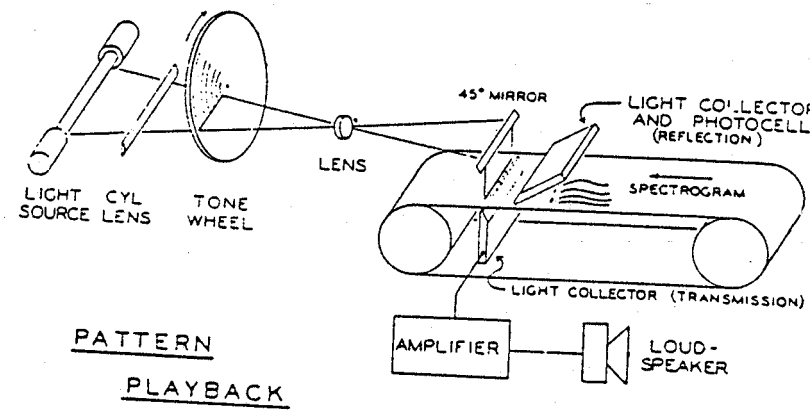

FIGURE 1

Operating principle of the pattern playback

For convenience in research the playback is designed to operate from either of two kinds of spectrograms (Fig. 2). In the one case, the spectrogram is a film transparency, and the sound is produced by the light which is transmitted through the relatively transparent portions of the film. These transmission spectrograms, so called, are photographic copies (on film) of an original produced from recorded sound by a spectrograph designed specifically for this purpose. The transmission spectrograms are most useful if one wishes to recreate the original sound as accurately as possible, or to make minor changes (especially deletions) in some detail of the spectrogram. In the other case, the playback operates from spectrograms which are drawn with white paint on a transparent plastic base, and the playback uses only the light which is reflected from the painted portions. The drawing is done by brush or pen, and the spectrograms can be prepared or modified in a variety of ways. These spectrograms are most

appropriate, then, if one wishes to make drastic changes in the sound, or, in the extreme case, to employ entirely synthetic patterns.

In general, the playback appears to be a most useful tool in research involving the experimental manipulation of speech sounds. By comparison with more conventional instruments for modifying the speech stream, the playback method is extremely flexible and convenient, and has the particular advantage that it allows considerable freedom in dealing with the dynamic or constantly varying aspects of speech. This was, indeed, the

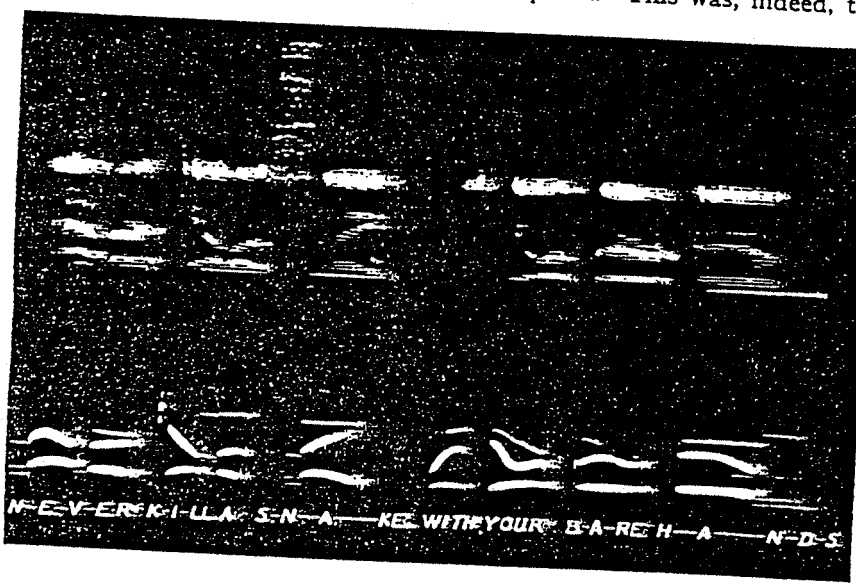

FIGURE 2

(*a*) Transmission spectrogram copied photographically from an original spectrogram without modification.

(*b*) Reflection spectrogram drawn by hand as a simplified version of the original spectrogram.

(*c*) Text of the sentence.

specific function for which it was designed. It is of course obvious that such a playback will be useful as a research tool only to the extent that it is able to produce intelligible speech which may then be degraded or dissected in various ways.

We have measured the intelligibility of the playback speech produced from transmission spectrograms, using for this purpose twenty standard test sentences, and have found it to be approximately 95%. The intelligibility of reflection spectrograms will obviously depend on how they are drawn. In drawing these spectrograms we have attempted, in a preliminary way, to determine the extent to which they could be simplified without

serious loss of intelligibility. The procedure was, first, to copy from an actual spectrogram the features which were most prominent visually, and then, largely on a trial-and-error basis, to make such further changes as were required for reasonable intelligibility. For the degree of schematization shown in figure 2 (*b*), the median intelligibility is about 85%, as determined with twenty test sentences. The experimental procedures and the nature of the simplifications will be reported in detail elsewhere.[3] For present purposes it is important merely to note that the simplified spectrograms are, in general, a reduction of the originals to their most obvious visual patterns. Although the first rough paintings were modified in many details in order to produce these highly simplified spectrograms, the modifications have not destroyed the over-all visual resemblance to the originals. To the extent that these similarities remain, and also to the extent that the simplified spectrograms are intelligible, these results provide a partial validation of the assumption, referred to earlier, that the spectrogram displays most prominently to the eye those acoustic features which are of greatest importance in auditory perception. If this were not the case, the spectrograph would not be so useful a tool in describing the sounds of speech, and, more significantly for our purposes, the playback would have no special advantage as a means of manipulating speech.

That the playback does have special advantages is indicated by our experience with it, and this fact seems, moreover, to have theoretical implications which deserve examination. It does not appear that the advantage is solely one of stopping time, that is, of converting a transitory sound into a stationary visible display which can be modified and then reconverted into sound for aural evaluation. This is an obvious advantage and an important one in experimenting with speech, but it is neither unique nor quite sufficient. For example, sounds can be represented to the eye by means of an oscillograph, and the oscillogram can be reconverted into sound by a device somewhat like a phonograph, yet the oscillographic representation is virtually useless as a basis for experimenting with the sounds of speech. The critical requirement, and the one which is not adequately met by the oscillogram, is that the display must provide for the eye information which is organized into patterns corresponding to the acoustic patterns on which auditory identifications depend; that is, the conversion must be from patterned acoustic information to patterned visual information. When this is so, the significant aspects of the acoustic pattern become comprehensible to the eye, and the display will have conceptual and also experimental utility in manipulating the sounds of speech. We believe that a reasonable approximation to the required conversion is represented by the spectrograph-playback combination, which interconverts the x and y coordinates of visual space with time and frequency in the acoustic domain (preserving intensity as a parameter in both cases), and that this

accounts for the special utility of these instruments as practical research tools.

On the theoretical side it would seem that the existence of this particular intersensory conversion, with its special advantages, may have interesting implications for the perceptual processes operating in vision and audition. Perhaps the most general implication is that there is an important similarity between visual and auditory perception, sufficient to permit an interconversion at the stimulus level such that patterns in the one sensory modality may, after conversion, be perceived as patterns in the other modality. The term pattern, as used here, implies an organization of stimuli which has the property of retaining its perceptual integrity in spite of gross and diverse changes in the absolute values of the several stimulus components. In vision, for example, a triangle will continue to be perceived as a triangle despite wide changes in size, position, etc. Similarly, in an auditory case, the perception of speech is to a certain extent independent of any fixed and isolated stimulus values, inasmuch as the sounds of speech can be considerably stretched or compressed in time, frequency and intensity without serious damage to intelligibility. These patterns of information, on which visual and auditory perception so largely depend, may then be defined operationally in terms of the particular alterations in stimulus which do and do not impair perceptual identification. By this definition, and according to our assumption about the general nature of the similarity between visual and auditory perception, it should be possible so to convert from vision to audition that varying the visual stimulus will affect the visual and the auditory identifications about equally, i.e., that stimulus changes which do or do not destroy the pattern in the one modality will, correspondingly, destroy it or not in the other. More simply, and only somewhat less precisely, stimuli which look alike can be made to sound alike, and stimuli which look different should then sound different.

It is most unlikely that visual and auditory patterns can be interconverted in complete detail and in all respects, and hence one must expect that the best of audio-visual transforms will be rather less than perfect. For all its inadequacies, however, such a transform will be of practical interest for its bearing on the problems of sensory prosthesis,[4] and for its application, as in the case of the playback, to investigations of the perception of a wide variety of sound patterns, including speech as a special and obviously important case.[5]

It appears, from a theoretical point of view, that the fact of intersensory pattern conversion implies functional similarities between vision and audition at a presumably high level of perception. Moreover, the specific nature of the similarities will be indicated by the precise way in which the best possible conversion is to be made. The transform concept also implies some degree of intersensory generality for certain perceptual laws, since

one may suppose that patterns will be interconvertible between vision and audition only to the extent that common principles determine the way in which visual and auditory stimuli are organized in perception. The further development and extensive testing of the transform concept, according to the criteria implicit in our assumptions, may therefore serve to reveal certain principles which are so basic as to have a special importance for theories about perceptual mechanisms in general.

We have taken a first and very tentative step toward testing the transform concept with materials other than speech sounds. In studying speech, we have been dealing with patterns which were quite obviously more

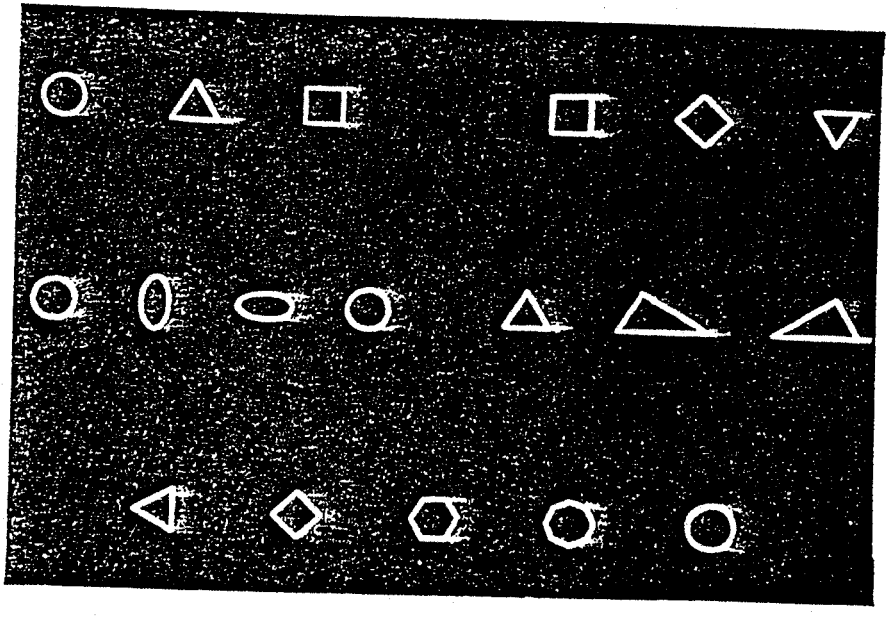

FIGURE 3

Three sequences of geometric figures used in testing the audio-visual transform.

familiar to the ear than to the eye. What will be heard when the playback is used to transform patterns which are familiar to the eye but not to the ear? The tests that we have so far made employed common geometrical forms such as circles, ellipses, triangles and squares (Fig. 3). As represented on the playback tape, each figure was varied in size and position, drawn in outline and in solid form, variously distorted in shape, and rotated about axes perpendicular to the tape. When these figures are converted by the playback, a listener hears sound which he rather readily puts into categories corresponding to the grouping of the figures in visual perception. All of the triangles, for example, sound somewhat alike, regard-

less of variations in size and position on the tape, and these sounds are quite distinctively different from the sounds corresponding to squares or circles. The one apparent exception is that rotation of many of the figures causes very large changes in the sound, and the listener tends not to put rotated forms of a rectangle, for example, into the same category. This may represent a limitation of the transform, though it must be remembered that the identifications of visual figures sometimes change when the figures are rotated, a familiar example being that of square and diamond.

These studies were intended to be exploratory and the results are most incomplete; there is not, of course, any special significance in the choice of geometric figures as test patterns. The experiments are described here because they illustrate how the playback transform can be tested according to our criterion of pattern interconvertibility, and because the results indicate in a general way the extent to which the playback transform, with its present imperfections, meets that test.

In summary, we have found it useful to assume, as a working hypothesis, that the perceptual processes in vision and audition exhibit important similarities in operating upon patterned information. A consequence of this hypothesis, put more explicitly and in operational terms, is that information which is perceived as a pattern when displayed to the eye can be so converted into sound for presentation to the ear that the pattern characteristics will be preserved, that is, that the sounds will be perceived as patterned information in audition. Conversely, patterned acoustic information can, by the reverse transformation, be displayed as visual patterns. The sound spectrograph and pattern playback are appropriate even though imperfect instrumental means of effecting such an interconversion, which in this case is that the x and y dimensions of the visual array are changed to the time and frequency dimensions of sound, with intensity remaining a parameter in both cases. Stated symbolically, the audio-visual transform then becomes $F_v(x, y, I) \rightleftharpoons F_a(f, t, I)$.

This conceptualization of a particular relationship between vision and audition, quite aside from its theoretical interest, provides a basis for understanding the demonstrated utility of the sound spectrograph in displaying the sounds of speech for visual study, and of the pattern playback as a tool for the manipulation and synthesis of speech.

* The research reported here was made possible by funds granted by the Carnegie Corporation of New York. The paper, as read before the Academy on October 10, 1950, employed recordings to illustrate various points in the discussion. Some revision of the text has therefore been necessary.

† Also, University of Connecticut.

[1] Potter, R. K., and Steinberg, J. C., *J. Acous. Soc. Am.*, 22, 803–823 (1950). Joos, Martin, *Acoustic Phonetics* (Language Monograph No. 23), Linguistic Society of America, 1948. Potter, R. K., Kopp, G. A., and Green, Harriet C., *Visible Speech*, D. van Nostrand, 1947.

[2] Potter, R. K., U. S. Patent No. 2,432,123 (1947). Cooper, F. S., Liberman, A. M., and Borst, J. M., *J. Acous. Soc. Am.*, 21, 461 (1949). Vilbig, F., *Ibid.*, 22, 754–761 (1950).

[3] Cooper, F. S., Borst, J. M., and Liberman, A. M., "Pattern Playback and Sound Spectrograph: Tools for Research in the Perception of Speech and Other Complex Sounds" (in preparation).

[4] Zahl, P. A., Ed., *Blindness: Modern Approaches to the Unseen Environment*, Princeton University Press, 1950. Potter, R. K., Kopp, G. A., and Green, Harriet, C., *Visible Speech*, D. van Nostrand, 1947. Cooper, F. S., *Physics Today*, 3, 6–14 (1950).

[5] Delattre, Pierre, *J. Acous. Soc. Am.*, 22, 678 (1950). Cooper, F. S., Liberman, A. M., and Borst, J. M., *Ibid.*, 22, 678 (1950). Cooper, F. S., *Ibid.*, 22, 761–762 (1950).