XIII. SPEECH COMMUNICATION

Prof. M. Halle	Jane B. Arnold	H. Fujisaki
Prof. K. N. Stevens	C. G. Bell	J. M. Heinz
Dr. A. S. House	P. T. Brady	C. I. Malme
Dr. T. T. Sandel	O. Fujimura	F. Poza
G. W. Hughes		G. Rosen

A. FORMANT TRACKING*

A method of formant tracking based on the use of a digital computer was described in Quarterly Progress Report No. 54, pp. 161-167. By this method, synthetic speech spectra, which are constructed within the computer from a catalog of elemental spectra, are compared with measured spectra of a speech signal. The method has been further refined and evaluated. Comparison of calculated and measured data has been facilitated by displaying the calculated data on the computer oscilloscope in the form of a "synthetic sonagram," with the aspect ratio of the oscilloscope axes identical to that of a real sonagram (see Fig. XIII-1).

Fig. XIII-1. Photograph of computer oscilloscope face during display of results of matching process for the utterance "now", spoken by a female. Each unit on the abscissa represents a time interval of 11 msec. Ordinate values above the time axis represent center frequencies of the analyzing filter bank. Ordinate values below the time axis represent the magnitude of discrepancy between the measured spectrum and the best synthetic spectrum, according to the chosen criterion.

*This work was supported in part by National Science Foundation.

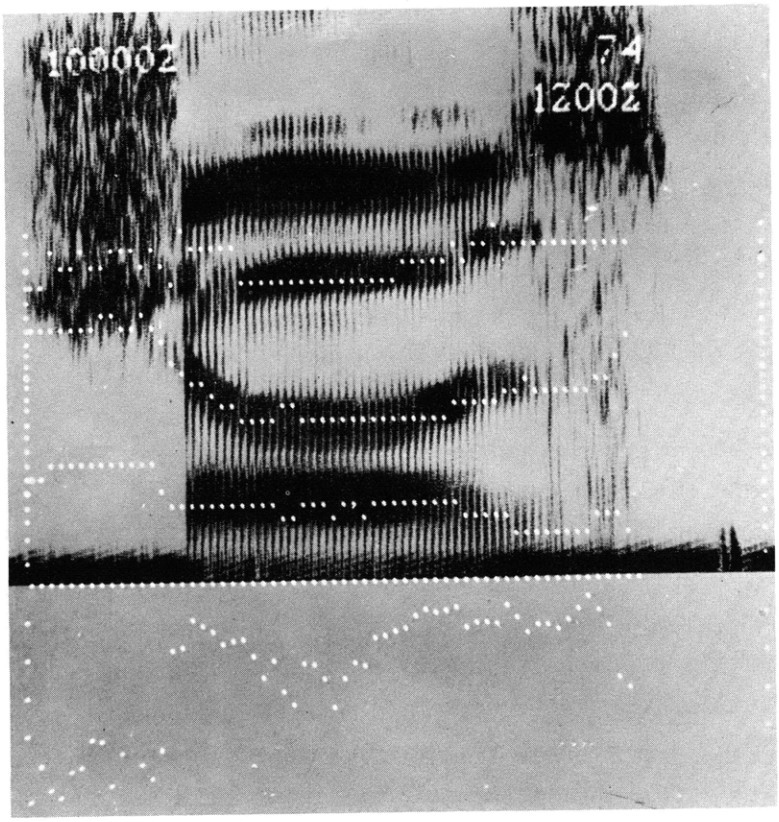


Fig. XIII-2. Superposition of computed results on an actual sonagram for the utterance "shows", spoken by a male. Computed results for the nonvowel segments should be ignored.

Utterances produced by a variety of speakers were subjected to formant tracking, and the results were evaluated by superposing an appropriately enlarged photographic negative of the oscilloscope face on the actual sonagram. Figure XIII-2 is an illustration of this technique. It was found that it is essential to choose appropriate formant positions for the initial comparison. The requisite information must be derived by some means (preliminary analysis) from the measured spectrum. Without such a preliminary analysis the method fails to overcome certain initial errors. It was noted, furthermore, that lack of detailed information concerning the glottal spectrum is a significant source of error and that the correction for higher poles deserves much more attention than was previously thought.

For further details see the author's thesis (1).

F. Poza

References

1. F. Poza, Formant tracking by machine, S.M. Thesis, Department of Electrical Engineering, M.I.T., Aug. 1959.

B. RECOGNITION OF SPEECH BY MACHINE*

This study has been completed. It was submitted as a thesis in partial fulfillment of the requirements for the degree of Doctor of Science, Department of Electrical Engineering, M.I.T., September 1, 1959.

G. W. Hughes

C. CHARACTERIZATION OF CURSIVE WRITING*

The problem of pattern recognition in speech is a very difficult one. Handwriting is a possible alternative physical representation of the set of messages that can be transmitted by speech. It appears to us that the analysis of handwriting is somewhat simpler than that of speech and that such an analysis might provide insights into the problem of speech recognition.

Handwritten texts represent a means of transmitting communications that can be as adequate as printed text under certain circumstances. The essential difference between handwriting and printed text lies in the fact that the latter is restricted to a finite set of typefaces. More particularly, it is restricted by the specification that each occurrence of a given constant (letter or number or punctuation mark) in a text be represented by the same symbol. In handwriting, however, the symbolic representation of the same set of constants is never unique; in fact, the set of symbolic representations for any constant is infinite. (This follows immediately from the fact that the symbol boundaries in handwriting are arbitrary cuts of functions that can be mapped continuously into a segment of the real line.) Many more serious variations (serious from the point of view of geometry or topology) occur without impairing the translatability of handwriting into printed text.

In order to analyze the structure of handwriting, recourse can be had to several strong analogies between handwriting and speech. In fact, it is possible to find a virtual point-for-point correspondence between the beginnings of the analysis described in the following paragraph and certain particular analytical procedures that have been proposed for spoken language. There is an obvious analogy between the concept of "phoneme" in speech and "letter" in handwriting. In the analysis of handwriting, the decision of identification is easier than the comparable decision in speech. Whereas there is some disagreement in the field of linguistics as to the nature of the phonemes

*This work was supported in part by National Science Foundation.

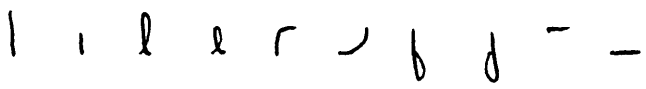
in a given language, there is no disagreement regarding the meaning of "letter" in the study of writing.

In a manner analogous to the distinctive features proposed for phonology by Jakobson and Halle, we devised a set of primitive symbols to characterize the cursive hand used by the majority of North Americans, as well as many Europeans and South Americans. Certain formation rules governing the concatenation of these primitive symbols or "strokes" into letters are suggested. Other rules are suggested for the concatenation of symbols, that is letters (which are well formed in terms of these first-order rules), into larger strings, a certain set of which are the permissible words of a language.

The set of primitive symbols and their designations are:

Designation: A^1 A_1 B^1 B_1 C^1 C_1 D^1 D_1 L^1 L_1

Symbol:



The rules are:

1. Each stroke is executed in a single movement without doubling back.
2. The strokes are written from the top down.
3. The strokes are written from left to right.

Rule 2 must be applied before rule 3. Thus the C strokes are written from the top down, although their general direction is right to left. Also note that rule 2 cannot apply to the B strokes, hence rule 3 applies.

4. Strokes are joined by placing the second stroke to the right of the right-most part of the first stroke.
5. If the terminal of the first stroke is on the same level as the initial end of the second stroke, then the join consists of the shortest straight line at that level.
6. If the direction of the terminal end of the first stroke is horizontal and the direction of the initial end of the succeeding stroke is vertical, the join is the shortest stroke that is concave upward.
7. If the direction of the terminal end of the first stroke is vertical and the direction of the initial end of the second stroke is horizontal, the join is concave downward.
8. A string is well formed if the first and last strokes are L strokes.

A list of the permissible strings, that is, the strings that correspond to letters of the alphabet follows.

a - $L^1C^1A_1L_1$		k - $L_1B^1C^1L_1$		t - $L_1A^1L_1$ + dash	
b - $L_1B^1L_1L^1$		l - $L_1B^1L_1$		u - $L_1A_1L_1A_1L_1$	
c - $L^1C^1L_1$		m - $L^1A_1L^1A_1L^1A_1L_1$		v - $L^1A_1L_1L^1$	
d - $L^1C^1A^1L_1$		n - $L^1A_1L^1A_1L_1$		w - $L_1A_1L_1A_1L_1L^1$	
e - $L_1B_1L_1$		o - $L^1C^1L^1$		x - $L^1A_1L_1$ + diagonal	
f - $L_1B^1D^1L_1$		p - $L_1D_1L^1C_1L_1$		y - $L^1A_1L_1D_1L_1$	
g - $L^1C^1D_1L_1$		q - $L^1C^1D^1L_1$		z - $L^1C_1D_1L_1$	
h - $L_1B^1L^1A_1L_1$		r - $L_1L^1A_1L_1$		Some nontrivial variants	
i - $L_1A_1L_1$ + dot		s - $L_1C_1L_1$		b - $L_1B^1C_1L_1$	
j - $L_1D_1L_1$ + dot				r - $L_1A_1C^1L_1$	

Of course, there is an additional number of trivial variants.

An examination of the handwriting of approximately 20 writers of diverse backgrounds leads to the following observations concerning the concatenation of letters into longer strings.

a. When the terminal stroke of one letter is $L_1(L^1)$ and the initial stroke of the succeeding letter is $L_1(L^1)$, then the ligature between the letters is equivalent to mapping $(\alpha)L_1 + L_1(\beta)$ into $(\alpha)L_1(\beta)$.

b. Ligatures joining letters of the form $(\alpha)L_1 + L^1(\beta)$ or $(\alpha)L^1 + L_1(\beta)$ have no standard form. In fact, either these ligatures are dropped by the writers or certain distortions are introduced into the strokes adjacent to the L strokes. For example, the string "be" is written either by lowering the L^1 stroke of the "b" below its usual level or by raising the B_1 stroke of the "e." These two distortions are illustrated below in the word "beast." It will be noted that the first distortion might suggest the misreading "least," and the second distortion suggest "blast."

The set of strokes provides us with ways of describing the degeneracies that appear in any handwriting that is written without conscious concern for the calligraphic rules. For example, a common instance of degeneracy is to write B strokes like A strokes. It will be noticed that, so long as the diacritic marks are retained, a degeneracy of this

sort introduces no ambiguity into the reading. Thus "ℓ" (l) is designated $L_1A^1L_1$, and "t" remains $L_1A^1L_1$ + dash. However, if the diacritic marks also are dropped, a large number of ambiguities occurs.

The purpose of an analysis of the structure of handwriting is to determine whether or not a meaningful text is readable; the clarity of all of the individual letters is relatively unimportant. It seems important, therefore, to be able to measure the readability, or adequacy, of a given text. A text will be called adequate (it is important to keep in mind that adequacy is always relative to some specified observer) if the observer can convert the handwritten text into printed text (e.g., by typing), that is, into a set of symbols constrained as described in the opening paragraphs of this report.

It is reasonable to conjecture that the readability of handwriting derives in some way from the redundancies of written language. Some of the more obvious categories of redundancy are:

1. Stylistic – The handwriting of a given individual appears to have idiosyncratic characteristics that limit the size of the set of representations of a constant; that is, the occurrence of some letter early in the text will provide the observer with some hints concerning the shape of that particular letter when it next occurs.

2. Semantic – Usually texts carry some meaning. It is to be presumed that when reading a text the observer will look for those interpretations of the symbols that "make sense."

3. Grammatical – It is likely that the observer will choose the interpretation of the symbols that makes the text grammatical.

4. Spelling – The observer might be said to have in his memory a finite set of words, and insofar as the handwritten text supplies him with the correct word boundaries, he would attempt to read the symbols as correctly spelled words.

5. Interletter constraints – Since the occurrence of all letter pairs (or, in general, n-tuples) is not equally likely (in fact, some are forbidden), the statistics of the language, insofar as the observer has an awareness of it, will permit him to recognize combinations of letters that he might otherwise miss. Preliminary tests, however, indicate that the statistics of the sequence is less important than the other categories of redundancy mentioned.

M. Eden, M. Halle

D. AUTOMATIC EXTRACTION OF FUNDAMENTAL PERIOD IN SPEECH SOUNDS*

Tentative schemes for the automatic extraction of the fundamental period in speech sounds have been devised. These methods are based on the analysis of the waveform

*This research was supported in part by the U.S. Air Force (Air Force Cambridge Research Center, Air Research and Development Command) under Contract AF19(604)-2061.

in the time domain, and have been simulated by means of the TX-0 computer.

The basic procedure was to apply 1-bit quantization to signal samples, and to follow this by a short-term autocorrelation analysis with an averaging time as short as one period of the lowest voice that is to be handled. The peaks of the autocorrelation function corresponding to the fundamental period were then detected. To ensure higher reliability, the immediately preceding answer, whenever it was definite, was used as the reference for the determination of the period.

A sampling rate of 10,000 cps and an averaging time of 14 msec were considered feasible on the basis of results of a preliminary test in which monopitched sustained vowels were used as signals. For the sake of simplicity, the fundamental period that is to be examined was limited arbitrarily to the range 2-16 msec. Determination of the period was based on the detection of the regular recurrence of major peaks of the short-term autocorrelation function. Several criteria were used for the detection of major peaks and for the detection of their regular recurrence. This procedure requires a signal section of approximately 40-msec duration for a single measurement of the instantaneous period. In the present version of this scheme, detection of the period can be made for every 50-msec section of speech.

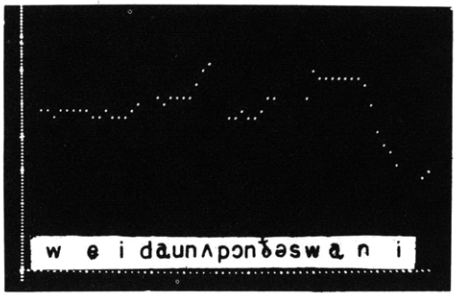


Fig. XIII-3. Result of analysis of the song, "Old Folks at Home." High-pitched female voice: horizontal axis, real time, 50-msec step; vertical axis, extracted period, 100-μsec step. The entire trace corresponds to the opening line, "Way down upon the Suwannee."

Results of the analysis of songs (see Fig. XIII-3) and isolated words show an acceptable accuracy and reliability, which suggest that the scheme could be applied to the analysis of connected speech without basic alteration.

Various elaborations of the detection scheme are being studied. Reduction of the sampling rate, together with appropriate prefiltering of the input signal, and the possibility of equalizing the signal spectrum will be examined in future studies.

H. Fujisaki