A Method of Analysis and Recognition for Voiced Vowels

MARK W. CANNON, JR.

Abstract—A method of speech analysis that has been shown to be capable of recognizing with high accuracy a set of seven voiced vowels spoken by twelve male talkers with various regional accents is described. The waveforms used in the recognition scheme are tapped from four points along a low Q dispersive delay line, which represents a model of the human cochlea. These four output signals are sampled for 4 ms at a rate of 25 000 points per second per output channel. Sampling time is synchronized with the onset of a glottal pulse. The 4-ms samples are autocorrelated on a digital computer, then cross-correlated (at zero delay only) with a set of stored prototype patterns to produce an array of cross-correlation coefficients. These coefficients are treated as components of a multidimensional vector that characterizes the input sound. The final decision as to which sound was spoken is made by a simple linear adaptive network that was trained to separate these multidimensional vectors into their proper classes. The network repeatedly alters a set of decision surfaces until correct classification has been obtained or until a specified number of trials has been exceeded. Successful training was attained in all cases indicating a linear separability of the vowel sounds in the space described by the correlation operations, and the short sampling time used points up the desirability of short time-signal analysis techniques in speech recognition work.

Manuscript received March 11, 1968. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass. The research reported in this paper was conducted by personnel of the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, Dayton, Ohio. This paper has been identified by Aerospace Medical Research Laboratories as AMRL-TR-67-185. Further reproduction is authorized for the internal use of the U. S. Government.

The author is with the Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio 45433

Introduction

IT IS WELL KNOWN that a voiced speech sound is a train of very similar pressure waveforms repeated over and over at the so-called glottal rate or fundamental frequency and that the shape of these waveforms is determined largely by the configuration of the vocal tract, the size of the mouth and throat cavities, and the position of the tongue.[1] Since these parameters vary from speaker to speaker, the speech waveforms for a given sound also vary from speaker to speaker, and even the speech waveforms produced by one speaker change slightly from one glottal period to the next. This variation has been enough to make speech recognition a formidable problem.

Despite this variation, the repeated speech waveforms are quite similar from one glottal period to the next, and there is considerable redundancy in voiced speech. If one considers the human speech production and receiving mechanism as an optimum system, it is apparent that this redundancy is a noise reducing technique. The human lives in a generally noisy environment. He is forced to use noisy elements (neurons) in his internal processing systems and therefore needs this redundancy for successful recognition. Proceeding further along this line of thought implies that one should be considering the repeated waveform as the basic unit of speech and that initial analysis of speech should be done at the glottal rate or faster. There are many indications that this approach leads to useful results. A number of recent approaches to speech recognition, as for instance by Martin, Nelson, and Zadell at RCA[2],[3] and by Piotrowski, Teacher, and Focht at Philco[4],[5] have achieved good speaker-independent results. Both of these methods have one thing in common: their first stage of processing is fast enough to analyze speech events at the glottal rate. In the case of the RCA system, the analysis filters have a Q of 2, insuring very short filter response times over the range of the speech spectrum. The Philco system picks a single equivalent formant by computing zero crossing intervals of the speech wave every glottal period.

This paper considers a somewhat different method of decoding speech, but it is one that takes advantage of short time analysis, which seems to be the most successful approach to speech recognition. The initial analysis is performed by an electronic analog ear[6] instead of the usual parallel filter array. This device, which will be described briefly in the next section, generated four outputs that were sampled in 4-ms segments synchronized with the onset of a glottal pulse. This is equivalent to about one half of a glottal period for a male speaker. These 4-ms samples were then autocorrelated, compared with stored prototype autocorrelation functions, and the results of this comparison were processed through a network that made a choice as to which vowel sound had been spoken. The very high recognition scores obtained by this method point out

that there is enough information contained in one half of one glottal period on a noiseless channel to identify a properly processed voiced speech sound.

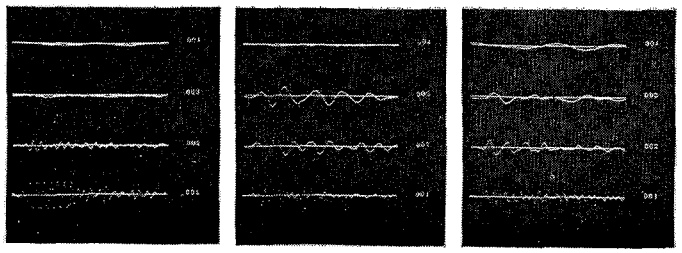
The Analysis System

The first stage of the analysis system, the electronic analog ear,[6] simulates the transformation performed by the middle and inner ear on a sound pressure wave impinging on the outer ear. The basilar membrane in the cochlea or inner ear, is simulated by a 24-section lumped parameter ladder network in which each section has a Q of 2.2, giving a broad frequency response. Cutoff frequencies of the ladder sections decrease logarithmically with distance from the input end. Highest frequencies of about 10 000 Hz are damped out close to the input end while 100-Hz components travel the full length of the ladder. Outputs are available from any of the 24 sections, and one can read out either basilar membrane displacement or velocity. The velocity responses of four sections of the analog basilar membrane were used in this work. These output points were chosen from filter elements equally spaced across the region of maximum activity for speech. The center frequencies of the points were 3000, 1470, 660, and 311 Hz.

Seven voiced vowel sounds were chosen for study. They were i (*see*), e (b*e*t), æ (b*a*t), ʌ (b*u*t), ɔ (*la*w), v (b*oo*k), and u (*too*). The velocity waveforms generated when these sounds were spoken were sampled by an analog-to-digital converter and multiplexer system and stored in a PDP-1 digital computer for further computation. The computer facility also contains a large screen cathode ray tube (CRT) and a digital tape system. Waveforms were sampled four channels at a time at 25 000 points per second per channel and displayed on the CRT before any computations were done by the computer. A total sample length of 16 ms was taken, and up to 8 ms of any portion of the sampled data could be displayed at any one time. Fig. 1 shows three sets of velocity waveforms as displayed on the CRT. These were sampled for 8 ms and adjusted so that the beginning of a glottal period is near the left end of the picture.

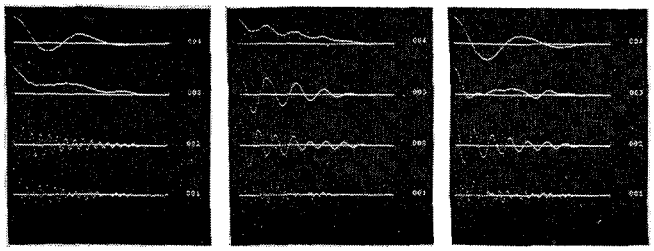
Adjustment of the waveforms was done visually by changing a program parameter that moved the CRT display to the right or left by as many sample points as desired. It is obvious from Fig. 1 that the beginning of a glottal period is well marked by the abrupt rise of the waveform on channel 1 (the bottom waveform in each picture). In each case, the zero crossing before the first large peak of a glottal period on channel 1 was placed within the first half-millisecond interval from the left edge of the CRT display. The position within this interval was not critical and rough visual judgments of its location were found to be adequate throughout subsequent experiments.

Fig. 2 shows three sets of autocorrelation functions computed from the velocity waveforms of Fig. 1 and



i ɔ u

Fig. 1. Basilar membrane velocity waveforms for i, ɔ, and u—sampled for 8 ms.



i ɔ u

Fig. 2. Autocorrelation functions for i, ɔ, and u—computed from the velocity waveforms of Fig. 1.

normalized to have the same amplitude at the origin. These normalized autocorrelation functions were used for analysis, because they were independent of speech amplitude and relatively insensitive to small fluctuations in the speech waveforms, including peak clipping. However, after many observations of the autocorrelation functions, variations in the positions of the peaks and zero crossings for delay times greater than 4 ms became evident for different samples. These differences appeared not only from talker to talker, but from one sample to the next for the same talker. But the portion of the autocorrelation functions for τ less than 4 ms appeared to be quite consistent. Because of this consistency, the analysis system was set up on the digital computer to compute the autocorrelation functions of 4-ms samples of the velocity waveforms. A set of autocorrelation functions was then prepared to serve as prototype patterns in a memory. The prototype patterns were built up from the utterances of six male speakers. Each speaker's mouth was placed 6 inches from an Altec 682 microphone that was connected directly to the analog ear. While the subject sustained the sound, the A-D converter system sampled the analog ear outputs for 16 ms and loaded the data into the digital computer. A 4-ms sample was extracted from each utterance by the visual adjustment process described previously, and autocorrelation functions were computed from these selected 4-ms samples. Each of the six talkers spoke a given speech sound five times. The samples were autocorrelated and then the five sets, of four autocorrelation functions each, were averaged together. The averaging of the ith autocorrelation function was given by

$$\phi_{iav}(\tau) = \sum_{j=1}^{5} \frac{\phi_{ij}(\tau)}{5}$$

where $i = 1 \cdots 4$, j is the repetition number, and τ goes from 0 to 4 ms. This produced one set of autocorrelation functions for each sound by each of the six talkers, that is, there were seven sets of autocorrelation functions for each of the six talkers. Finally, each of the seven sounds was averaged across talkers to give a group of seven averaged autocorrelation sets. The averaged autocorrelation function sets for i, ɔ, and u are shown in Fig. 3. At this stage in the averaging process, only those features common to all six talkers were left in the autocorrelation functions of a given sound, since the averaging would tend to treat speaker-dependent features as noise and cancel them out. It is interesting to note from Fig. 3 that only the first 2 ms show any appreciable amplitude in the high-frequency channels (1 and 2). This implies that the consistency lies in the first 2 ms of the glottal period. Storing these averaged autocorrelation sets on digital tape generated the memory to which autocorrelation functions of subsequent sounds could be compared.

The mechanism of comparison used was cross-correlation at zero delay ($\tau = 0$). When a new sound was spoken, its four velocity waveforms were autocorrelated and compared by cross-correlation with each of the seven stored sets of autocorrelation functions. The result of each cross-correlation was a number between 0 and 1 in increments of 1/10. The total output was 28 such cross-correlation coefficients, four from the comparison with each stored autocorrelation set. It is from these cross-correlation coefficients that the decisions are made for recognition. Each time a sound was spoken, the cross-correlation coefficients were printed out by the computer in order to keep a record of each comparison.

The operation of cross-correlating two autocorrelation functions has more significance than might be realized at first glance. If we consider each of the 28 memory autocorrelation functions as an autocorrelation function of the impulse response of a linear filter, then the cross-correlation, at zero delay, of the autocorrelation function of the input to the filter with the autocorrelation function of the filter impulse response is the mean-squared output of the filter.[7] Thus, we can consider this cross-correlation process as a filtering process, where the outputs of the analog cochlea are fed into an array of matched vowel filters, as shown in Fig. 4, and the recognition decision is made by operating on the power output values of the filters with some decision network. Since the autocorrelation functions exist only out to $\tau = 4$ ms, the unit impulse responses of such filters would also go to zero in 4 ms. This implies that a true analog system containing such vowel filters would be continuously analyzing a 4-ms interval of the speech waveform at any given time. The outputs from these filters could be averaged over a number of glottal periods to smooth out noise and then fed to a recognition network for a decision. Such a system, if adapted to a wide variety of phonemes, would avoid any speech segmentation sampling problems because it would be making continuous computations.

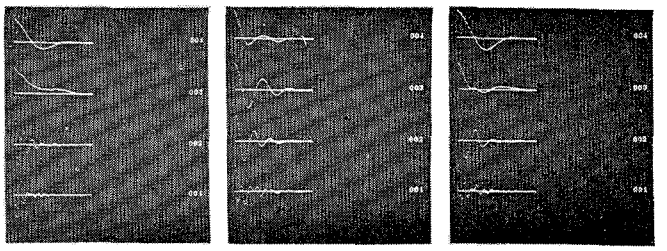


Fig. 3. Autocorrelation functions of i, ɔ, and u—computed from 4-ms samples and averaged over six talkers.

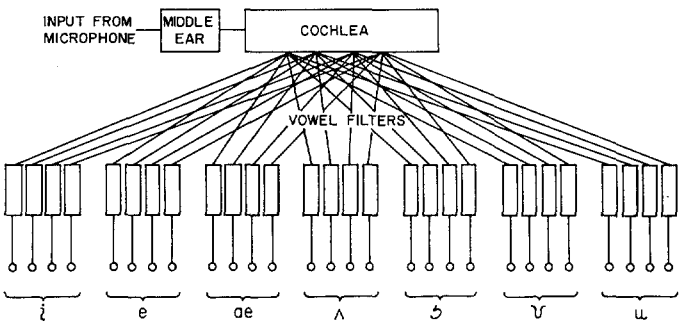


Fig. 4. Block diagram of the speech analysis system.

The Decision Process

For convenience, let us refer to the cross-correlation process as filtering, the stored autocorrelation functions as vowel filters, and the cross-correlation coefficients as filter outputs. Two different recognition methods were tried, one empirical and the other learned by use of a linear adaptive machine simulated on the digital computer. When the basilar membrane velocity waveforms for a given vowel pass through the array of vowel filters, the four filters that are matched to that particular vowel should have the largest overall outputs. The first method used for recognition then was to sum the filter outputs in groups of four as suggested by Fig. 4. This produces seven numbers, each of which indicates the degree to which the particular filter set is matched to the incoming waveforms. The filter set with the maximum output can, therefore, identify the input sound. By simply picking the maximum, an overall recognition rate of 92 percent for 356 samples from eight male talkers was obtained as shown in Table I. There were 20 ties in the output amplitudes. These ties were caused by the fact that the individual cross-correlation coefficients were computed in increments of 1/10 between 0 and 1. After studying the data for a short time, it became obvious that the recognition score could be improved by breaking ties in the following way.

tie	choice
i-u	i
ae-e	ae
ʌ-ɔ	ʌ
ʌ-v	ʌ
v-u	v

The number of talkers was expanded to 10, and when the decision rules for tie breaking were in force, the overall recognition rate rose to 95 percent. Table II gives the breakdown of recognition scores for the individual vowel sounds showing that the lowest recognition score was 92 percent.

One more obvious feature that would have corrected the five u errors was picked out. These particular samples were mistaken for e or ae, because the output sum of the e or æ filter set was slightly larger than the output for the u filter set. However, any time an e or æ was spoken, the output of the u filter was below a certain threshold. When u was spoken, it was always above this threshold.

One more decision rule, which checked the output of the u filter and compared it to the threshold value, would have picked out all the u's correctly and given a 100 percent recognition rate for this vowel, raising the overall recognition rate to 96 percent for 424 samples from ten male talkers.

A certain amount of information that could be used for recognition was distributed throughout the array of 28 cross-correlation coefficients or filter outputs. Therefore, a pattern recognition approach was tried using a linear adaptive machine[8] to learn the proper decision rules for classifying the 28-dimensional vectors represented by the cross-correlation coefficients. This machine, simulated on a digital computer, contains one element for each of the sounds to be recognized. Each element performs the dot product of the 28-dimensional input vector and a 28-dimensional weight vector. This dot product gives the distance from any input vector to a certain hyperplane in this 28-dimensional space. For training, the machine processes a set of vectors with each vector identified as to which of the seven vowels it represents. If the input is an i and the dot product computed by the i unit is ≥ 0 with all other units having outputs <0, the machine has made a correct choice. If this occurs during training, the machine goes on and looks at the next sound. If the machine makes a mistake during training, all the weights are adjusted until the dot products have the correct sign. This learning process may continue for hundreds of times through the training set until all vectors have been correctly identified. Incidentally, a particular weakness of this type of linear machine is that if there is some overlap in the regions of this 28-dimensional space occupied by a pair of sounds, the training procedure will never converge to a solution. Weight vectors will never be found for sound classes not separable by hyperplanes.

TABLE I

RECOGNITION SCORES OBTAINED BY PICKING THE MAXIMUM VOWEL FILTER OUTPUT

Sound	Recognition Rate
i	98%
e	91%
ae	85%
ʌ	87%
ɔ	93%
v	93%
u	93%

Total Number of Sounds—356
Total Number of Errors—11
Total Number of Ties—20
92 Percent Overall Recognition Rate with 8 Male Talkers

TABLE II

RECOGNITION SCORES OBTAINED BY THE ADDITION OF TIE BREAKING LOGIC

Sound	Number of Samples	Number of Errors	Recognition Rate
i	59	0	100%
e	57	3	95%
ae	58	3	95%
ʌ	62	1	99%
ɔ	58	4	93%
v	70	3	96%
u	60	5	92%

Total Number of Sounds—424
Total Number of Errors—19
95 Percent Overall Recognition Rate with 10 Male Talkers

When the machine has been trained on a specified number of patterns, the learning mode is terminated and the recognition mode initiated. The weight vectors are frozen, and the machine is given an entirely new set of sounds to recognize. In this mode, it will classify all incoming sounds as one of the set of seven it "knows" or will reject if all dot products are negative, giving the message "sound not recognizable." Fifty-six sounds from four male talkers composed the first training set. After learning to classify these vectors correctly, the machine was put into recognition mode and given a new set of correlation coefficients produced by five male talkers. The results shown in Table III gave the rather low overall recognition rate of 86 percent. However, after five training sessions in which the number of sounds in the training set was increased to 250 and the number of talkers represented rose to 10, the machine showed considerable improvement in the recognition mode, as can be seen in Table IV in which the recogni-

TABLE III

RECOGNITION SCORES OF THE ADAPTIVE NETWORK AFTER THE FIRST TRAINING SESSION

Sound	Number of Samples	Number Correct	Number of Errors	Number Not Recog-nizable	Recognition Rate
i	12	10	2	0	84%
e	10	10	0	0	100%
ae	11	10	0	1	91%
ʌ	14	11	2	1	79%
ɔ	12	10	0	2	84%
v	13	10	1	2	77%
u	10	10	0	0	100%

86 Percent Overall Recognition Rate with 5 Male Talkers

TABLE IV

RECOGNITION SCORES OF THE ADAPTIVE NETWORK AFTER THE FIFTH TRAINING SESSION

Sound	Number of Samples	Number Correct	Number of Errors	Number Not Recog-nizable	Recognition Rate
i	10	10	0	0	100%
e	10	10	0	0	100%
ae	10	9	1	0	90%
ʌ	11	10	1	0	91%
ɔ	10	10	0	0	100%
v	10	10	0	0	100%
u	10	10	0	0	100%

97 Percent Overall Recognition Rate with 5 Male Talkers

tion rate has risen to 97 percent. There is not enough data available yet to compare the adaptive recognition method to the empirical one, since the sample size of the latter is so much larger. However, the 100-percent scores for i, ɔ, and u have held up since the third training sequence, and since then at least 30 samples of each sound have been presented to the machine in recognition mode. Also, the five u sounds missed by the maximum output method were easily learned by the adaptive machine.

CONCLUSIONS

The concept of analyzing the cochlear processed speech waveforms for times less than one glottal period was proved quite successful in identifying the seven voiced vowels considered. Comparing the autocorrelation functions of the waveforms to a stored set of autocorrelation functions has been shown to be essentially the same as running these waveforms through an array of matched vowel filters. In a sense, construction of these vowel filters is a learning process of the simplest kind—the computation of an average filter response for a given sound and a variety of speakers. An array of filters so constructed from the analog ear responses to six talkers has been shown capable of accurately selecting the correct sound by the maximum selection criterion, even when the number of talkers was expanded to ten. This implies that averaging the original six talkers removed much of the speaker-dependent features from the memory autocorrelation functions. These filter outputs have also been shown to be sufficient for recognition by a linear adaptive machine at a potentially higher recognition rate.

REFERENCES

[1] H. Fletcher, *Speech and Hearing in Communication*. New York: Van Nostrand, 1953.

[2] T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech recognition by feature-abstraction techniques," RCA, Camden, N.J., DDC AL-TDR-64-176, August 1964.

[3] T. B. Martin, A. L. Nelson, H. J. Zadell, and R. B. Cox "Recognition of continuous speech by feature abstraction," RCA, Camden, N.J., DDC AFAL-78-66-189, May 1966.

[4] C. F. Teacher and C. F. Piotrowski, "Voice sound recognition," Philco Advanced Communications Lab., Rept. RADC-TR-65-184, April 1965.

[5] L. R. Focht and C. F. Piotrowski, "Voice sound recognition," Philco Advanced Communications Lab., Rept. RADC-TR-66-507, October 1965.

[6] J. L. Stewart, "Speech processing with a cochlear neural analog," Santa Rita Technology, Menlo Park, Calif., DDC Rept. AMRL-TR-66-229, February 1967.

[7] Y. W. Lee, *Statistical Theory of Communication*. New York: Wiley, 1963, ch. 13, pp. 334–336.

[8] N. J. Nilsson, *Learning Machines*. New York: McGraw-Hill, 1965.

[9] W. A. Grimm, "Perception of segments of English-spoken consonant–vowel syllables," *J. Acoust. Soc. Am.*, vol. 40, pp. 1454–1461, 1966.

Mark W. Cannon, Jr., was born in Pittsburgh, Pa., on June 17, 1936. He received the B.S. and M.S. degrees in physics from the University of Pittsburgh, Pittsburgh, Pa., in 1958 and 1962, respectively. He entered the Air Force in 1962 and was assigned to the Biodynamics and Bionics Division of the Aerospace Medical Research Laboratories at Wright-Patterson Air Force Base, Ohio. In 1965 he became a civilian employee of this organization. He has been active in developing computer simulations for neural models and speech recognition techniques using adaptive components.