Computer Recognition of Speech Using Zero-Crossing Information

G. D. EWING, Member, IEEE
JOHN F. TAYLOR, Member, IEEE
Naval Post Graduate School
Monterey, Calif.

Abstract

In the present age of scientific discovery, man has become more and more dependent on the use of electronic computers. As this powerful tool becomes more universally important in man's day-to-day existence, it becomes increasingly more annoying that he has to speak to it in its mode of communication, paper tape or punch cards; and not in his own, the spoken word. Even today in the very infancy of the computer age, the time required to do many computations is less than the time required to instruct the machine in how to do them. All this points to the need of a method of achieving machine recognition of speech.

In this paper an electronic speech processor is described which provides an analog voltage output based on the difference signal between the first speech formant and the second. Machine recognition of numbers zero through nine was very good when the speech processor output was sampled and compared with previously recorded memory-stored data in a small digital computer.

Manuscript received October 5, 1968; revised December 16, 1968.

I. Nature of Clipped Speech

If a speech wave is viewed in the time domain, that is, a plot of amplitude versus time, an obvious characteristic is the great dynamic range of amplitudes that are present. Variations of up to 60 dB are not uncommon in normal speech. In particular, it is noted that vowels are on the average 12 to 28 dB higher than the consonants. This wide dynamic range of normal speech presents problems in speech processing for transmission, since a transmitter would have to work at a very low average power (and, of course, lower range) if the exact shape of the speech wave were preserved. In order to increase the average power, work has been done in the area of speech clipping.

It has been found by various researchers that clipping the original speech waveform up to 12 dB has little effect on intelligibility [1]. Clipping of about 12 dB sounds as if the speaker were enunciating carefully.

Actually, by reducing the peaks which are primarily associated with the vowels, the process serves to enhance the relative power in the consonants. Since the consonants are much more transitory in nature than the vowels, it is appealing to say from an information theory point of view that they are the primary information-bearing elements in the signal, and to increase their relative power is to increase the emphasis on the information content of the speech wave. It also follows from this that the individual speaker characteristics are contained more in the vowels than in the consonants. One would, therefore, expect clipped speech to be somewhat less indicative of speaker voice traits, and this is an experimentally proven fact.

The reason that clipped speech is intelligible can further be seen from a frequency spectrum point of view. It is a basic fact of nature which is easily proven that the human ear is insensitive to phase. It has been common to assert that the information content of a speech wave is contained in the energy spectrum of its various frequency components. If the relative phases of these components are varied within limits, thereby producing a wholly different amplitude versus time pattern, the ear would notice no difference.

The importance of clipped speech to the work in this paper is the relationship of its zero-crossing rate to the formants of the speech sounds. Chang *et al.* have demonstrated mathematically that the average rate of zero crossing of the undifferentiated speech wave is very nearly a measure of the first formant frequency [2]. Furthermore, the average rate of zero crossing of the differentiated wave is a measure of the second formant frequency.

From the above interrelation of clipped speech and the first two formants, which are strongly believed to be the information-bearing elements of the speech wave, it is proposed that equipment could be designed to obtain the formant frequencies via the clipped speech zero-crossing rate. These formant frequencies could then be used together, or perhaps with the assistance of some other speech parameters, to distinguish spoken words for at least a limited vocabulary and for a variety of speakers.

This proposal is based on the appealing assumptions that the formant frequencies do contain the information and that the individual speaker characteristics can be eliminated by going to the formants via the clipped speech zero-crossing approach. We have found no indications in the literature, save that discussed below, to show that anyone else has attempted to verify Chang's conclusions for speech sounds.

The Vector Display

In a report on a government-sponsored research effort on signal processing by infinite clipping, Pyron and Williams discussed a vector display unit which they had developed for visually displaying voice and other short-time highly transient signals [3]. They found that an analog signal proportional to the short-time running average of the zero-crossing rate of the original or differentiated speech wave was quite similar for the same sound by many speakers and distinctly different for other sounds. Another analog was produced proportional to the smoothed envelope of the amplitude of the original waveform and used as a second coordinate for an oscilloscope display. This display, called a vector display by the originators, consisted of the averaged zero-crossing analog applied to the vertical plates and the amplitude analog applied to the horizontal plates of a storage-type oscilloscope.

The authors reported the patterns produced by the vector display had a tendency to correlate well for spoken words and seemed independent of individual voice characteristics. In particular, the patterns produced using the differentiated waveform seemed to give the most distinctive shapes. This was as expected since the differentiated waveform is felt to carry more intelligence than the original for infinitely clipped speech.

As a beginning point in our investigation we constructed the vector display as presented by Pyron and Williams and attempted to verify and improve upon their results. The objective was a series of distinctly different patterns for different spoken numbers, but reasonably alike for the same number by various speakers. If such could be achieved, the ultimate objective was to use a digital computer for recognition of the patterns as the numbers they represent.

Our experience with this display was not encouraging for the goals we had in mind. The patterns obtained were not unique enough for individual sounds nor consistent enough for the same sound by various speakers. To be sure, some sounds do have quite distinctive features, particularly those containing plosives or fricatives such as "ship" or "tooth," but there seemed to be too many exceptions to make such a system workable.

In the course of working with the vector display the idea suggested itself that a display of averaged zero-crossing rate of the original waveform versus that of the differentiated wave should be of greater interest. This type display was appealing for the following reasons.

1) It would be a pattern defined by the first and second formants of the speech wave (as derived theoretically and demonstrated experimentally by Chang *et al.* [2]) which are held to be the information-carrying elements of the speech wave.

2) It would eliminate the amplitude parameter from the somewhat promising vector display, a parameter whose phase dependence made its value suspect from the very beginning.

Memoscope displays of first formant versus second formant were next generated for the numbers zero to nine by several male speakers. It became immediately obvious that this type pattern, although most promising in theory, left a trace that was too confusing for worthwhile analysis. It was apparent that if anything useful was to be obtained from this combination, another parameter would have to be included to spread the formant versus formant excursions of the trace out more from the origin of the axes.

Time was the obvious other parameter chosen and it was included in the present plot by applying a time sweep to both sets of deflection plates of the memoscope, producing a time sweep diagonally rising across the scope.

Patterns produced by this type of processing were very promising, but for the numbers zero to nine there was a need for more individuality in some of the patterns, especially those not containing plosives, fricatives, or stop consonants. Band limiting of the input wave to either or both channels offered possibilities of improvement, as did asymmetrical weighting of the formant channels. The second formant is felt by many speech researchers to be the principal carrier of information and so it seems reasonable to give it more emphasis in this kind of plot. More work was not done with this type of display because the similar approach to be discussed next gave much more interesting results at the same level of investigation.

Formant Two Minus Formant One versus Time

While working with the processing method just discussed, it became apparent to us that the display being studied was a vector sum of the two formants with time (not mutually orthogonal vectors). A simple arithmetic difference display had been overlooked and with no justifiable reason. Such a combination would have the effect of cancelling the portions of the formants that are similar at the same time.

Patterns were generated and studied for the numbers zero through nine by three male speakers. It was noted that this type of pattern had to a fair degree the desired simplicity and uniqueness needed for the goal of machine recognition of speech. The patterns showed very good consistency for a given sound by various speakers, and, although not absolute uniqueness for different sounds, enough variance to make more work here feasible.

Rather than rely on visual consideration as done previously, it was decided at this stage of the investigation to feed the analogs being generated into a computer for comparison. Since such work would be best done in a real time environment, both from the standpoint of study

and as an ultimate machine recognition capability, it was decided to do this work on a small, but more readily available computer, the Control Data 160. To be sure, forsaking the capabilities of the larger computers available at the Naval Postgraduate School, the IBM 360 and the SDS 930, required greater effort in programming and provided a lesser degree of potential operations. However, for testing and evaluating a system such as this the advantage of working in a real-time situation cannot be overestimated. Also, the value of any processing scheme is increased if the required computer capacity is held to a minimum.

In selecting a method of pattern comparison for the computer to execute, it is immediately suggested to one who has studied some communications theory that a correlation technique should be used. Cross correlation is defined as a graph of the similarity between two waveforms as a function of the time shift between them. However, if cross correlation is considered as a matched filtering process, which it is, then the uselessness of this method in this work becomes apparent.

When a signal is cross correlated with another, it is equivalent to an autocorrelation of that signal with itself plus noise. The effect is that the process acts as a filter and only allows through those frequencies which are in the signal. Thus, this method of signal processing is very powerful where you have a high-frequency signal buried in wide-band noise, such as the radar problem, but of little use when the signals of interest are band limited to below 300 Hz. As a comparator, correlation gives an average measure of the similarity between two waveforms; it is quite insensitive to local differences in the amplitudes of the two waveforms. Since local differences of the analog waves generated are the precise means by which we have attempted to perform machine recognition, correlation techniques would not work.

The poor performance of correlation techniques in a low frequency problem has been experimentally shown by Bezdel [4]. He noted poor results using correlation methods, although he does not explain why. A little thought on the matter makes one realize it would have been an anomaly if his results had been good, since the tool has little power at these frequencies.

Bezdel and Chandler indicated that they had success in their comparison work using an Euclidean distance measurement, that is, a point-by-point difference calculation between corresponding points on the "unknown" sound and a previously stored "dictionary" sound [5]. The dictionary word yielding the least total difference from the unknown would be selected as the best comparison. This technique was employed in our work since it was simple in concept and easily programmed within the limitations of the CDC 160 computer.

Conclusions and Recommendations

The numbers zero through nine were recorded by five male speakers. Computer recognitions were tried using each of the voices as a dictionary. Results were excellent for the same voice against itself, as would be expected. Attempts at intervoice comparison were not consistently successful for certain of the numbers, as discussed below.

Numbers six, seven, and eight were very unique in their patterns and were easy to match for many different voices. The remaining numbers were different enough to provide good identification if the speakers spoke normally and clearly. It was noted that people frequently try to speak very clearly (and usually so much so that it is unnatural) when asked to speak into a microphone for testing. This presented a problem in identifying the number three. Some speakers said "th -ree" and this gives a different pattern from the monosyllabic version of the word. Aside from such anomalies results were very good, especially if each speaker heard how the others pronounced the words.

From the above testing it was apparent that some work would be required to make the patterns that were close in shape more unique, so more leeway might be allowed for individual speaker mannerisms. The input bandwidth was varied for each channel with the hope of increasing the difference between the patterns. Since the first formant is expected to exist somewhere below 1 kHz, this channel was bandlimited between this frequency and 300 Hz. The second formant exists somewhere between 800 and 4000 Hz, and so this channel was limited to that frequency range. Other bands were also tried, but these settings seemed to do as well or better than any others and are reasonable for the parameters being extracted.

Under these new conditions results obtained for the same five voices as above were improved, but problems still existed for the numbers one, four, five, and zero. Results indicated that the real key to discriminating between the patterns rests on the substantial excursions caused by the plosives, affricatives, and fricatives, and those words which contain none are inherently in trouble. With practice all the speakers tested began saying even these words the same and higher recognition scores were realized. In this regard the system functioned well as a speech training aid because all found it quite easy to enact the speaking enunciation of the best speaker and to obtain his good patterns. For people who speak clearly and crisply, results for this ten-word vocabulary would be very good.

After testing the display as discussed above, it was evident that this scheme as it now stands is not sufficient for dependable computer recognition of speech. There is enough information available in these patterns to render them far from valueless, and to add support to the theory of interconnection of the zero-crossing rate and the formant frequencies, but more information is needed for errorless identification of speech.

Better results would seem possible for this type of comparison if an average of several voices were used as a dictionary. This would tend to minimize particular voice characteristics and accentuate the general sameness of the words being spoken. Time did not permit us to explore this possibility. It is recommended to anyone who wishes to pursue this work further.

If one is to hold the theory that the formants contain the information, and further, that the zero-crossing rate is a measure of the formant frequencies, then it is logical to say that sufficient information is available here for error-free speech identification, and the problem lies in the way this information is being handled. It was not proposed at the outset of this work that an Euclidean comparison of the patterns was the optimal way of performing recognition, and the results tend to indicate that it is far from satisfactory. Since the speech signals are statistical in nature, it is reasonable to expect that any comparison system that does not allow for such a nature will not be adequate. It is proposed by these researchers that statistical methods be employed in future comparison work.

In conclusion, it seems that the first and second formant analogs, as extracted from the speech sounds here, are a worthy measure of the intelligence being transferred, and more work in their processing is warranted.

REFERENCES

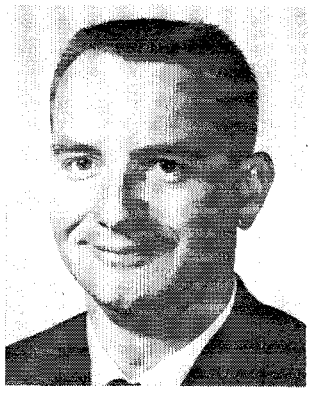
[1] J. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping on the intelligibility of speech," *J. Acoust. Soc. Am.*, vol. 20, pp. 42–51, January 1948.

[2] S. Chang, G. E. Pihl, and M. W. Essigmann, "Representations of speech sounds and some of their statistical properties," *Proc. IRE*, vol. 39, pp. 147–153, February 1951.

[3] B. Pyron and F. Williams, Jr., "Signal processing by infinite clipping and related techniques," Georgia Institute of Technology, Atlanta, Ga., Final Report, Project A-727, U. S. Govt. Contract DA 49-092-ARO-21, April 1964.

[4] W. Bezdel and H. J. Chandler, "Results of an analysis of recognition of vowels by computer using zero-crossing data," *Proc. IEE* (London), vol. 112, November 1965.

[5] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*. New York: Macmillan 1962.



Gerald D. Ewing (M'60) was born in Alliance, Nebr., on January 6, 1932. He received the A.A. degree in engineering from the College of Marin, Kentfield, Calif., in 1955, the B.S. and M.S. degrees in electrical engineering from the University of California, Berkeley, in 1959, respectively, and the professional degree of Electrical Engineer and the Ph.D. degree in electrical engineering from Oregon State University, Corvallis, in 1962 and 1964, respectively.

From 1956 to 1958, he was a designer of nuclear event counting and field measuring equipment and instruments for the University of California Lawrence Radiation Laboratory. From 1958 to 1960, he was employed by Sylvania Electric Products, Electronics Defense Laboratory, Mountain View, Calif., as an Electronics Engineer doing research and development in the field of electromagnetic countermeasure receiving devices. From 1960 to 1961, he was a Semi-Conductor Application Engineer for Rheem Semiconductor Corp., Mountain View. In 1961 he joined Shockley Transistor, Unit of Clevite Transistor, Stanford Industrial Park, Palo Alto, Calif., as Supervisor of Application Engineering. In August, 1963, he came to his present post as Associate Professor of Electrical Engineering at the Naval Postgraduate School, Monterey, Calif., where he teaches and does research in electronic circuits and systems.

Dr. Ewing is a member of Sigma Xi. He has received the following awards: Collins Radio Company Scholarship, 1961, Foundation for Instrumentation Education and Research Scholarship (FIER Award for 1962 and again for 1963), NASA—Stanford University—American Association for Engineering Education Summer Fellowship, 1964.

John F. Taylor (S'66–M'68), was born in Providence, R. I., on April 5, 1939. He received the B.S. degree in physics from Providence College in 1961, and the degree of Electrical Engineer from the Naval Postgraduate School, Monterey, Calif. in 1968.

Since 1961 he has been on active duty as a Naval Flight-Officer serving until 1965 as an aerial navigator and antisubmarine warfare tactical coordinator on patrol-type aircraft. From 1965 to 1968 he studied communications engineering at the Naval Postgraduate School, where the work on this paper originated. He is now a lieutenant undergoing training for reassignment to an operational fleet patrol squadron.