The Use of the IBM 704 in the Simulation of Speech-Recognition Systems

G. L. SHULTZ[†]

Introduction

THE TERM speech-recognition device usually brings to mind a machine capable of duplicating the function of a human listener. Such a device not only would have to be capable of receiving and classifying acoustic stimuli, but also would have to be able to extract from these stimuli the message the speaker intended. To do this, a recognition machine must be familiar with the language statistics, and indeed, the entire human environment. Many years of investigation will be required before such a human replacement can be achieved.

However, more limited man-to-machine communications systems can be defined, and *acoustic* (as distinguished from speech) recognition devices can be built in the near future. Articulation tests with nonsense syllables show that listeners can agree upon, and classify, speech sounds on the basis of the acoustic signal with little use of language redundancy. Surely, then, a set of measurements exists by which a machine could likewise classify these sounds. Since classification of speech sounds is a necessary part of even the most comprehensive recognition system, our efforts are first turned to this task.

The study by introspection how we, ourselves, classify speech sounds is not very successful. We need to "look at" speech. The sound spectrograph was developed by Bell Laboratories to produce visible speech. An example of its display is shown in Fig. 1. The more familiar voltage amplitude vs time function is shown at the top of the figure. Directly below, and aligned in time, is the sound spectrogram of the same utterance. The frequency is scaled along the ordinate. Intensity at a given frequency and time is depicted by the blackness of the mark. Rules for reading these displays have been developed and reported in the literature by Bell Telephone Laboratories, M.I.T., Haskins, and others. These rules are being presented to us in a unified course conducted by Prof. Morris Halle at M.I.T.

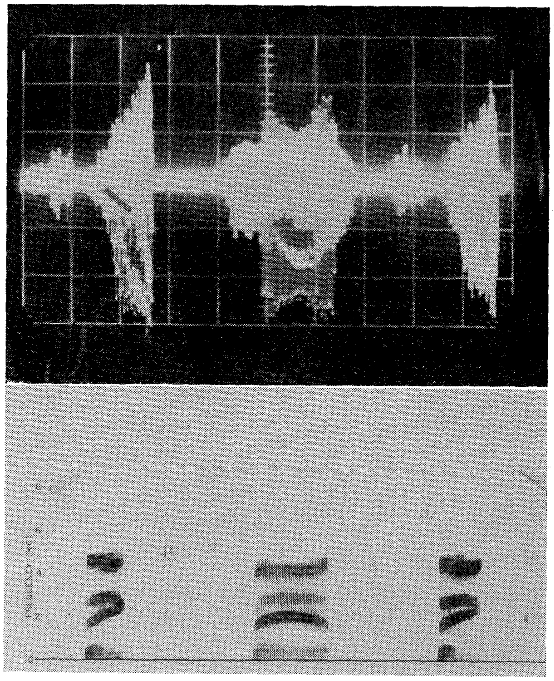

Fig. 1—Comparison of amplitude vs time function and sound spectrogram.

Note that the spectrogram is divided into segments of no activity, horizontal bar structure, and areas of striation. The horizontal bars, termed formants, characterize the vowels, (*r*), (*l*), and nasal consonants(*m*), (*n*), (ŋ). The irregular striated areas characterize the fricative or noise-like consonants (*s*), (ʃ). A vertical blank area, followed by a sharp vertical line, is the distinguishing property of stops or plosives (*t*), (*p*), (*k*). The presence of a heavy horizontal bar at the bottom of the spectrogram indicates voicing. Shown here, the first word consists of an unvoiced fricative, a changing vowel sound, and a terminating unvoiced stop. The second word is voiced throughout and begins and ends with a nasal sound characterized by abrupt transitions in the formants of the middle vowel. The third word is the same as the first. Although we have only mentioned some rules for classifying vowels, fricatives, and stops, such rules have also been developed for subdividing these three classes into the approximately 40 basic elements of speech.

These qualitative rules must be operationally or quantitatively defined in solving the problem of mechanical recognition. For example, just what circuit would be able to identify a striated area? Even after quantitative rules have been defined, a set of physical properties will result whose ranges of variation with context and speaker must be determined. This calls for a statistical approach with its consequent data-handling problems. Further, the number and complexity of these speech properties require a versatile system of analysis. To accomplish this analysis, special advantage is taken of techniques made possible by the advent of the large-scale digital computer.

A computer can be programmed to duplicate any of the measurements of speech signals now used in speech studies.

[†] IBM Corp., Yorktown Heights, N.Y.

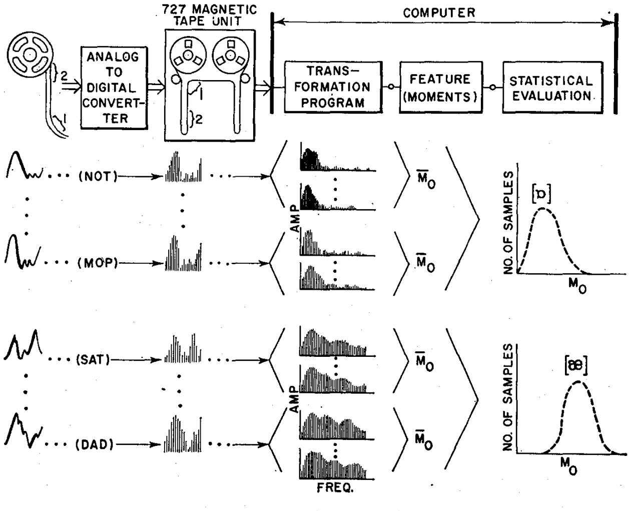

Fig. 2—System of analysis.

The computer is especially well suited for the large-scale reduction and analysis of data. The flexibility of the computer enables a recognition system to be evolved through easily inserted program modifications. Further, the inertia associated with the construction of specialized equipment is eliminated.

When a set of speech properties has been found, and a successful system based on this set has been duplicated in the computer, *then* the system can be embodied in the circuits of a speech-recognition machine.

System of Analysis

Using the computer as a central tool, we have built up the system of analysis as outlined in Fig. 2. In order to gather a large number of like speech events, a device is required to edit these events from continuous speech. Once a speech event has been selected, the acoustic wave must be converted to a digital form satisfying the input requirements of the computer.

The initial program routines are designed to aid in determining which speech properties are most significant with regard to recognition. First, the proposed property is computed for a large number of speech events. Then, the statistical distribution of these measurements is estimated and listed by like speech event.

We have divided the measurement of speech properties into two operations. The transformation block in Fig. 2 contains a basic program which yields a quantitative pattern of the acoustic signal. The many properties of this pattern are then explored by a set of simpler features programs.

Analog-to-Digital Converter

The input system of equipment required for the computer analysis consists of two machines, an Editor and a Coder, each with dual-track, audio-tape devices. Fig. 3 is a photograph of the Editor (left) and Coder (right). The Editor aids in the selection of speech events from continuous speech. The Coder converts these selected events to digital form. Care has been taken to make this system automatic so that the main concern of our work can be the study of speech events rather than their preparation. The desired speech event is selected by the Editor under push-button control. Editing is accomplished by transferring the section of speech containing the speech events of interest to an endless tape, or loop. Synchronizing pulses are placed on the second tracks of both the input and endless tapes as the speech is transferred. When the loop is read, two electronic delays, as shown in Fig. 4, are initiated by the loop synchronizing pulse. One delay is adjusted to extend from the loop synchronizing pulse to the start of the speech event. A second delay, oscilloscope-sweep length, is then adjusted to the duration of the speech event. This second delay also keys the selected portion into earphones, permitting simultaneous sight and sound adjustments.

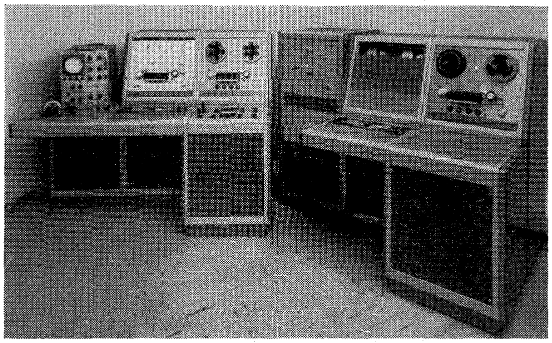

Fig. 3—Editor and Coder.

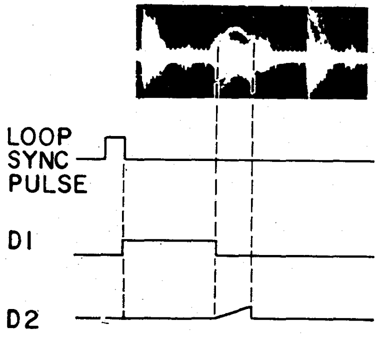

Fig. 4—Edit timing delays.

After the delays are properly set in the endless tape operation, they are initiated once more by the synchronizing pulse on the second track of the input tape. This final timing sequence properly writes an editing pulse on this track. Hence, the result of the editing operation is an audio tape with speech recorded on one track and editing pulses located on the second track opposite the selected speech events.

Fig. 5 shows the entire Editor-Coder operation in block diagram form. It has been split here for convenience into the audio and pulse tracks. As mentioned, the audio signal is recorded once on the Editor, and during the editing op-

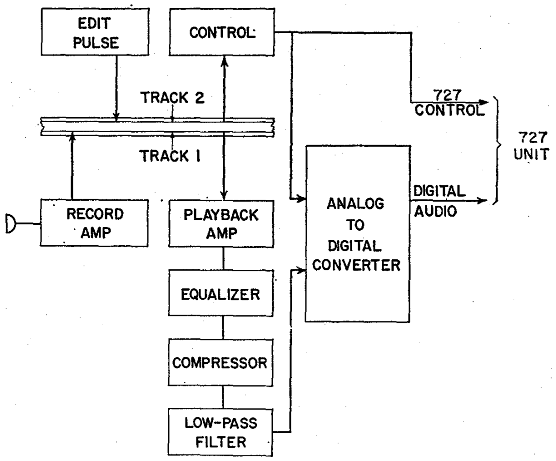

Fig. 5—Editer-Coder audio and timing.

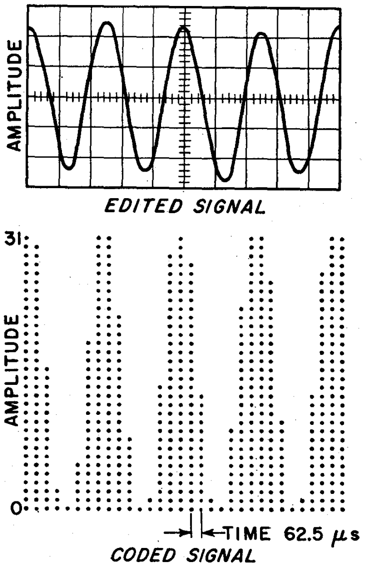

Fig. 6—Selection accuracy.

eration a pulse is affixed on the second track of this tape opposite the speech events selected. This tape is then played back in the coder where the audio information is *continuously* passed through a playback amplifier, an equalizer, a compressor, and a low-pass filter before entering the analog-to-digital converter. Upon reading a pulse from the second track of the tape, the converter, externally triggered at the prescribed 727-tape character rate of 16 kc, converts and speech-signal amplitude at each sample point to an 11-bit binary number. Simultaneously, the editing pulse causes control circuits to bring the 727-tape unit up to speed and properly write the converted speech signal.

One 727-tape record is made for each speech event selected. At present, only the five most significant bits are written on digital tape since this accuracy yields a sufficiently low quantizing noise value. The speeds (7½ or 15 ips) of the input tape in the Editor and the playback tape of the Coder can be arranged to produce effective sampling rates of ½, 1, or 2 times the tape character rate.

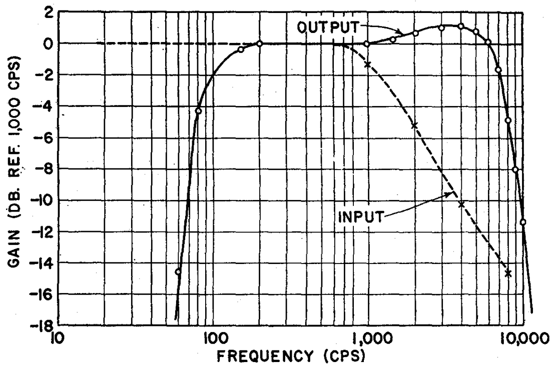

Fig. 7—Frequency responses of Editor-Coder.

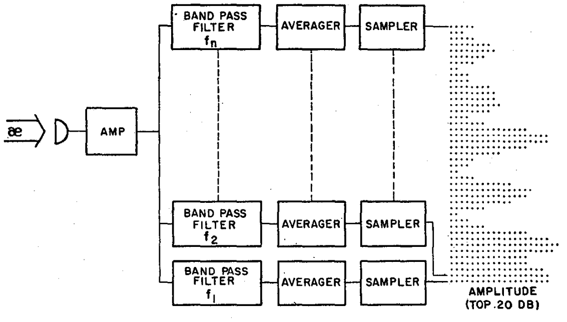

Fig. 8—Diagram of filter bank with spectrum.

The selection accuracy of the Editor is shown in Fig. 6. The 2-kc sine wave is shown at the top of the figure as it was selected in the editing operation. Below is shown a print out of this sine wave sampled at 16 kc. The accuracy of selection is primarily limited by tape-speed variation during delay 1. For a delay of 1 second from the synchronizing pulse to the start of the speech event, the start is located with an accuracy of approximately 2.5 msec.

The audio system response is shown in Fig. 7. Here the input signals were attenuated at a rate of 6 db per octave above 1000 cycles. With this input and the high frequency emphasis circuit inserted, the over-all audio response between half-power points is from 85 to 7500 cycles.

System of Programs

A set of programs has been written to implement the general analysis system. These programs have been tested together on a set of vowel phonemes.

The first transformation program was written to compute spectra. Essentially this program simulates a bank of band-pass filters as depicted in Fig. 8. The output of each filter is averaged for a certain period of time and this average output is sampled periodically. At each sample time the output amplitudes of the entire bank of filters are plotted as a function of the center frequency of each filter. The resulting graph is shown to the right of the figure. In this graph, then, we have two of the constituents of the spectrogram, namely, frequency and amplitude.

A series of these graphs at adjacent sample times would display amplitude and frequency as a function of time. The

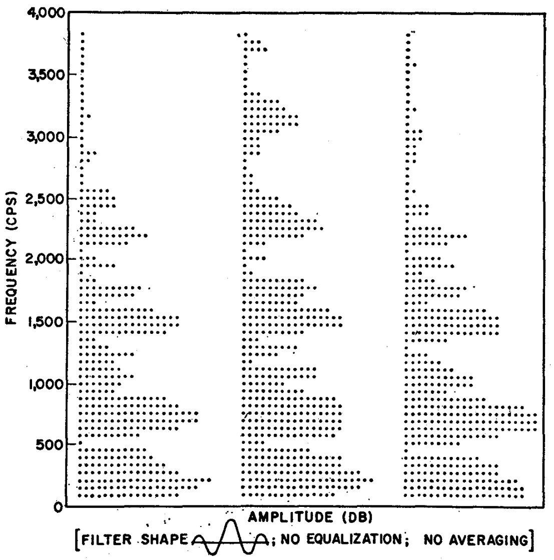

Fig. 9—Spectra produced by original program.

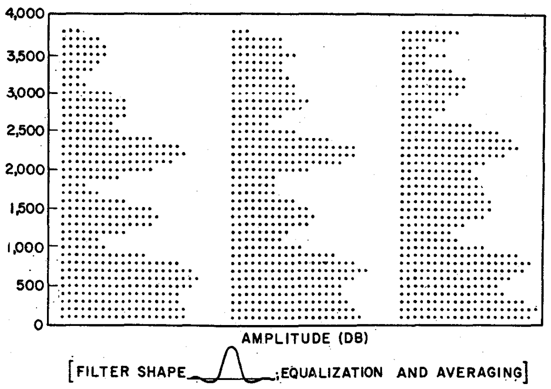

Fig. 10—Spectra produced by modified program.

program was so written that the number of filters, the center frequencies of each filter, the shape of each filter, the averaging time, and the sampling rate could be modified by program parameter cards.

The flexibility of this arrangement was of great advantage in obtaining curves that provide a much-improved display of formant structure. Spectra resulting from the program as originally written are shown in Fig. 9. The filter widths were 200 cps and the weighting function was a rectangle which produced a high side-lobed frequency response. Note that formant structure is masked by the additional contributions of these side lobes. Furthermore, formant position with time is not uniform since there was no time averaging. Finally, the amplitudes decrease with higher frequencies, making the third formant quite low. By contrast, Fig. 10 shows a much-improved display. Here, a cosine weighting function produced a filter frequency response with low side lobes. The formant structure is well defined for this 184-cps width. Time averaging of 20 msec resulted in a smooth flow of formant position with time. High-frequency emphasis of 6 db per octave above 1000 cps yielded a well-defined third formant. The

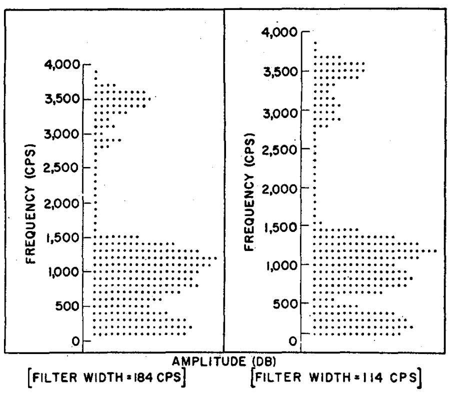

Fig. 11—Spectra compared for two filter widths.

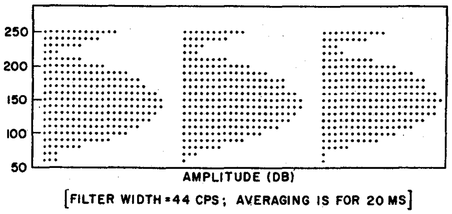

Fig. 12—Use of spectrum program for pitch finding.

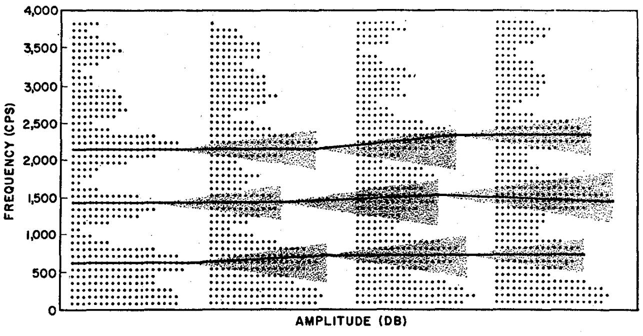

Fig. 13—Formant tracking.

general equations computed by this program are listed in Appendix I.

Further investigations of filter widths were made using this modified program. Fig. 11 compares a spectra obtained with filter widths of 114 cps and 184 cps. The (ꝏ) sound has a first and second formant so closely situated in frequency as to become one broad peak for filter widths of 184 cps. The 114-cps width resolves this broad peak into the two formant peaks.

We are continuing to modify and improve this program. Conceivably, an ultimate filter might be varied dynamically in accordance with the speaker's pitch so as to provide the best display for each speaker. Fig. 12 represents a pitch-finding effort already made. To get the display shown, we used the spectrum-analysis program described to simulate a filter bank of closely-spaced, narrow-width filters.

The first feature program was written to compute the low-order moments of the spectra. This concise mathematical computation (Appendix II) permitted the system of programs to be tested while our input equipment was being constructed. Studies indicated that moments might provide a means of distinguishing the front from back vowels. A sufficient number of samples has yet to be tested.

In contrast to the mathematical statement of the moment program, a formant-tracking program can be defined only after a series of program modifications has been made based on experience of a rudimentary program.

Such a "starting" program is now being written. Fig. 13 depicts the procedure followed in this program. First, a peak is defined. Then, the peaks are examined further to determine what peaks are considered formants. In order to detect a "bar" structure, the peaks must be tracked with time; that is, after a major peak is located in a given spectrum, it must be confirmed in succeeding spectra before it is established as a formant. We can emphasize this by pointing out that only after a threshold of formant duration has been determined through experience, can we ignore spurious indications such as the third peaks appearing here in the second and third spectra. By taking full advantage of the 704, we hope to evolve a highly-detailed, formant-extraction method.

The final stage of our system is the statistical evaluation program. For a collection of measurement values, the statistical evaluation program, as it is now written, can develop the frequency table, can sample mean and standard deviation, cumulative probability function, and probability density function. The term evaluation will have more meaning as experience with this program grows. For example, when we find ourselves consistently performing further data reduction, and routine evaluation tasks, then these tasks should be inserted in the statistical program.

Conclusion

I have assumed, here, the role of correspondent, reporting the result of the highly cooperative effort of my associates, Messrs. Welch, Wimpress, and Wilser. During the past year, we have built the equipment and written a first system of programs. This effort has been supported in part by the Office of Naval Research.

It is our belief that this system of analysis will provide an efficient means of experimental study. Through this study we hope to contribute to the understanding of speech.

Appendix I

The spectrum-analysis program solves the following:

$$A_i = 10 \log_{10} \left\{ \frac{E_i}{N} \sum_{j=1}^{N} \left[\left(\sum_{k=1}^{S} \sigma_k W_k \sin \frac{2\pi k f_i}{R} \right)^2 + \left(\sum_{k=1}^{S} \sigma_k W_k \cos \frac{2\pi k f_i}{R} \right)^2 \right] \right\}.$$

The power-frequency characteristics are determined in this equation through the simulation of the power outputs from a bank of band-pass filters. The weighting functions of the filters are products of the function, W_k, which determines the filter shape, and the sinusoid, sin $(2_\pi k f_i)/R$, which determines the location of the pass band. S is the segment, or summation, interval in input-time samples, and σ_k is the ac amplitude of the kth input-time sample of the acoustica signal. The filter power output, A_i, for the ith frequency, f_i, is averaged for N segments of speech, which corresponds to approximately two pitch periods (20 msec). These coefficients A_i are further modified by E_i, the frequency equalization factor. R is the sampling rate in samples per second. The equations used to compute the two filter shapes discussed are:

Shape	W_k	*Width**
Rectangular	1	$\frac{R}{2S} \sim f \cdot$
Cosine	$1/2\,[1 + \cos(2\pi k/S - \pi)]$	$\frac{R}{S} \sim f$

* Width between 3-db points of main lobe.

Appendix II

The moment program solves the following:

$$M_p = \frac{\sum_{i=1}^{N} (Q_i - Q_0)^p A_i}{\sum_{i=1}^{N} A_i},$$

where M_p is the pth moment about point Q_o in a spectrum of Q equidistant coefficients, A_i, located at points N_i.

Discussion

M. Martin (General Electric Co.): What is the sampling rate used? How many samples are used with each filter to determine the frequency spectrum?

Mr. Shultz: The audio is sampled at the required 727-tape character rate of 16 kc. Since the two audio tape systems each have two speeds, *effective* sampling rates of 8, 16, and 32 kc can be achieved. The number of samples depends on sampling rate and filter width. For the 44-cycle filter for extracting pitch and for a sampling rate of 8 msec, 160 samples are required.

J. R. Barley (Du Pont): Can you transfer the numbers 0 to 9 to digital tape?

Mr. Shultz: Yes. Speech events from 30 msec to 5 seconds can be edited from continuous speech. A microphone input has been provided on the CODER to allow direct conversion of speech to digital tape.

L. S. Bearce (U.S. Naval Research Lab., Washington, D.C.): Would you comment in regard to the feasibility and practicality of a speech recognition system that would operate in real time?

Using your program in the 704, how much increase, if any, in computing speed and complexity do you think will be necessary?

Mr. Shultz: A bank of analog filters would operate in real time. Once we have completely investigated and specified a filter bank through this program, then it might be desirable to shorten our analysis time by building a bank of analog filters.

The spectrum analysis program produces spectra at a rate of 200 times real time. For example, sound of 200-msec duration would require 40 seconds of computation. This delay is due largely to serial computation of the power coefficient of each frequency.