Automatic Recognition of Speech*

THOMAS MARILL†, MEMBER, IRE

Summary—**Research in the field of automatic speech recognition is reviewed. Despite the considerable effort which has been devoted to this field, the results are still quite limited.**

The devices reviewed have achieved only very small recognition vocabularies (the ten spoken digits, for example). Those devices which do not make use of a computer are further restricted to the recognition of the speech of a particular speaker, for whose voice they have been specially adjusted. This limitation is being overcome by the more complex equipment which incorporates a digital computer.

Introduction

THE ultimate goal of workers in the area of automatic speech recognition is the construction of a device which, in effect, replaces a secretary taking dictation: the desired machine is to produce accurate printed copy from the speech of anyone with a reasonable command of the language.

A machine of this kind is desirable for a number of reasons. In the first place, an automatic speech recognizer would represent the major component of a system that reached nearly the ultimate in speech bandwidth compression. A high-fidelity speech communication system is an expensive commodity. Faithful reproduction of speech demands a communication channel having a bandwidth of some 5000 cps and a signal-to-noise ratio of at least 30 db. In terms of amount of information, such a communication channel has a capacity of about 50,000 bits per second. On the other hand, if the speech could be "recognized," *i.e.*, transformed into a string of discrete letters, and if codes for these letters were then transmitted, the rate of information transmittal would be of the order of 50 bits per second or less [3]. An automatic speech recognizer operating ahead of the transmitter could, therefore, "compress" the information by approximately three orders of magnitude. To be sure, some of the information in the speech wave pertaining to the identify, mood, sex, and perhaps other attributes of the speaker [10] would be lost. For many purposes, however, this loss does not present a serious disadvantage.

If it is desired that the communication system produce speech rather than printed text as its output, one may envisage hooking the receiver to a speech synthesizer which performs an approximation to the inverse of the operation of the recognizer. Such synthesizers have been studied; they will not, however, be discussed in the present paper.

* Received by the PGHFE, October 15, 1960. This work was performed under contract USAF 19(602)-2235 with Rome Air Dev. Ctr., Griffiss AFB, Rome, N. Y.

† Bolt, Beranek and Newman, Inc., Cambridge, Mass.

The construction of an automatic speech recognizer would represent an important step in achieving an efficient, natural communication link between men and machines [11]. Man-machine communications have tended to create serious bottlenecks, particularly in large military systems involving digital computers; in such systems, the operator is faced with myriads of insertion switches, knobs, light guns, joy sticks, etc., which he must manipulate in order to influence the behavior of the system. A natural mode of communication, made possible by a speech input, would greatly alleviate this condition.

Much effort is, of course, expended on merely transcribing the spoken word to written form, and machine assistance in this field would be welcome. It is difficult, however, to foresee in the near future an automatic speech recognizer which would successfully compete with a secretary in regards to cost, versatility, and charm.

Aside from practical considerations, the design of a speech recognizer is interesting from the theoretical standpoint. Along the road to its construction, we are forced to come to grips with many important questions relating to the operation of the human nervous system, or, at any rate, relating to the operation of devices that perform functions hitherto performed only by the human nervous system. The problem of automatic speech recognition represents a major challenge currently facing the young discipline of Artificial Intelligence. The solution of this problem, or contributions toward its solution, cannot fail to enrich the discipline as a whole.

In the next section of this paper, we discuss three principal concepts underlying the construction of such machines for automatic speech recognition as have been built to date. In the section following that, we examine briefly those machines which have been built and exhibited.

As we shall see, the achievements to date have been modest. In a final section, we discuss some of the major difficulties in the way of achieving a workable automatic speech recognizer.

Three Fundamental Concepts

Intensity-Frequency-Time Analysis

The sound pressure wave-form measured at the ear of the listener unambiguously defines the acoustic stimulus. This representation, however, has not proved to be particularly apt for the study of speech. If the sound is represented instead in terms of the three dimensions of time, frequency, and amplitude, we obtain a more interesting analysis. Such a representation can be obtained with a device called the *sound spectrograph* [17], which per-

forms an analysis of the sound by passing it through a bank of band-pass filters and determining at each moment of time how much energy there is in each of the bands. (This kind of process is well known today, and no further description of the techniques is given here.) There are at least two reasons for believing that the kind of representation provided by the sound spectrograph is a natural one for use with speech sounds: 1) People can be taught to "read" spectrograph records with fair accuracy, indicating that the information necessary for speech recognition is contained in the records in a form which is decodable by the human nervous system. 2) There is reason to believe that this kind of frequency-time-intensity analysis is also performed by the human auditory system prior to the processing by the central nervous system [12]. It is for these reasons that most attempts at making automatic speech recognizers start by transforming the sound-pressure waveform into a time-intensity-frequency representation by means of a bank of band-pass filters, with smoothing on the output of each channel to give a slowly varying voltage proportional to the power in that channel.

When we attempt to recognize speech visually from spectrograph records, we quickly learn to pay particular attention to the heavy black portions of the record, which indicate the regions along the frequency scale that contain the maximum energy concentration. These regions, called *formants*, appear to have characteristic interrelations in various speech sounds, these interrelations being more or less invariant for different speakers. In particular, it has been found that the position of the second formant, measured in relation to that of the first formant, is highly indicative of the nature of the speech sound, especially in the case of vowels [16].

Articulatory Analysis

A second fundamental concept stems from investigations of how speech is generated in the human vocal apparatus. Speech sounds are produced by exciting in one of several ways what is essentially a modifiable system of cavities. The particular nature of the sounds may be characterized by stating the kind of excitation and some of the main facts relating to the system of cavities at the time of excitation. Thus, by specifying some dozen or so parameters, we may specify the important facts about the state of the vocal apparatus at the time the sound is generated. The important aspects of speech may, therefore, be described by the time variation in these parameters [21].

Linguistic Analysis

While the first of our three fundamental concepts dealt with acoustical facts, and the second dealt with physiological facts underlying the production of speech sounds, the third derives from more abstract linguistic considerations. It has been argued that each phoneme (roughly speaking, each meaningful speech sound) may be uniquely characterized by a small set of "distinctive features" [8]. Phoneme x is different from phoneme y in that the former possesses certain distinctive features which are absent from the latter. Since each of the properties or distinctive features is either present or absent, the phonemes may be represented as an encoding of sequences of binary decisions. A phoneme may, for example, be characterized by being voiced or unvoiced, nasal or non-nasal, etc.[1]

Devices Realized in Hardware

C. P. Smith

One of the earliest attempts to build hardware for automatic speech recognition was made in 1951 by C. P. Smith [20], based on an earlier suggestion of Licklider's [13]. According to Smith, Licklider's idea was as follows: Produce spectograph patterns of all the sounds to be recognized and place each over the face of a cathode-ray tube. Now, generate electronically, on all of the CRT's in parallel, the spectograph pattern of the sound currently to be recognized. Measure with a photocell the amount of light transmitted from each CRT. The one having the greatest output corresponds to the one showing the greatest degree of correlation between the mask and the sound currently under analysis. Smith's device is "perhaps only a sophisticated version of the system proposed by Licklider." The device does not use cathode-ray tubes, nor photocells, but instead passes the input speech through a bank of 32 band-pass filters. Each band-pass filter-channel terminates in a full-wave rectifier and a smoothing circuit with 1/50th-second time constant. The 32 signals are then sent to a switch selector panel "by means of which signals chosen from the total set can be added, subtracted, or ignored by positioning the switches, thus specifying the pattern of energy distribution which is to be detected" [20]. Thus, a distinctive output can be obtained from the switch panel whenever the input speech exhibits energy concentrations in distinctive regions of the spectrum. Auxiliary circuits for determining the fundamental pitch of the voice and for determining the voiced or unvoiced character of the sound were also provided. Although the circuits described were of an experimental nature and the results obtained were tentative, Smith's device did achieve rudimentary distinctions between speech sounds.

Davis, Biddulph and Balashek

Another recognition device was reported to the conference of speech analysis held at the Massachusetts Institute of Technology, Cambridge, in June, 1952, by Davis, Biddulph and Balashek of the Bell Telephone

[1] While we have distinguished between the concepts of intensity-frequency-time (acoustic) analysis, of articulatory analysis, and of linguistic analysis, it must be emphasized that they are intimately related. Thus, the distinctive features may be given definition in terms of articulatory position, and the latter may be studied in relation to the acoustic intensity-frequency-time analysis.

Laboratories [1]. The device described was able to recognize the ten spoken digits with approximately 98 per cent accuracy, provided these were spoken by the person for whose speech the device had previously been adjusted. For others than "his master's voice" the device failed.

The operation of the device is as follows. The speech signal is first separated into two frequency bands, the lower band containing the frequencies up to 900 cps and the higher band containing the frequencies above 900 cps. In each band the signals are amplitude-limited; the frequency changes in each band are "tracked" by axis-crossing counters. Two quantities are thus generated which carry information about the location, in the frequency domain, of the regions of high-energy concentration in the two bands. Roughly, these two quantities represent, as a function of time, the frequency of the first two formants of the speech. The two quantities are made to generate a trace on a CRT, the x deflection being provided by the quantity derived from the low-frequency band, the y deflector by the quantity derived from the high-frequency band. During the speaking of a word, the trace moves about the scope in a characteristic fashion.

The face of the scope is partitioned into 30 squares by selecting six intervals along the x axis and five along the y axis, though only 28 of the 30 squares are actually used. During the speaking of a word, the trace spends various amounts of time in each of the 28 squares. These 28 durations are determined; each spoken word, then, yields a vector of 28 numbers. Representative vectors are determined for each of the ten spoken digits; these are stored as charges on condensers. When a new word is to be identified, its vector is generated and correlated with the ten stored vectors. The highest correlation yields the decision as to which digit was spoken.

Olson and Belar

In 1956, Olson and Belar described an experimental "phonetic typewriter" [15]. This device achieved a vocabulary of ten words: "are, see, a, I, can, you, read, it, so, sir." When sentences were fashioned out of these words and spoken into the machine by the talker for whom the machine had previously been adjusted, the words were typed out with 98 per cent accuracy. It is mentioned that care had to be exercised to insure clear enunciation. The recognition was performed essentially one syllable at a time, with pauses between the syllables.

The system operates as follows. The sound to be recognized is assumed to be about 0.2 seconds long, and a time-intensity-frequency plot (quantized) of the sound is formed. Frequency is quantized into eight adjacent bands, time into five consecutive periods, and intensity into two levels. The resultant time-intensity-frequency plot may be thought of as an eight-by-five matrix of cells, each cell containing a zero (energy below the threshold), or a one (energy above the threshold). The information in this matrix is then fed into a decoder, which has been preset to a given person's speech in such a way that a given pattern in the matrix corresponding, *e.g.*, to the word "can," will cause the typewriter to type out this word. Relay storage devices were used to activate the proper keys on the typewriter, once a given syllable had been detected.

Wiren and Stubbs

The following attempt is based on the distinctive feature theory of Jacobson, Fant and Halle [8]. While this attempt has not, to date, been successful, it is nonetheless worth mentioning here.

In 1957, Wiren and Stubbs reported some preliminary results in their attempt to classify speech sounds by testing for the presence or absence in each sound of properties that identify distinctive features [22]. They succeeded in building more or less independent circuits that were reasonably successful in testing for the following properties: voiced-unvoiced, turbulent-nonturbulent, stop-fricative, acute-grave. Considerably more work would, of course, have to be done along this line before a practical speech recognizer, even with a small vocabulary of sounds, could be realized.

Fry and Denes

A device developed by the British scientists Fry and Denes [5], [6] has a repertoire of 14 sounds (four vowels, nine consonants and "space"). The machine will give an indication, at each moment of time, of its best guess as to which of the 14 sounds is currently being spoken. Thus, for a language consisting only of these 14 sounds (approximately 35 per cent of the necessary repertoire for natural English), the machine is capable of real-time phonetic transcription of speech. When used by the talker for whom the machine was adjusted, it was found that the recognition of the 14 sounds proceeded with approximately 72 per cent accuracy. For connected discourse, using an artificial language consisting of these 14 sounds only, the word recognition accuracy was 44 per cent.

The machine operates in two stages. In the first stage, the acoustic input is analyzed, and a best guess among the 14 sounds is picked. The second stage makes use of linguistic knowledge. Stored within this stage are the digram frequencies, that is, the relative frequencies of occurrence of successive pairs of sounds as they occur in the artificial language. This second stage generates a guess as to the currently spoken sound on the basis of knowledge of the previously recognized sound, and of knowledge that this sound has, in the past, most frequently been followed by some other given sound. Then, finally, using the results of the acoustic analysis of the first stage and of the probability analysis of the second stage, an over-all best guess is generated. The acoustic analysis in the first stage is performed by passing the speech through a bank of 20 band-pass filters, the outputs of which are rectified and smoothed.

> The occurrence of a given sound in a speech input is found to give rise to a maximum of energy in a particular pair of filters. . . . The outputs of two filters in a pair are multiplied together and this gives 14 voltage products [5].

The greatest of these 14 products is then selected and yields the decision of the first stage. A new sound is assumed to be present whenever the maximum of the voltage products shifts from one of the 14 categories to another.

Dudley and Balashek

A new scheme for the recognition of spoken digits was discussed in 1958 by Dudley and Balashek [2]. This device was able to recognize the ten digits with high reliability when spoken by the person for whom the device was originally adjusted. As was the case with the previously discussed digit recognizer, the device works badly with other than "his master's voice."

The device works in two stages, a phoneme-recognition stage, and a word-recognition stage. In the first stage, the device yields, at each moment of time, a best guess as to which of ten speech sounds (six vowels and four consonants) is presently being spoken. This is achieved by passing the speech through a bank of ten band-pass filters, the output of which is rectified and smoothed, and which goes into one side of a ten-by-ten resistance matrix. The output from the other side of the resistance matrix is, in effect, a set of ten different weighted sums of the original ten input voltages. The weights are so chosen that the first weighted sum will be, on the average, the greatest among the ten weighted sums whenever the first of the ten speech sounds is actually present; and similarly for each of the other nine. Thus, the greatest of the ten outputs from the resistance matrix is picked and yields the best guess, regarding the ten sounds in the repertoire, at each moment of time. From the output of the first stage, the second stage computes the amount of time, relative to the entire spoken word, during which the best guess was yielded for sound number one, sound number two, etc. The second stage, in other words, forms the distribution over time of the various best guesses that occurred during the spoken word. Since there are ten possible sounds, this distribution is represented by ten quantities. These ten quantities then feed into a second resistance matrix which forms ten different weighted sums, the weights being so chosen that sum number five, say, will be, on the average, the highest when the spoken digit was the numeral five.

Speech Recognition by Computers

It is only within the last two or three years that investigators in this field have turned to the use of digital computers. In particular, three projects should be mentioned,[2] that of Forgie and Forgie of the Lincoln Laboratory [4], that of Shultz of IBM [19], and that of Stevens of the Massachusetts Institute of Technology [21]. These three projects represent fairly ambitious undertakings; however, being new, the results are as yet fragmentary.

All three of the contemplated techniques are similar in that they start by analyzing the speech through a bank of band-pass filters, the output of which is rectified, smoothed, and sampled at periodic intervals. Forgie and Forgie use 35 channels, each one of which is sampled 180 times per second. Stevens uses 36 channels sampled 120 times per second. In the Shultz system, the filters are not realized in hardware, but are instead simulated by means of the computer program itself. Thus, the number of filters and the sampling rate is variable.

The program of Forgie and Forgie has so far achieved the recognition of ten English vowels when these vowels are imbedded in words of the form b-vowel-t. The recognition accuracy was 93 per cent for the speech of eleven male and ten female speakers, *with no adjustment required for the individual speakers.* The technique is based essentially on the fact, discussed previously, that the relative position of the first two formants is indicative of the particular vowel spoken. However, additional information is also used to make the discrimination: the fundamental pitch of the voice, the position of the third formant, and certain additional facts about the spectrum are taken into consideration. The method of Forgie and Forgie, as reported so far, represents the implementation on the digital computer of fairly standard, well-understood principles in speech recognition.

Shultz' program has achieved the recognition of spoken digits with an accuracy of 97 per cent when the digits were spoken by 50 speakers (25 males and 25 females), *with no adjustment required for individual speakers.* The exact method for doing this has not been disclosed. It appears, however, that rather less reliance is placed on the known properties of speech, and rather more on sophisticated statistical decision procedures.

The philosophy underlying the work of Stevens' group is based on the articulatory analysis of speech, that is, on analysis in terms of the dozen or so parameters which characterize the state of the speech-producing mechanism at the time the sound is generated. The method advocated for determining the values of these parameters is known as "analysis by snythesis." In this procedure, we generate the spectrum of a speech sound by means of a model for speech production controlled by a dozen parameters. The synthetic spectrum thus obtained is then compared with the spectrum of the speech to be recognized; if the match is not satisfactory, the parameters are readjusted to create a better match. After several iterations of this process, very good matches can be achieved in many cases. The system has so far been proved capable of handling the analysis of approximately a dozen vowels from the speech of various speakers.

Some Major Difficulties

As we have seen, the achievements in the field of automatic recognition of speech, despite considerable efforts, are as yet quite modest. The general problem has proved

[2] A fourth project is that of Hughes and Halle [7]. Since these authors do not yet report any actual data on recognition, their project has not been discussed in the present paper.

to be remarkable refractory. About the best that has been done so far is the recognition of the spoken digits and of some of the vowels in isolated context.

The difficulties arise in the following four areas: 1) the problem of segmentation, 2) the choice of proper measurements, 3) suitable statistical decision procedures, and 4) difficulties of handling large information-processing tasks in real time. Concerning the last of these, very little can be said within the confines of the present paper; therefore, let us address ourselves to the other three.

Speech is essentially continuous; it does not have readily distinguished boundaries between segments. It is generally agreed that in order to analyze speech, it is first necessary to break up the continuous stream into separate segments. It is conceivable that one would wish to recognize in units of words or perhaps even phrases, but to date the most popular approach has been to use the phoneme as the basic unit. Even the validity of segmentation in terms of phonemes has been questioned, however. According to Ladefoged [9],

> Many sounds have clear cut boundaries; but some sequences (*e. g.*, that in the word "trying") can be segmented only by *ad hoc* judgments which cannot be stated in the form of rigorous procedures that are always applicable.

The second fundamental difficulty is that of picking an adequate set of "measurements" of the acoustic segment to be recognized. In attempting to recognize a given segment, it would be inordinately cumbersome to have to deal with the actual wave-form of that segment (or a quantization of this wave-form in terms of 50,000 bits per second). The attempt is rather to select a small, judiciously chosen set of measurements of the wave-form, basing the recognition procedure on the results of these measurements.

There is no question, for example, that it is extremely useful, for the recognition of vowels, to measure the position on the frequency scale of the first few formants. Indeed, a good portion of the job of recognizing vowels has been done right there and then. The discovery of such highly informative measurements is probably the most critical step in the construction of automatic speech recognizers. The model of Stevens [21] represents, today, the closest approach to an adequate framework within which we may look for guidance in the development of such measurements. A more adequate psychological theory of speech perception than exists at the moment would also help to provide such a framework.

Assuming we had solved the segmentation problem and the problem of selecting adequate measurements on which to base our recognitions, we would still be faced with the question of how to combine these various measments, which may be highly interdependent, in such a way as to issue forth with a best decision as to the proper classification of the segment. This problem is one in the area of decision making. The statistical approach to this problem, for recognition processes in general, has been discussed by Marill and Green [14]. It is possible that as we become more proficient in the area of statistical decision making, we shall be able to relax somewhat our demands for highly discriminating measurements. The application of sophisticated principles of parallel processing and of adaptive machines would unquestionably be of considerable use here [18]. As yet, however, our understanding of valid procedures by means of which we may base reasonable conclusions on many various pieces of information, each of which by itself may be inconclusive, is still very rudimentary.

Bibliography

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recogntion of spoken digits," *J. Acoust. Soc. Am.*, vol. 24, pp. 637–642; November, 1952.

[2] H. Dudley, and S. Balashek, "Automatic recognition of phonetic patterns of speech," *J. Acoust. Soc. Am.*, vol. 30, pp. 721–732; August, 1958.

[3] G. Fant, and K. N. Stevens, "Systems for speech compression," in "Fortschritte der Hochfreguenztechnik," Bd. 5, Friedrich Rühmann, Karlsruhe-Durlach, Germany; 1961.

[4] J. W. Forgie, and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Am.*, vol. 31, pp. 1480–1489; November, 1959.

[5] D. B. Fry, and P. Denes, "The solution of some fundamental problems in mechanical speech recognition," *Language and Speech*, vol. 1, pp. 35–38; 1958.

[6] D. B. Fry, and P. Denes, "An analogue of the speech recognition process," *Proc. Symp. on Mechanisation of Thought Processes*, Natl. Phys. Lab., Teddington, Eng., Her Majesty's Stationery Office, London; 1959.

[7] G. W. Hughes, and M. Halle, "On the recognition of speech by machine," *Proc. Internatl. Conf. on Information Processing*, UNESCO, Paris, France; June, 1959.

[8] R. Jakobson, C. G. M. Fant, and M. Halle, "Preliminaries to Speech Analysis," Acoustics Lab., Mass. Inst. Tech., Cambridge, Tech. Rept. No. 13; 1952.

[9] P. Ladefoged, "The perception of speech," *Proc. Symp. on Mechanisation of Thought Processes*, Natl. Phys. Lab., Teddington, Eng., Her Majesty's Stationary Office, London; 1959.

[10] J. C. R. Licklider, and G. A. Miller, "The perception of speech," in "Handbook of Experimental Psychology," S. S. Stevens Ed., John Wiley and Sons, Inc., New York, N. Y.; 1951.

[11] J. C. R. Licklider, "Man-computer symbiosis," IRE Trans. on Human Factors in Electronics, vol. HFE-1, pp. 4–11; March, 1960.

[12] J. C. R. Licklider, "On the process of speech perception," *J. Acoust. Soc. Am.*, vol. 24, pp. 590–594; November, 1952.

[13] J. C. R. Licklider, "Compression of information in voice communication," unpublished memo.; 1949.

[14] T. Marill, and D. M. Green, "Statistical recognition functions and the design of pattern recognizers," IRE Trans. on Electronic Computers, vol. EC-9, pp. 472–477; December, 1960.

[15] H. F. Olson, and H. Belar, "Phonetic Typewriter," *J. Acoust. Soc. Am.*, vol. 28, pp. 1072–1081; November, 1956.

[16] G. E. Peterson, and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, pp. 175–184; March, 1952.

[17] R. K. Potter, G. A. Kopp, and H. C. Green, "Visible Speech," D. Van Nostrand Co., Inc., New York, N. Y.; 1947.

[18] O. G. Selfridge, "Pandemonium: a paradigm for learning," *Proc. Symp. on Mechanisation of Thought Processes*, Natl. Phy. Lab., Teddington, Eng., Her Majesty's Stationery Office, London; 1959.

[19] G. L. Shultz, "Investigation Procedures for speech recognition," *Proc. Seminar on Speech Compression and Processing*, AF Cambridge Res. Ctr., Bedford, Mass., Tech. Rept. No. 198; September, 1959.

[20] C. P. Smith, "A phoneme detector," *J. Acoust. Soc. Am.*, vol. 23, pp. 446–451; July, 1951.

[21] K. N. Stevens, "Toward a model for speech recognition," *J. Acoust. Soc. Am.*, vol. 32, pp. 47–51; January, 1960.

[22] J. Wiren, and H. L. Stubbs, "Electronic binary system for phoneme classification," *J. Acoust. Soc. Am.*, vol. 28, pp. 1082–1091; November, 1956.