Speech Recognition: A Model and a Program for Research*

M. HALLE† AND K. STEVENS‡, MEMBER, IRE

Summary—**A speech recognition model is proposed in which the transformation from an input speech signal into a sequence of phonemes is carried out largely through an active or feedback process. In this process, patterns are generated internally in the analyzer according to an adaptable sequence of instructions until a best match with the input signal is obtained. Details of the process are given, and the areas where further research is needed are indicated.**

THE FUNDAMENTAL problem in pattern recognition is the search for a *recognition function* that will appropriately pair *signals* and *messages*. The input to the recognizer generally consists of measured physical quantities characterizing each signal to be recognized, while at the output of the recognizer each input signal is assigned to one of a number of categories which constitute the messages. Thus, for instance, in machine translation, the signals are sentences in one language and the messages are sentences in another language. In the automatic recognition of handwriting, the signal is a two-dimensional curve and the message a sequence of letters in a standard alphabet. Similarly, research on automatic speech recognition aims at discovering a recognition function that relates acoustic signals produced by the human vocal tract in speaking to messages consisting of strings of symbols, the phonemes. Such a recognition function is the inverse of a function that describes the production of speech, *i.e.*, the transformation of a discrete phoneme sequence into an acoustic signal.

This paper proposes a recognition model in which mapping from signal to message space is accomplished largely through an active or feedback process. Patterns are generated internally in the analyzer according to a flexible or adaptable sequence of instructions until a best match with the input signal is obtained. Since the analysis is achieved through active internal synthesis of of comparison signals, the procedure has been called "analysis by synthesis."[1]

The Process of Speech Production

In line with the traditional account of speech production, we shall assume that the speaker has stored in his memory a table of all the phonemes and their actualizations. This table lists the different vocal-tract configurations or gestures that are associated with each phoneme and the conditions under which each is to be used. In producing an utterance the speaker looks up, as it were, in the table the individual phonemes and then instructs his vocal tract to assume in succession the configurations or gestures corresponding to the phonemes.

The shape of man's vocal tract is not controlled as a single unit; rather, separate control is exercised over various gross structures in the tract, *e.g.*, the lip opening, position of velum, tongue position, and vocal-cord vibration. The changing configurations of the vocal tract must, therefore, be specified in terms of parameters describing the behavior of these quasi-independent structures.[2] These parameters will be called *phonetic parameters*.[3]

Since the vocal tract does not utilize the same amount of time for actualizing each phoneme (*e.g.*, the vowel in *bit* is considerably shorter than that in *beat*), it must be assumed that stored in the speaker's memory there is also a schedule that determines the time at which the

* Received by the PGIT, September 25, 1961; revised manuscript received, October 12, 1961. This work was supported in part by the U. S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research; in part by the National Science Foundation; and in part by the Air Force (Electronic Systems Division) under Contract AF 19(604)-6102.

† Research Laboratory of Electronics and Department of Modern Languages, Massachusetts Institute of Technology, Cambridge, Mass.

‡ Research Laboratory of Electronics and Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[1] The relevance of such analysis procedures to more general perceptual processes has been suggested by several writers. See, for example:

D. M. MacKay, "Mindlike behavior in artefacts," *Brit. J. for Philosophy of Science*, vol. 2, pp. 105–121; 1951.

G. A. Miller, E. Galanter, and K. H. Pribram, "Plans and the Structure of Behavior," Henry Holt and Co., New York, N. Y.; 1960.

M. Halle and K. N. Stevens, "Analysis by synthesis," *Proc. of Seminar on Speech Compression and Processing*, W. Wathen-Dunn and L. E. Woods, Eds., vol. 2, Paper D7: December, 1959.

[2] This view was well understood by the founder of modern phonetics, A. M. Bell, who described utterances by means of symbols ("Visible Speech and The Science of Universal Alphabetics," Simpkin, Marshall and Co., London, Eng.; 1867) from which the behavior of the quasi-independent structures could be read off directly. The subsequent replacement, for reasons of typographical economy, of Bell's special symbols by the Romic of the Internatl. Phonetic Assoc. has served to obscure the above facts and to suggest that phonemes are implemented by controlling the vocal tract as a single unit.

[3] We cannot discuss in detail at this point the nature of the phonetic parameters, and we do not take sides here in the present discussion between proponents of the Jakobsonian distinctive features (R. Jakobson and M. Halle, "Fundamentals of Language," Mouton and Co., The Hague, The Netherlands; 1956) and those of more traditional views ("The Principles of the International Phonetic Association," University College, London, England; 1949). We insist however, that the control of the vocal-tract behavior must be described by specifying a set of quasi-independent phoentic parameters.

vocal tract moves from one configuration to the next, *i.e.*, the time at which one or more phonetic parameters change in value. The timing will evidently differ depending on the speed of utterance—it will be slower for slower speech and faster for faster speech.

Because of the inertia of the structures that form the vocal tract and the limitations in the speed of neural and muscular control, a given phonetic parameter cannot change instantaneously from one value to another; the transitions from one target configuration to the next must be gradual, or smooth. Furthermore, when utterances are produced at any but the slowest rates, a given articulatory configuration may not be reached before motion toward the next must be initiated. Thus the configuration at any given time may be the result of instructions from more than one phoneme. In other words, at this stage in the speech production process, discrete quantities found in the input have been replaced by continuous parameters. A given sequence of phonemes, moreover, may produce a variety of vocal-tract behaviors depending upon such factors as the past linguistic experience of the talker, his emotional state, and the rate of talking.

The continuous phonetic parameters that result from a given phoneme sequence give rise in turn to changes in the geometry and acoustic excitation of the cavities forming the vocal tract. The tract can be visualized as a time-varying linear acoustic system, excited by one or more sound sources, which radiates sound from the mouth opening (and/or from the nose). The acoustic performance of this linear system at a given time and for a given source of excitation can be characterized by the poles and zeros of the transfer function from the source to the output, together with a constant factor.[4] For voiced sounds the vocal tract is excited at the glottis by a quasi-periodic source with high acoustic impedance. Its fundamental frequency varies with time, but the waveform or spectrum of each glottal pulse does not change markedly from one speech sound to another. In addition, the vocal tract may be excited in the vicinity of a constriction or obstruction by a broad-band noise source or by sound.

In the process of generating an acoustic output in response to a sequence of phonemes, a talker strives to produce the appropriate vocal-tract configurations together with the proper type of source, but he does not exert precise control over such factors as the detailed characteristics of the source or the damping of the vocal tract. Consequently, for a given vocal-tract configuration the shape of the source spectrum, the fundamental frequency of the glottal source, and the bandwidths of the poles and zeros can be expected to exhibit some variation for a given talker. Even greater variation is to be expected among different talkers, since the dimensions of the speech-production apparatus are different for different individuals. This variance is superimposed on the already-mentioned variance in articulatory gestures.

Reduction of the Continuous Signal to a Message Consisting of Discrete Symbols; the Segmentation Problem

The analysis procedure that has enjoyed the widest acceptance postulates that the listener first segments the utterance and then identifies the individual segments with particular phonemes. No analysis scheme based on this principle has ever been successfully implemented. This failure is understandable in the light of the preceding account of speech production, where it was observed that segments of an utterance do not in general stand in a one-to-one relation with the phonemes. The problem, therefore, is to devise a procedure which will transform the continuously-changing speech signal into a discrete output without depending crucially on segmentation.

A simple procedure of this type restricts the input to stretches of sound separated from adjacent stretches by silence. The input signals could, for example, correspond to isolated words, or they could be longer utterances. Perhaps the crudest device capable of transforming such an input into phoneme sequences would be a "dictionary" in which the inputs are entered as intensity-frequency-time patterns[5] and each entry is provided with its phonemic representation. The segment under analysis is compared with each entry in the dictionary, the one most closely resembling the input determined, and its phonemic transcription printed out.[6]

The size of the dictionary in such an analyzer increases very rapidly with the number of admissible outputs, since a given phoneme sequence can give rise to a large number of distinct acoustic outputs. In a device whose capabilities would even remotely approach those of a normal human listener, the size of the dictionary would, therefore, be so large as to rule out this approach.[7]

The need for a large dictionary can be overcome if the principles of construction of the dictionary entries are

[4] G. Fant, "Acoustic Theory of Speech Production," Mouton and Co., The Hague, Neth.; 1960.

[5] The initial step in processing a speech signal for automatic analysis usually consists of deriving from the time-varying pressure changes a sequence of short-time amplitude spectra. This transformation, which is commonly performed by sampling the rectified and smoothed outputs of a set of band-pass filters or by computing the Fourier transform of segments of the signal, is known to preserve intact the essential information in the signal, provided that suitable filter bandwidths and averaging times have been chosen.

[6] A model of this type was considered by F. S. Cooper, *et al.*, "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Am.*, vol. 24, p. 605; November, 1952.

"The problem of speech perception is then to describe the decoding process either in terms of the decoding mechanism or—as we are trying to do—by compiling the code book, one in which there is one column for acoustic entries and another column for message units, whether these be phonemes, syllables, words, or whatever."

[7] This approach need not be ruled out, however, in specialized applications in which a greatly restricted vocabulary of short utterances, such as digits, is to be recognized. See, for example:

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

P. Denes and M. V. Mathews, "Spoken digit recognition using time-frequency pattern matching," *J. Acoust. Soc. Am.*, vol. 32, pp. 1450–1455; November, 1960.

G. S. Sebestyen, "Recognition of membership in classes," IRE Trans. on Information Theory, vol. IT-6, pp. 44–50; January, 1961.

known. It is then possible to store in the "permanent memory" of the analyzer only the rules for speech production discussed in the previous section. In this model the dictionary is replaced by *generative rules* which can synthesize signals in response to instructions consisting of sequences of phonemes. Analysis is now accomplished by supplying the generative rules with all possible phoneme sequences, systematically running through all one-phoneme sequences, two-phoneme sequences, etc. The internally generated signal which provides the best match with the input signal then identifies the required phoneme sequence. While this model does not place excessive demands on the size of the memory, a very long time is required to achieve positive identification.

The necessity of synthesizing a large number of comparison signals can be eliminated by a *preliminary analysis* which excludes from consideration all but a very small subset of the items which can be produced by the generative rules. The preliminary analysis would no doubt include various transformations which have been found useful in speech analysis, such as segmentation within the utterance according to the type of vocal-tract excitation and tentative identification of segments by special attributes of the signal. Once a list of possible phoneme sequences is established from the preliminary analysis, then the internal signal synthesizer proceeds to generate signals corresponding to each of these sequences.

The analysis procedure can be refined still further by including a *control* component to dictate the order in which comparison signals are to be generated. This control is guided not only by the results of the preliminary analysis but also by quantitative measures of the goodness of fit achieved for comparison signals that have already been synthesized, statistical information concerning the admissible phoneme sequences, and other data that may have been obtained from preceding analyses. This information is utilized by the control component to formulate strategies that would achieve convergence to the required result with as small a number of trials as possible.

It seems to us that an automatic speech recognition scheme capable of processing any but the most trivial classes of utterances must incorporate all of the features discussed above—the input signal must be matched against a comparison signal; a set of generative rules must be stored within the machine; preliminary analysis must be performed; and a strategy must be included to control the order in which internal comparison signals are to be generated. The arrangement of these operations in the proposed recognition model is epitomized in Fig. 1.

Processing of the Speech Signal Prior to Phoneme Identification

In the analysis-by-synthesis procedure just described, it is implied that the comparison between the input and the internally generated signal is made at the level of the time-varying acoustic spectrum. It is clear, however, that the input signal of Fig. 1 could equally well be the result of some transformation of the acoustic spectrum carried out at a previous stage of analysis. Indeed, in any practical speech recognizer, it is essential to subject the spectral pattern to a certain amount of preliminary processing before entering the phonemic analysis stage. The necessity for initial transformations or simplifications stems from the fact that many acoustic signals may correspond to a given sequence of phonemes. To account for all the sources of variance or redundancy in one stage of analysis is much too difficult an undertaking. Through a stepwise reduction procedure, on the other hand, variance due to irrelevant factors can be eliminated a small amount at a time.

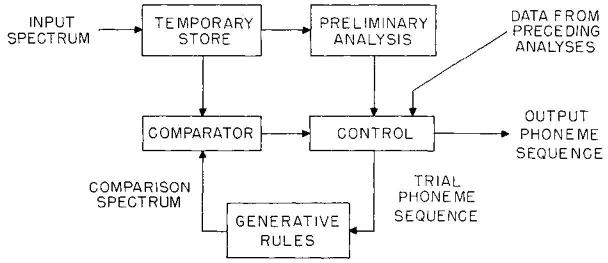

Fig. 1—Block diagram of analysis-by-synthesis procedure for extracting a phoneme sequence from a time-varying input spectrum. The input spectrum, which may be placed in temporary storage pending completion of the analysis, is compared in the comparator with signals synthesized by the generative rules. Instructions as to the phoneme sequences to be tried are communicated to the generative rules by the control component, which bases its decisions on the results of a preliminary analysis of the input signal and on the output of the comparator for previous trials, as well as on other information as noted in the text. When a best match is obtained in the comparator, the control component reads out the phoneme sequence which, through the generative rules, produced that match. This figure also serves to show the arrangement of components in the proposed model for the reduction of speech spectra to continuous phonetic parameters.

The proposed procedure for speech processing contains two major steps. In the first stage the spectral representation is reduced to a set of parameters which describe the pertinent motions and excitations of the vocal tract, *i.e.*, the phonetic parameters. In the second stage, transformation to a sequence of phonemes is achieved. These steps provide a natural division of the analysis procedure into one part concerned primarily with the physical and physiological processes of speech, and the other concerned with those aspects of speech primarily dependent on linguistic and social factors. In the first stage, variance in the signal due to differences in the speech mechanism of different talkers (or of a given talker in different situations) would be largely eliminated. The second stage would account for influences such as rate of talking, linguistic background or dialect of the talker, and contextual variants of phonemes.

Many of the problems involved in the first analysis stage are not unlike those encountered in reducing an utterance to a phoneme sequence. It is not feasible to store all possible spectra together with the corresponding articulatory descriptions. Since, however, rules for generating the spectrum from the articulatory description are known, it is possible to use an analysis-by-synthesis procedure[8] of the type shown in Fig. 1.

The output of this stage is a set of phonetic parameters (rather than the phoneme sequence shown in Fig. 1). The heart of this first-stage analyzer is a signal synthesizer that has the ability to compute comparison spectra when given the phonetic parameters, *i.e.*, an internal synthesizer in which are stored the generative rules for the construction of speech spectra from phonetic parameters. A strategy is required to reduce the time needed to match the input spectrum and the comparison spectrum. The strategy may again depend on the results of a preliminary approximate analysis of the input signal, and on the error that has been computed at the comparator on previous trials. It may also depend on the results that have been obtained for the analysis of signals in the vicinity of the one under direct study. Some of the instructions that are communicated by the control component to the generative rules remain relatively fixed for the matching of spectra generated by a given talker in a given situation. When signals generated by a different talker are presented, the strategy must be able to modify this group of instructions automatically after sufficient data on that talker's speech have been accumulated. The analysis-by-synthesis procedure has the property, therefore, that its strategy is potentially able to adapt to the characteristics of different talkers.

Summary of Model for Speech Recognition

The complete model for speech recognition discussed here takes the form shown in Fig. 2. The input signal is first processed by a peripheral unit such as a spectrum analyzer. It then undergoes reduction in two analysis-by-synthesis loops, and the phoneme sequence appears at the right. In order to simplify the diagram, the group of components performing the functions of storage, preliminary analysis, comparison, and control have been combined in a single block labeled *strategy*.

The procedure depicted here is suitable only for the recognition of sequences of uncorrelated symbols, such as those that control the generation of nonsense syllables. If the speech material to be recognized consists of words, phrases, or continuous text, then the output of the present analysis scheme would have to be processed further to take account of the constraints imposed by the morphological and syntactic structure of the language.

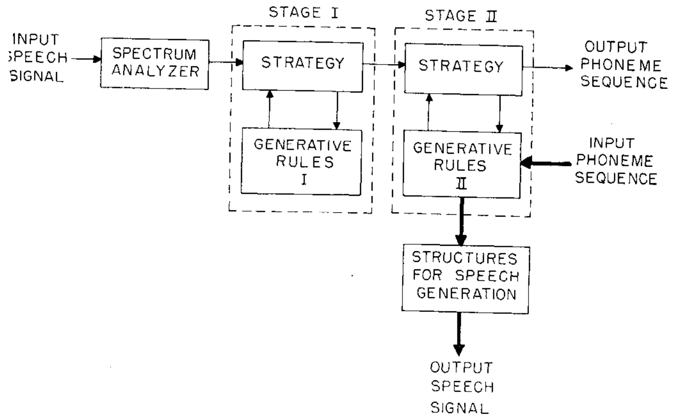

Fig. 2—Block diagram of two-stage scheme for speech processing. Following processing by a spectrum analyzer, the input speech signal is reduced in Stage I to a set of quasi-continuous phonetic parameters, which are processed in Stage II to yield an output phoneme sequence. An analysis-by-synthesis procedure is used for processing the signal at each stage. The heavy lines indicate the operations that are involved in generating a speech signal from a phoneme sequence.

The final analysis stage of Fig. 2 includes, of course, the generative rules for transforming phoneme sequences into phonetic parameters. These are precisely the rules that must be invoked in the production of speech. During speech production the output from these stored rules can be connected directly to the speech mechanism, while the input to the rules is the phoneme sequence to be generated. Addition of peripheral speech-generating structures to Fig. 2 then creates a model that is capable of both speech recognition and speech production. The same calculations are made in the second set of generative rules (and in the generative rules at possible higher levels of analysis) whether speech is being received or generated. It is worthwhile observing that during the recognition process phonetic parameters are merely calculated by the "generative rules II" and direct activation of the speech structures is nowhere required.[9]

For the recognition of continuous speech it may not always be necessary to have recourse to analysis-by-synthesis procedures. A rough preliminary analysis at each of the stages in Fig. 2 may often be all that is required—ambiguities as a result of imprecise analysis at these early stages can be resolved in later stages on the basis of knowledge of the constraints at the morphological, syntactic, and semantic levels.[10]

[8] Partial implementation (or models for implementation) of the analysis-by-synthesis procedure applied at this level, together with discussions of the advantages of the method, have been presented in:

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

L. A. Chistovich, "Classification of rapidly repeated speech sounds," *Sov. Phys. Acoustics*, vol. 6, pp. 393–398; January–March, 1961 (*Akust. Zhur.*, vol. 6, pp. 392–398; July, 1960).

S. Inomata, "Computational method for speech recognition," *Bull. Electro-Tech. Lab.* (Tokyo), vol. 24, pp. 597–611; June, 1960.

M. V. Mathews, J. E. Miller, and E. E. David, Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Am.*, vol. 33, pp. 179–186; February, 1961.

C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Am.*, vol. 33; December, 1961.

[9] This point was discussed by A. M. Liberman ("Results of research on speech perception," *J. Acoust. Soc. Am.*, vol. 29, pp. 117–123; January, 1957) who suggested that speech is perceived with reference to articulation, but that "the reference to articulatory movements and their sensory consequences must somehow occur in the brain without getting out into the periphery."

[10] Knowledge of constraints imposed on phoneme sequences by the structure of the language has been incorporated in the design of an automatic speech recognizer described by Fry and Denes (D. B. Fry, "Theoretical aspects of mechanical speech recognition," and P. Denes, "The design and operation of the mechanical speech recognizer at University College, London," *J. Brit. IRE*, vol. 19, pp. 211–234; April, 1959.

Implementation of the Model: Problems for Research

While certain components in both major stages of analysis can be designed from present knowledge, further research is necessary before the remaining components can be realized and before the system can be designed to function as a whole.

In the first stage of analysis, one of the major problems is to devise a procedure for specifying in quantitative terms the "phonetic parameters." These must describe the behavior of structures that control the vocal-tract configuration as well as activities of the lungs and vocal cords. A great deal is known about some parameters, *e.g.*, parameters that relate to voicing, nasalization, interruptedness, and labialization. For others, such as tenseness or the so-called point of articulation, our knowledge is still far from adequate.

A second task is to establish the generative rules describing the conversion of phonetic parameters to time-varying speech spectra. These rules involve a series of relations, namely, those between 1) the phonetic parameters and the vocal-tract geometry and excitation characteristics, 2) the transformation from vocal-tract geometry to the transfer function in terms of poles and zeros, and 3) the conversion from the pole-zero configurations and pertinent excitation characteristics to the speech spectra. The last two of these, which involve application of the theory of linear distributed systems, have been studied in some detail,[6,11,12] whereas the first is less well understood.

The generative rules of the second stage are made up of several distinct parts. First, they embody the relation between what linguists have called a "narrow phonetic transcription of an utterance" and its "phonemic or morphophonemic transcription." The nature of this relation has received a fair amount of attention in the last 30 years and a great deal of valuable information has been gathered. Of especial importance for the present problems are recent phonological studies in which this relation has been characterized by means of a set of ordered rules.[13] Secondly, the generative rules II must describe the utilization of those phonetic parameters that are not governed by the language in question, but are left to the discretion of the speaker. Thus, for instance, it is well known that in English speech, voiceless stops in word final position may or may not be aspirated. The precise way in which individual speakers utilize this freedom is, however, all but unknown. Thirdly, the generative rules II must specify the transformation from discrete to continuous signals that results from the inertia of the neural and muscular structures involved in speech production. There are wide variations in the delay with which different muscular movements can be executed, but the details of the movements are not understood. The study of these problems, which essentially are those of producing continuous speech from phonetic transcriptions, has just begun in earnest. We owe important information to the work of Haskins Laboratory on simplified rules for speech synthesis.[14] This work must now be extended to take physiological factors into consideration more directly, through the use of cineradiography,[15] electromyography, and other techniques. Contributions can also be expected from studies with dynamic analogs of the vocal tract.[16]

Finally, for both stages of analysis, the design of the strategy component is almost completely unknown territory. To get a clearer picture of the nature of the strategy component, it is useful to regard the generative rules as a set of axioms, and the outputs of the generative rules as the theorems that are consequences of these axioms. Viewed in this light the discovery of the phonemic representation of an utterance is equivalent to the discovery of the succession of axioms that was used in proving a particular theorem. The task of developing suitable strategies is related, therefore, to a general problem in mathematics—that of discovering the shortest proof of a theorem when a set of axioms is given. It should be clear, however, that the powerful tools of mathematics will be at our disposal only when we succeed in describing precisely and exhaustively the generative rules of speech. Until such time we can hope only for partially successful analyzers with strategies that can never be shown to be optimal.

[11] T. Chiba and M. Kajiyama, "The Vowel, Its Nature and Structure," Tokyo-Kaiseikan, Tokyo, Jap.; 1941.

[12] H. K. Dunn, "The calculation of vowel resonances, and an electrical vocal tract," *J. Acoust Soc. Am.*, vol. 22, pp. 740–753; November, 1950.

[13] M. Halle, "The Sound Pattern of Russian," Mouton and Co., The Hague, The Netherlands; 1959. N. Chomsky and M. Halle, "The Sound Pattern of English," to be published.

[14] A. M. Liberman, F. Ingemann, L. Lisker, P. Delattre, and F. S. Cooper, "Minimum rules for synthesizing speech," *J. Acoust. Soc. Am.*, vol. 31, pp. 1490–1499; November, 1959.

[15] H. M. Truby, "Acoustico-cineradiographic analysis considerations," *Acta Radiologica*, (Stockholm), Suppl. 182; 1959.

[16] G. Rosen, "Dynamic Analog Speech Synthesizer," Res. Lab. of Electronics, Mass. Inst. Tech., Cambridge, Tech. Rept. No. 353; February 10, 1960.