26

# ON THE RECOGNITION OF SPEECH BY MACHINE

G. W. HUGHES and M. HALLE

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts, U.S.A.

The following essay is devoted to a discussion of the problems involved in constructing a device capable of recognizing speech. If we have dwelt at length on theoretical issues, this is due to our conviction that at present these constitute the most serious obstacle to a satisfactory solution. In the discussion we have tried to show not only the difficulties inherent in the problem, but also to outline an orderly method of attack that would yield results having practical utility even in the event that a complete solution is not achieved. The results of the theoretical discussion have been implemented in an actual computer program for speech recognition which is described in the last section of the paper.

The ultimate aim of the studies reported here is to duplicate in a machine an essentially human function, that of recognizing spoken messages. Specifically, we are interested in describing a procedure that will distinguish among utterances that are normally recognized as different by a native speaker of a particular language. We do not propose to simulate all the complex processes involved in the understanding of speech. In particular we do not consider the question of how the identification of spoken messages might be learned by a device or by man. Unlike investigators concerned with machine learning, we are in a position therefore to take full advantage of all that is known about speech, regardless of whether or not we are able to describe a mechanical procedure for deriving this information from the data.

We propose to differentiate among all utterances which are not repetitions of each other. Our device will, therefore, class as the same such physically different utterances as, e.g., the French formidable! spoken by a soprano and by a baritone.

It will not distinguish among semantically different utterances that are homophonous, as the past tense of the English verb *lead* and the name of the chemical element Pb. It will, however, be able to distinguish among physically quite similar utterances, as the English *pitch* and *peach*, or the French *l'un* and *l'ain*. We shall call the speech events among which we propose to distinguish, the set of *phonetically different utterances*.

It has been proposed to attack the problem of distinguishing among the phonetically different utterances by first separating one utterance from all others. When this is achieved, the repertoire might be increased to two, three, four, etc., utterances. Consider, for example, the case of a language containing *n* phonetically different utterances. In order to identify any given utterances in the set--i.e., to differentiate it from the remaining *n-1*--as many as *n-1* differences may have to be employed. In order to identify all *n* utterances as many as *n(n-1)/2* differences may have to be taken into account, which is a very large number for ensembles of even moderate size. Furthermore, for a natural language, it is impossible in principle to specify an upper bound on *n*.[1] This approach is, therefore, not feasible if a general solution is desired.

The problem of identifying physical phenomena belonging to an unbounded ensemble is one of several in which speech recognition presents interesting parallels with chemical analysis. In solving this problem both chemistry and the study of language have had recourse to essentially the same solution: the complex phenomena to be classified were regarded as configurations of simpler entities, of which there is only a strictly limited number.[2] In the case of chemical compounds, these are the elements whose significant properties are determined in turn by subatomic particles. In the case of speech events, the elementary entities are the *phonemes*, which themselves are composed of a number of special binary attributes, the *distinctive features*.[3] A speech event is then regarded as a mapping into an acoustically continuous signal of a string of phonemes, each characterized by a particular set of distinctive features. The number of phonemes in the different languages is small: it varies between 15 and 100. The set of distinctive features is even more restricted: some 12 to 15 distinctive features suffice to characterize all phonemes of all languages of the world. No single language, moreover, utilizes all distinctive features; most languages use between 8 to 12 features.

The view that utterances are strings of discrete segments is of course inherent in all alphabetic writing systems. Few writing systems decompose the temporally discrete segments into more basic constituents.[4] Yet even superficial consideration of the facts indicates that such a decomposition reduces not only the number of independent variables that have to be taken into account, but also provides a natural means for expressing regularities in linguistic behavior for which there is no simple description if the temporally discrete phonemes are viewed as the ultimate constituents of language. As an example consider the six diffuse vowels of standard German whose feature characterization is given in the following table:

| | |
|---|---|
| /u/ as in *Glut* "blaze" | tense, flat, grave |
| /ü/ as in *glüht* "blazes" | tense, flat, acute |
| /i/ as in *Glied* "member" | tense, nonflat |
| /U/ as in *Gluck* (proper name) | lax, flat, grave |
| /Ü/ as in *Glück* "fortune" | lax, flat, acute |
| /I/ as in *Blick* "look" | lax, nonflat |

In German an important rôle is played by a phonetic process called *Umlaut*, which consists in the replacement in cognate forms of /u/ and /U/ by /ü/ and /Ü/ respectively; e.g., *Zug* "train"--*Züge* "trains"; *Mutter* "mother"--*Mütter* "mothers". In terms of phonemes no reason can be advanced why the correspondences are between /u/ and /ü/ and between /U/ and /Ü/, and not vice versa. In terms of distinctive features, Umlaut can be described simply as a replacement of grave vowels by their acute cognates, all other features remaining unaffected. We find thus that phonemes sharing certain distinctive features exhibit similar behavior in particular contexts. The parallel to the groups of elements in the periodic table is apparent.

In yet another respect there is an interesting parallel between speech recognition and chemical analysis. In both cases the objective is to find abstract representations for physical phenomena. In speech recognition, alphabetic (phonemic) representations of different acoustical events (utterances) are sought. In chemical analysis, the purpose is to discover the chemical formula of a substance. The science of chemistry states the relationship between the properties of the elements and those of their compounds, just as a grammar of a language describes the relationship between the feature composition of the phonemes and the physical properties of sentences. In neither case are the properties of the complex phenomena to be analyzed--i.e., of the chmical substances and of the linguistic utterances, respectively--simple sums of the properties of their components. Consequently in neither case will analysis procedures necessarily be based on the detection of properties which are directly associated with basic constituents. There is no need to expatiate here on the indirect relationship between the measurable properties of compounds and the defining properties of the chemical elements. With regard to the analogous relationship in the case of speech recognition, the following facts may, however, be noted.

The set of acoustical properties measurable in the speech waveform can, in general, not be interpreted directly in terms of distinctive features. There is certainly no one-to-one

The emphasis put here on the indirect nature of the relationship between the abstract phonemic representation and the measurable properties of the speech signal must not obscure the lawful nature of this relation. For every properly formed utterance--i.e., excluding slips of the tongue, foreign accent, etc.--there exists a set of statements which map the phonemic representation into the acoustical signal. As a matter of fact, in many specific instances this set of statements is of a simple, direct type. For example, in vowels the feature "grave-acute" is correlated directly with the difference in the frequency positions of the lowest two vocal tract resonances (formants): in grave vowels this difference is small; in acute vowels it is large. As will be shown below a relatively simple measurement procedure has been implemented to extract the relevant information about this feature from the signal.

Unfortunately the existence of a simple relationship between an acoustical measurement and a distinctive feature does not of itself guarantee a workable measurements procedure. On the one hand, engineering difficulties often arise in instrumenting such procedures. On the other hand, for certain distinctive features there do not as yet exist fully adequate descriptions of the acoustical correlates.

It is evident that at present a complete solution to the problem of speech recognition cannot be offered. It is, however, possible even now to instrument an identification scheme capable of tracking a restricted set of features in an utterance. Such an incomplete identification scheme will be able to differentiate only among some of the phonetically different utterances; it will "confuse" many phonetically different utterances. For instance, if we can distinguish only three different frequency positions of the first vowel formant and have no information about the frequency of the second formant, we can separate all English vowels into three classes: the tense diffuse /i/ and /u/ (*peel* and *pool*), the compact /a/ /æ/ and /ʌ / (*pot*, *pat* and *putt*), and a third class containing all remaining vowels.

Such an incomplete identification scheme can have practically useful results, since we know in advance how it will classify any given input utterance. It is then possible, for instance, to select the input utterances to a voice-operated device in such a manner as to make communication between the operator and the device depend completely on features which the identification scheme is capable of handling reliably. If a device capable of performing the above tripartite classification of the vowels were also able to locate vowel sounds in the signal,[7] it could distinguish among three monosyllabic utterances, nine bisyllabic utterances, 27 trisyllabic utterances, etc., provided, of course, that these be suitably preselected. A computer might thus be instructed by means of the following utterances: start, send, cease, keep two, read one, jump three, set cores, hold last, cut three, stop shift, add up, etc.

The above theoretical considerations have guided our experimental efforts which so far have been directed towards machine tracking of the acoustical properties most evident in a speech signal. For this purpose the digital computer Whirlwind I at the M.I.T. Lincoln Laboratories was employed. As this is being written the programs outlined below were actually in operation.

The signal is fed to a bank of 35 bandpass filters. The outputs of the filters are rectified, smoothed and sampled by a rotary switch 180 times a second. After this information has been converted to digital form, it may be held by the computer for immediate analysis or stored digitally on magnetic tape for later processing. A more complete description of the input system has been given elsewhere.[6] All programs operate solely on this spectral information which enters the computer at the rate of 69,300 bits per second.

The program in use may be described as follows:

1. The relative amount of high-frequency energy in the sound is calculated. On the basis of this calculation the segment is classified as sonorant or nonsonorant, i.e., as having formant structure or not. If classified as a sonorant the segment is tested further for formant position (see 2). If classified as nonsonorant the sound level data determine whether it is to be tested for class of fricative or stop (see 5), or identified as a silence produced by a complete closure of the vocal tract preceding stop phonemes.

2. The segments classified as sonorants are then subjected to a formant-tracking program which establishes the frequency position of the two lowest vocal tract resonances. A block diagram outlining the logic of this program is shown in Fig. 1. After peaks are located in the spectra, constraints are applied which to some extent smooth out abrupt changes in the formant locations and prevent confusion of closely spaced formants. These programmed constraints are an attempt to simulate some of the physical limitations on the possible configurations of the vocal tract and their rate of change. This program forms the basis for other programs dealing with sonorant segments.

3. In vowel segments the difference in frequency between the locations of the first and second formant is computed and the frequency position of the first formant classified as "low", "medium", or "high". The boundaries of the latter three regions are determined, in part, by the results of the frequency difference computation.

4. A calculation is made of the amount of change in formant frequency and sound level during successive intervals of about 40 msec throughout the utterance. If the total of the changes exceeds a threshold, a boundary between phonemes is marked.

5. The segments classified as fricatives or stops are subjected to measurement procedures outlined elsewhere.[9] These procedures eventuate in a tripartite classification.

So far some 70 isolated words each spoken by 15 native speakers have been processed. The words used were selected so as to place each phoneme of English in as many environments as possible. The memory capacity of the computer limits the maximum length of each utterance to 800 msec.

Results to date show that although the individual acoustic feature tracking programs perform quite well (on a 90-95% confidence level), there are at least two types of error which are now the subject of an intense investigation.

Inadequacies in equipment and programs result in excessive sensitivity to small perturbations and noise. With the present system it is necessary to specify thresholds which are relatively inflexible. As a result unforeseen and informationally unimportant changes in the input signal are sometimes sensed by the analysis programs. For example, a temporary drop in the second formant level may force the formant tracking program into mistakes from which it is difficult to recover even when the level returns. In many cases, the effort to correct such difficulties has been successful. There is, however, always the danger of overcorrection with a concomitant insensitivity to legitimate variations.

The second source of error is due to incomplete use made of the interdependence of acoustical features. The programs currently employed utilize to some extent previous measurement results to help specify the procedures that follow. For example, the frequency position of the first formant will in large part determine the frequency range in which the second formant is to be located. The measurement of the difference in frequency between the first and second formant (grave-acute) helps to specify the frequency boundaries of the first formant classes "low", "medium" and "high" (compact-diffuse). In the programs much more use must be made of flexibility in classification procedures. The interpretation and specification of measurements must be tied more closely to other measurements which themselves may or may not be fully interpreted. Much of our effort in the near future will be centered on this problem. The solution, we feel, will be a major contribution towards the development of reliable overall identification procedures which of necessity must be based on imperfect programming and input hardware.

# NOTES

[1] Chomsky, N., "Three Models for the Description of Language," IRE Transactions on Information Theory, IT - 2, No. 3 (1956): 113-124.

[2] The parallelism between the problem of describing the unbounded variety of entities in the material world and that of describing the totality of linguistic utterances was brought out very clearly by the Greek philosophers. In Plato, Aristotle and their immediate followers, the word στοιχεῖον means both "speech sound" or "letter" as well as "elementary unit of matter, element, atom," and the word συλλαβή means "syllable" as well as "combination or compound of elements." Thus Eudemus, a disciple of Aristotle, writes in his treatise on physics: "For elements would seem to be present [in physical objects] just as individual letters are present in articulated speech." Cf. Diels, H., Elementum, Leipzig, B. G. Teubner, 1899, p. 35. The above parallelism between linguistics and physics was brought to our attention by our colleague Roman Jakobson.

[3] On the concept of the phoneme, see Sapir, E., "Sound Patterns in Language," and "The Psychological Reality of Phonemes," in Mandelbaum, D. G., ed., Selected Writings of Edward Sapir Berkeley and Los Angeles, University of California Press, 1949, pp. 33-61; Troubetzkoy, N. S., Grundzügeder Phonologie = Travaux du Cercle Linguistique de Prague, VII (1939); and Martinet, A., Phonology as Functional Phonetics, London, Oxford University Press, 1949.

For a detailed discussion of the distinctive features see Jakobson, R., Fant, C. G. M., and Halle, M., Preliminaries to Speech Analysis = M.I.T. Acoustics Laboratory Technical Report, No. 13 (1952); Jakobson, R., and Halle, M., Fundamentals of Language, The Hague, Mouton & Co., 1956; see also Halle, M. H., "Three Lectures on Linguistics," Il Nuovo Cimento (in press); and Cherry, E. C., "Roman Jakobson's Distinctive Features as the Normal Coordinates of Language," For Roman Jakobson, The Hague, Mouton & Co., 1956, pp. 60-64.

[4] An interesting way of decomposing phonemes into distinctive features is reflected in the shape of the letter of the standard Korean alphabet.

[5] Menzerath, P., and de Lacerda, A., Koartikulation, Stauerung, und Lautabgrenzung, Berlin-Bonn, 1933.

[6] For more examples of this type see Fischer-Jørgensen, E., "What Can the New Techniques of Acoustic Phonetics Contribute to Linguistics?", Proceedings of VIII International Congress of Linguists (Oslo, 1958), pp. 470-478.

[7] Devices capable of performing this partial identification have been proposed and even constructed; see, for example, Radley, J.-P. A., *M.I.T. R.L.E. Quarterly Progress Report* (April, 1954), p. 70, and Hughes, G. W., *et al.*, *ibid.* (October, 1956), p. 109.

[8] Forgie, J. W. and Hughes, G. W., "A Real-Time Speech Input System for a Digital Computer," *JASA* 30 (1958): 668.

[9] Hughes, G. W. and Halle, M., "Spectral Properties of Fricative Consonants," *JASA* 28 (1956): 303-315; Halle, M., Hughes, G. W., and Radley, J.-P. A., "Acoustic Properties of Stop Consonants," *JASA* 29 (1957): 107-116.