AUTOMATIC SPEECH RECOGNITION

D. B. FRY

The problem of automatic speech recognition has occupied the attention of communication engineers for some fifteen or twenty years and during this time has aroused a certain amount of interest in the world of linguistics and phonetics. The papers that have been published on the subject can be divided into two broad classes which one might perhaps label the "philosophical" and the "practical". The first type of paper deals with fundamental aspects of the problem and says in effect that in order to carry out the process of automatic speech recognition one would need a certain part of the machine to perform such and such an operation, another part to do a second operation, a third to do a third and so on. The second type of paper is that which reports experiments already carried out and which gives an idea of how far in practice we have progressed. Owing to the complexity of the problem, many of the latter may well deal with only a section of the total process. In recent years and increasingly as time goes on, the tendency has been to carry out experiments with digital computers and reports of this work form a sub-class of the practical papers.

Although the published material on the subject is not very extensive, it will not be possible in this short paper to review it adequately and we shall therefore attempt rather to take stock of the present situation and to indicate the directions in which progress is most likely to be made.

It is not necessary at this stage to discuss the purposes of automatic speech recognition but it will be as well to say once more what we understand by the term. The process consists in the transformation of the sound-waves of speech into signals representing linguistic units, that is to say having a one-to-one correlation with them. The mode of operation of a particular mechanical recognizer will determine the choice of linguistic units but any machine which produces signals correlated with phonemes or letters, with words or sentences is performing automatic speech recognition.

The success of a mechanical recognizer can be measured by the accuracy with which it effects the transformation from sound-waves into quasi-linguistic signals. For any speech input sequence, we can determine the number and the order of the phonemes (or letters), words or sentences that it contains. What is required of the automatic recognition process is that each time a given linguistic unit occurs in the input, the same signal should appear at its output; any occasion on which this does

not happen constitutes an error on the part of the machine and the number of errors reflects the success of the operation.

It should be noticed that the use made of the output signals is not relevant to the fundamental problem of recognition. They may be made to produce visual signals, synthesized speech or electrical signals for a variety of purposes. In each case a further transformation of the recognizer output takes place and the form chosen becomes important only when we consider the success of the over-all system. The number of errors that can be tolerated in automatic recognition, as we showed by experiments some years ago, will depend on the purpose to which the quasi-linguistic signals are put and the ways in which they are further transformed.

A machine which could carry out automatic recognition on running speech in any one language, even supposing we knew how to design such a machine, would be of enormous complexity and considerable size, even with present-day electronic techniques. In all the experiments that have been done up to the present, efforts have been made to bring the problem within manageable bounds by reducing the variability of the speech input, limiting it to the speech of one or perhaps several speakers, and by severely restricting the ensemble of linguistic units that the machine is asked to recognize, that is by using in effect an artificial language. Success in recognition, that is the proportion of errors the mechanical recognizer is likely to make, is related in a very direct manner to the size of the ensemble and this fact has to be taken into account in evaluating the performance of any particular system.

Any act of recognition, whether it be performed by a human being or by a machine, consists in taking in information, or a "stimulus", and classifying it by reference to information already stored in the system. In speech reception, the incoming stimulus for both men and machine is provided by the sound-waves. In the case of the human being, the classifying is done on the basis of stored knowledge about the language system and such is the degree of redundancy of natural languages that the weight attached to the incoming acoustic information is low compared with the weight given to stored information. It is this feature of human speech recognition which has so far been difficult to reproduce in a mechanical recognizer and which high-speed digital computers offer a greater possibility of simulating.

In most of the mechanical recognizers that have been used as a basis for experiment, the principle of operation has been the comparing of incoming acoustic patterns with stored patterns representing acoustic features. The machine looks for the best match between the incoming and the stored patterns and makes a decision based entirely on acoustic information. Obviously such an operation must form one element in any process of automatic speech recognition. The recognizer must operate on the acoustic features of the speech input as a basis for any subsequent procedure. The question is how far acoustic recognition alone can be a successful method of automatic recognition. When the linguistic ensemble is very restricted, the relative weight of acoustic and linguistic information is, of course, very materially altered. A number of experiments have been done, for example, on the automatic recognition of spoken

numbers, 0 to 9. In this case, the fact that the language system is restricted to ten words has great force, but the weight attached to linguistic constraits *within the system* is very slight. At the word level, there is generally no sequential dependency between one number and the next; at the phonemic level, there are strong constraints within words but with such a small repertory, the phonemic sequential dependencies are faithfully reflected in the acoustic constraints since every phoneme occurs in only one or two, or at the most three, different contexts. Here phonemic invariance goes with relatively little acoustic variability.

Two sets of experiments with vocabularies of this size have been reported. The first is the work on the recognition of spoken numbers by the Bell Telephone Laboratories (Davis, Biddulph, Balashek and Dudley) in which series of numbers, uttered by one speaker, were recognized by an automatic recognizer operating on a purely acoustic basis with an accuracy of 97%. In the second case, the machine constructed and tested by Olson and Belar had a vocabulary of ten monosyllabic English words and these were recognized, in the speech of one speaker, with an accuracy of 98%. Both of these sytems carried out recognition at the word level. Although the high scores attained are promising at first sight, the methods used to deal with these very small ensembles do not represent a great contribution to the general study of automatic recognition; that is to say it is not possible, in either case, to double the vocabulary, to use the same methods of recognition and still to obtain anything like the same level of accuracy. This does not, of course, detract from their possible usefulness in connection with specific practical problems such as voice dialling for telephone systems.

The principle which seems still to hold the greatest promise of a general solution to the problem is recognition at the phoneme level. If a machine could recognize all the phonemes in one language in a wide variety of contexts, with the required level of accuracy, then the addition of *words* to the ensemble used would entail no serious decrease in accuracy; one would in fact be achieving the economy of means to be found in natural languages. It is perhaps worth noticing that Dudley and Balashek experimented with the digit recognizer as a phoneme recognizer and although they give no figure for the accuracy of phoneme recognition, it is clear from their published data that the level is very much lower than for the ten *words* in its vocabulary. Phoneme recognition has also been attempted in the system of Dreyfus-Graf and in the mechanical recognizer designed by Denes and built at University College, London. In Dreyfus-Graf's machine, the ensemble of units was reported to be all the phonemes of French, but his published papers do not, unfortunately, give any value for accuracy of recognition. The recognizer at University College dealt with an ensemble of 13 English phonemes and had a potential vocabulary of all English words containing only these phonemes. It was tested repeatedly with a vocabulary of 200 words uttered by one speaker and although accuracy in acoustic recognition was not the principal objective of the experiments (see p. 317), it achieved an accuracy of 60% phoneme recognition. The corresponding word score was 24% and there is clearly

a great difference between this level of accuracy and that achieved in the experiments with ten-word ensembles. Here again the same qualification must be made as in the case of the digit recognizer: any addition to the phoneme repertory of the machine would affect the recognition of the phonemes in the existing ensemble. On the other hand, the recognizer was dealing with about one-third of all the English phonemes and this represents a considerably greater proportion of the required total than ten words would represent in comparison with the size of any word ensemble that would be useful for general purposes.

It is clear that successful acoustic recognition will call for more sophisticated methods than have yet been used. All the practical systems so far constructed have employed some method of frequency analysis and there is no doubt that this will always form a part of the process of acoustic recognition. A consideration of results obtained in experiments on the recognition process in the human listener indicates, however, certain directions in which the acoustic methods need to be improved. The salient feature of these results is the demonstration that a listener uses a variety of cues in making one recognition. In particular, cues in the time domaine, the spacing of events on the time axis, and relatively rapid changes, such as formant transitions, are seen to have great importance. Improvements in acoustic recognition in the mechanical recognizer can probably be brought about only by making use of a number of acoustic criteria at the same time. This technique was used to a very limited extent in the mechanical recognizer at University College in which spectral information was combined, for certain recognitions with information about over-all intensity and for others, with information about duration.

The application of a multiplicity of acoustic critera has been described by Wiren and Stubbs who based their system on the detection of *distinctive features*. Their work brings out a most important point about acoustic recognition and that is that we are not likely to get the best results if acoustic criteria are applied in the form of a binary decision tree. As these authors themselves say, such a system will have a low error rate only if, at each branching of the tree, a correct decision is certain or very highly probable. It is safe to say that there is no acoustic decision about speech that we can make with certainty; in the acoustic, just as in the linguistic, sphere we are always dealing with probabilities and we cannot in fact afford to throw away the information that is discarded each time a binary decision is made. What is required is that whenever a criterion is applied, the output of that particular stage should take the form of an expression of probabilities with respect to all the phonemes in the machine repertory. A final acoustic decision would then be made by computing the combined acoustic probabilities for *all phonemes* and for *all criteria* and selecting the phoneme showing the highest combined probability.

The mechanical recognizer at University College used more than one acoustic criterion for certain decisions but the criteria were applied serially and not in the manner just suggested. The only practical system which has made any use of this principle is that of Sakai at Kyoto University in which spectral analysis and zero-crossing counts are carried out and certain decisions made on the basis of the coincidence of features revealed by the methods of examination. A much more thoroughgoing application of the principle will probably be necessary before the maximum accuracy in acoustic recognition is attained.

It is clear, however, if we consider speech recognition in the human listener that acoustic recognition alone will not provide an automatic recognizer with a satisfactory level of performance. In the human being, the role of primary or acoustic recognition is only to provide a scaffolding on which the complete message is constructed by applying linguistic constraints at all levels. The sounds that are perceived are classified as phonemes, the phoneme sequences are formed into morphemes, the morpheme sequences into words and the word sequences into sentences. At each of these levels the appropriate constraints are applied both in the act of recognition and in order to detect errors. While it is true, therefore, that we have to perceive sounds in order to take in speech, the part played by perception is very much less than is generally believed and, as we said earlier, the weight given to incoming acoustic information is less than that given to stored linguistic information. We know, for example, that a listener to English can take in continuous messages if he recognizes correctly 50% of the phonemes and that when he is in doubt, he will resolve ambiguities according to linguistic constraints rather than on the basis of the acoustic signal. The listener in fact has a very powerful system for carrying out recognition in the face of great variability in the acoustic input and for correcting errors at all levels of the decoding process. An automatic recognizer with even a small proportion of the adaptability and accuracy of the human recognizer will not be achieved without recourse to methods similar, in principle at least, to those used by the listener.

The only attempt so far to implement these principles is to be found in the recognizer built at University College which was in fact expressly designed to demonstrate the feasibility and the desirability of using linguistic information in a mechanical recognizer. The linguistic information was of the most elementary kind – knowledge of phoneme digram frequencies – and yet its application increased the proportion of words correctly recognized from 24%, the value given above for acoustic recognition, to 44%. This means that with the same speech input and with exactly the same process of acoustic recognition, the accuracy was almost doubled by the application of linguistic constraints of a very low order.

It is clear that there are many difficulties in the way of using extensive linguistic constraints in automatic speech recognition but the use of digital computers with their ready-made storage has put a new complexion on experiments in this field. It should be possible to simulate, at least for a restricted language, the hierarchic structure of the natural recognition process. This would involve phoneme recognition with the aid of digram, trigram and possibly tetragram frequencies. It would probably not be very profitable to include a morpheme level in the process but sequences of phonemes would be fed forward for identification as words, making use of word sequential probabilities. At each level linguistic constraints would be used both in

making recognition decisions and in correcting errors. For example, if phoneme trigram information were used, then in making each phoneme decision sequential probability to three places would operate as one factor. When a phoneme had been recognized, the sequence formed by this phoneme and the two preceding ones would be checked against the complete trigram inventory to test whether the sequence was a highly probable or at least a possible one. In a restricted language such a procedure might lead to the detection of a substantial number of errors. In a case where an error was detected, an attempt at correction would, of course, involve reference back to the output of the acoustic recognizer. A similar process would be repeated at the word level where word sequential probability would be used to identify words and each complete word checked against the word inventory. If necessary, sentences could be checked against a sentence inventory, though it is doubtful whether it would be worth while to go to this stage.

It is evident that to perform even a part of this process would present serious storage problems if one were trying to deal with the whole of current speech in one language. The immediate task is to attempt some parts of it in experiments with a language which is restricted but is nonetheless large enough to offer some justification for generalizing from the results. We suggest that it is to this aspect of automatic speech recognition that the resources of modern digital computers can most profitably be applied.

University College
London

BIBLIOGRAPHY

Biddulph, R., "Short-term autocorrelation analysis and correlatograms of spoken digits", *J. acoust. Soc. Amer.*, 26 (1954), 539.
Chang, S–H., "Two schemes of speech compression system", *J. acoust. Soc. Amer.*, 28 (1956), 505.
Chao, Y. R., "Linguistic prerequisites for a speech writer", *J. acoust. Soc. Amer.*, 28 (1956), 107.
David, E. E., "Voiced-actuated machines: problems and possibilities", *Bell Labs. Record*, 35 (1957), 281.
Davis, K. H., Biddulph, R., Balashek, S., "Automatic recognition of spoken digits," *J. acoust. Soc. Amer.*, 24 (1952), 637.
Denes, P., Mathews, M. V., "Spoken digit recognition using time-frequency pattern matching", *J. acoust. Soc. Amer.*, 32 (1960), 1450.
Dreyfus-Graf, J., "Phonétographe et phonétique", *Folia Phoniat.*, 5 (1953), 223.
Dreyfus-Graf, J., "Le sténo-sonographe phonétique", *Bull. Tech. PTT*, 1950, No. 3.
Dudley, H., Balashek, S., "Automatic recognition of phonetic patterns in speech", *J. acoust. Soc. Amer.*, 30 (1958), 721.
Dudley, H., "Phonetic pattern recognition vocoder for narrow band speech transmission", *J. acoust. Soc. Amer.*, 30 (1958), 733.
Forgie, J. W., Forgie, C. D., "Results obtained from a vowel recognition computer program", *J. acoust. Soc. Amer.*, 31 (1959), 1480.
Fry, D. B., Denes, P., "On presenting the output of a mechanical speech recognizer", *J. acoust. Soc. Amer.*, 29 (1957), 364.

Fry, D. B., Denes, P., "The solution of some fundamental problems in mechanical speech recognition", *Language and Speech*, 1 (1958), 35.
Fry, D. B., Denes, P., "An analogue of the speech recognition process", *Mechanization of Thought Processes*: *National Physical Laboratory Symposium No.* 10, 275 (London, H. M. Stationery Office, 1959).
Hughes, G. W., "Identification of speech sounds by means of a digital computer", *J. acoust. Soc. Amer.*, 31 (1959), 113.
Olson, H. F., Belar, H., "Phonetic typewriter", *J. acoust. Soc. Amer.*, 28 (1956), 1072.
Sakai, T., Inoue, S–I., "New instruments and methods for speech analysis", *J. Acoust. Soc. Amer.*, 32 (1960), 441.
Smith, C. P., "A phoneme detector," *J. acoust. Soc. Amer.*, 23 (1951), 446.
Smith, C. P., "The analysis and automatic recognition of speech sounds", *Electronic Eng.*, 24 (1952), 368.
Stevens, K. N., "Towards a model for speech recognition", *J. acoust. Soc. Amer.*, 32 (1960), 47.
Wiren, J., Stubbs, H. L., "Electronic binary selection system for phoneme classification", *J. acoust. Soc. Amer.*, 28 (1956), 1082.