Some Experiments with a Simple Word Recognition System

JOHN N. SHEARME
PETER F. LEACH

Abstract—This paper describes some pilot experiments in a project to recognize selected words spoken by any talker. For this initial work, a population of 10 talkers is used and the 32 test words are spoken in isolation. In the experiments, a word is represented by a set of regularly spaced time samples of the normalized spectrum envelope, and recognition is achieved by comparison of such a set of samples with a library of stored sets. Limitations on computer storage necessitate a compact method of coding a spectrum sample. Two methods have been compared, one of which classifies spectra into a very small number of types, while the other uses a 24-bit representation. The most serious problem encountered in spectrum matching is the well-known lack of synchronism between corresponding spectral events when phonemically identical words are spoken by different talkers. This paper describes an attack on this problem, which relies on using a number of sets of spectrum samples to represent each word in the stored library of words to be recognized. There are a number of ways of using the scores obtained from matching an unknown word with a library of known words. Some of these are described and some results are given both for the case where the unknown word is known to be in the stored library, and for the case where there is no such limitation.

Manuscript received September 25, 1967. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass.

The authors are with the Joint Speech Research Unit, Ruislip, Middlesex, England.

I. Introduction

THIS PAPER is concerned with a conceptually simple automatic word recognition process, based on comparing the short-term spectra of input words with those of a stored library of words.

After spectrum analysis, amplitude normalization, and time sampling, a digital computer is used to code the resulting sequence of spectrum cross-section samples representing an input word into a sequence of identifying numbers. This sequence is compared with a library of such sequences representing words which are to be recognized, and a recognition decision is based on the closeness of match.

For the experimental work described here, the input was a list of 32 words spoken in isolation and with reasonable care by 10 male talkers. The words come from a computer control context and are not chosen to satisfy any particular phonetic constraints. They include the spoken digits zero to nine.

The spectrum analysis is performed with a bank of 20 simulated bandpass filters and envelope detectors to give a short-term spectrum cross-section (or profile) described by 20 ordinates. The profile is sampled 100 times per second. Each sample (consisting of 20 ordinates) is amplitude normalized by taking the logarithm of each ordinate and subtracting the mean of the 20 logarithms from it. The normalized ordinates then represent the spectrum shape of each sample independently of amplitude level. Amplitude information is, of course, potentially useful for recognition, but for simplicity it was decided to see what could be done initially with spectrum shape only.

II. Spectrum Profile Coding

A system of the type to be described becomes more tractable, and a necessary saving in computer storage space can be achieved, if the 20-ordinate spectrum profiles are coded into a simpler form.

The 20 ordinates of each spectrum profile sample can be considered as defining a point in a 20-dimensional spectrum space. The type of spectrum coding which forms the basis of the work specifies a number of prototype spectra which define fixed points in the space. Each of these points is labeled, and a spectrum profile sample is coded by quoting the label of the "nearest" prototype spectrum. (The "distance" between two points in the space is calculated by taking the difference between corresponding ordinates and summing over all 20 such differences, ignoring sign.)

The label 0 is used to designate silence (i.e., signal below a specified amplitude threshold), and has to be separately derived since amplitudes have been normalized out of the spectrum profiles.

The choice of prototype spectra is obviously of crucial importance; the prototypes must be distributed efficiently throughout regions of the space that are occupied by speech spectra. The method of choosing prototype spectra will not, however, be described here. In our experiments, 20 prototypes labeled 1–20 (excluding silence) have been used, and it is perhaps confusing that this is the same as the number of ordinates describing a spectrum profile; there is, of course, no connection between the two.

With 20 prototypes, a spectrum profile is classified by about 4.3 bits; this can be compared with typical coding in a 20-channel vocoder, where the spectrum would be coded by 60–80 bits. It is not, of course, suggested that speech synthesized from 4.3-bit spectra compares favourably with speech from a vocoder, but it is, nevertheless, fairly intelligible, and the considerable reduction in data has significant advantages for recognition purposes; 20 classes can be easily handled, $2^{60} (\approx 10^{18})$ cannot.

III. Word Matching

The string of coded spectrum profile samples derived from a word is represented in the computer as a succession of prototype spectrum labels. For example, the word "eight" might be represented by the sequence: 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 1, 2, 1, 2, 0, 0, 0, 0, 0, 15, 15, 14, 15, 15, 15. Here, the early part of the sequence is due to the diphthong and is followed by the silent gap represented by zeros and the noise burst of the /*t*/.

Such sequences can be made time independent by taking only the first member of any string of repetitions, e.g., the sequence above becomes 3, 2, 1, 2, 1, 2, 0, 15, 14, 15. It had been hoped to use this property to achieve freedom from time variations in word matching. A difficulty arises, however, which is illustrated by the sequence above. The excursions between prototype 1 and 2 (or between 14 and 15) are almost certainly not definitive of the word; it just happens that this particular speaker produces a spectrum near the boundary between prototypes 1 and 2 for the last part of the diphthong. Another speaker may well produce the (time independent) sequence 3, 2, 0, 15 for this word, and it becomes very difficult to match sequences which differ in this way. It was also found to be impracticable to make rules for producing acceptable sequences by rejecting unimportant members. Because of these difficulties, it was decided to use the complete sequences for word matching.

For recognition, we have to match the sequence representing an incoming word with sequences from a stored library of words, and thus to produce matching scores which will enable us to identify the incoming word. The library sequences are called templates to distinguish them from incoming sequences. In our experiments, goodness of match is measured by summing the distances in 20-dimensional space between corresponding members of the two sequences, with distance defined as in the prototype spectrum coding. As two sequences being matched will, in general, contain different numbers of members, the operation is carried out starting from the first member of each sequence and continuing over all members of the template. If the input sequence is shorter than the template, zeros (silence) are added to make up the length (indicated by asterisks in the example below). If it is longer, it is truncated. An example is given in Table I.

TABLE I

	Shorter Input Word									
Input sequence	3	3	2	1	8	0*	0*	0*		
Template	3	2	2	2	7	7	8	7		
Distance =	0	$+d_{32}$	$+0$	$+d_{12}$	$+d_{87}$	$+d_{07}$	$+d_{08}$	$+d_{07}$		
	Longer Input Word									
Input sequence	3	3	2	1	8	7	7	7	6	6
Template	3	2	2	2	7	7	8	7		
Distance =	0	$+d_{32}$	$+0$	$+d_{12}$	$+d_{87}$	$+0$	$+d_{78}$	$+0$		

It will be noted that a good match is indicated by a low score. In order that the score be independent of template length, it is divided by the number of members of the template. The distances between prototype spectra d_{01}, d_{12}, etc., can all be tabulated owing to the small number of prototypes. These distances are calculated from the ordinates of the prototype spectra at present, but arbitrary distances could be used based, for example, on judged phonetic distances between prototypes. For purposes of distance measurement, silence is assumed to have a flat spectrum (i.e., all ordinates are zero after normalization).

At this stage, we have to consider what templates should be used in the library. We could, for instance, represent each word to be recognized by a template which was some sort of average sequence over all talkers and conditions for the word concerned. Alternatively, a number of templates representative of the word (as spoken by different talkers, for instance) could be used. Examination of some actual sequences showed that the latter method is likely to be much more powerful because of the wide differences between sequences due to different talkers. An investigation of the number of templates required per word is the main experimental work described in the paper.

IV. Identification of a Word from Matching Scores

In identifying an input word from matching scores, there are two cases to consider: Case I is where the input word is known to be one of the stored library words (although of course not necessarily from the same population of talkers that generated the templates). Case II is where there is no such restriction.

An obvious procedure in Case I is to identify the input word with the library word corresponding to the best fitting (i.e., lowest scoring) template. A slight elaboration on this, which has been found worthwhile when each library word has several templates, is to take a majority vote of the N best fits with votes weighted inversely as the score. Where each library word has been represented by nine templates, the optimum value for N has been found to be about 4.

An estimate of the confidence that can be placed in an identification can be obtained experimentally by counting the proportion of words correctly identified, when known input words are used.

In Case II, we have not merely to find the best fitting library word, but we must have an estimate of whether this fit is good enough to make it likely that the input word is in fact the "same" as the library word. Such an estimate can only be obtained by experiment with a particular library, and a sufficiently large population of talkers and input words (ideally all the words in the language). Fig. 1 illustrates the results of plotting separately the distribution of scores obtained by matching "same" input words (i.e., input words which are, in fact, in the library), and "different" words (i.e., those input words which are not in the library). The curves are normalized probability density curves of matching score for the same and different word conditions. For any particular score, the ratio of the ordinates of the two curves gives the relative probability (or odds) of such a score having been caused by a same word or different word match. Note that the curves refer to matches between input words and a single template. Better discrimination can be expected if the best score due to matching with several templates of the same word is taken. Inspection of these curves is informative about the effectiveness of a recognition system; obviously, wide separation of the curves indicates an effective system. Furthermore, the curves enable a threshold to be set to satisfy any desired criterion as to the relative frequency of occurrence of the two possible types of error (i.e., identifying an input word as a particular word when it is not, and failing to identify an input word as a particular library word when it is).

From the curves, we can associate with a best match score an absolute measure of the probability of it being a "same word" match. This would also provide a means of weighting the results of several independent processes to produce a combined match.

At this point, we are in a position to draw a block diagram of the recognition system under discussion; it is shown in Fig. 2, and we can now go on to describe some experimental investigations of the parameters of such a system.

V. Experiments on the Reduction of Talker Variation Effect

It has been suggested in Section III that talker variation effects might be reduced by increasing the number of templates representing each word in the comparison library. Some experiments have been carried out to test this idea, using a vocabulary of 32 words spoken by each of 10 talkers. The same utterances (320 in all) were used both as input words and as library templates, but in the experiments, identical utterances (i.e., same word and talker) were excluded, since they would necessarily match perfectly. Thus, for a given utterance the maximum number of templates available for matching was nine.

For this experiment, each of the templates representing a word was derived from a different talker. No attempt was made to select the talkers to be representative of as wide a class of talkers as possible. They were selected at random but excluded speakers with strong regional accents.

The proportion of correct identifications in a Case I situation is shown on Fig. 3 for various numbers of templates per library word.

Where the number of templates per word is small (say ≤ 3), the variation in correct identifications is considerable according to which talkers out of the 10 available are selected to provide templates. The mean number of correct identifications increases rapidly with number of templates up to 3 or 4, and then more slowly, but still very consistently, from 4 to 9. Plainly, further work is necessary to include up to, say, 20 templates to see where the curve flattens off.

When nine templates are used per library word, the Case II performance is shown in Fig. 1 (upper curves).

The preceding results were obtained with the prototype spectrum method of coding spectrum profiles. There was some fear that this extreme form of compression might jeopardize good recognition right from the start. To investigate this point, some of the work has been repeated using much more precise coding.

The new coding represents the spectrum profile by a 24-bit code. The first ordinate is coded to 5 bits, and the remaining ordinates by 1-bit Δ-coding, starting with the first ordinate as reference. The bit economy achieved by this code relies on the fact that outputs from neighbouring channels of a filter bank analyzer with a speech input are highly correlated. Very intelligible speech is produced by synthesis from this coding, and although it is less easy to handle, for some purposes, than prototype spectrum coding, it does at least allow a spectrum profile to be stored in one computer word. The recognition process remains unaltered, but each profile sample is represented by a 24-bit number, instead of a number in the range of 1 to 20. Silence is still represented by zero, and distances are calculated from the ordinates after decoding (as opposed to prototype spectrum coding, where distances can be obtained from a small lookup table with considerable saving in calculation).

The Δ-coded data yielded an almost identical proportion of correct recognitions in Case I (91 percent against 90 percent for prototype spectra). The Case II curves are shown in the lower part of Fig. 1, and although they differ somewhat from the curves obtained with prototype spectrum coding, they are not appreciably better. This result was interpreted as indicating that inadequate representation of spectrum profiles due to using prototype spectrum coding was not the main cause of recognition failure.

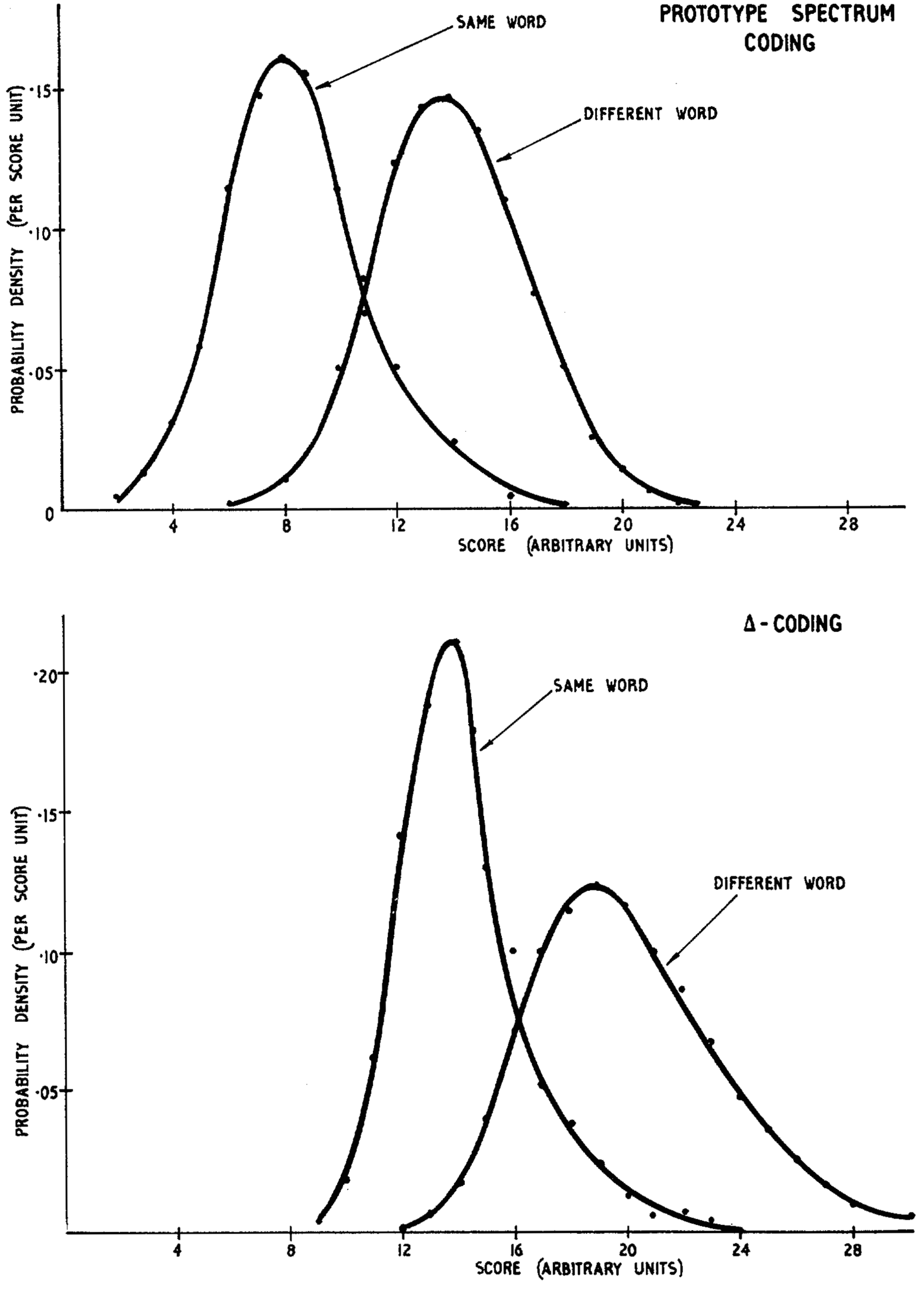

Fig. 1. Distribution of matching scores for same words and different words, with two types of coding.

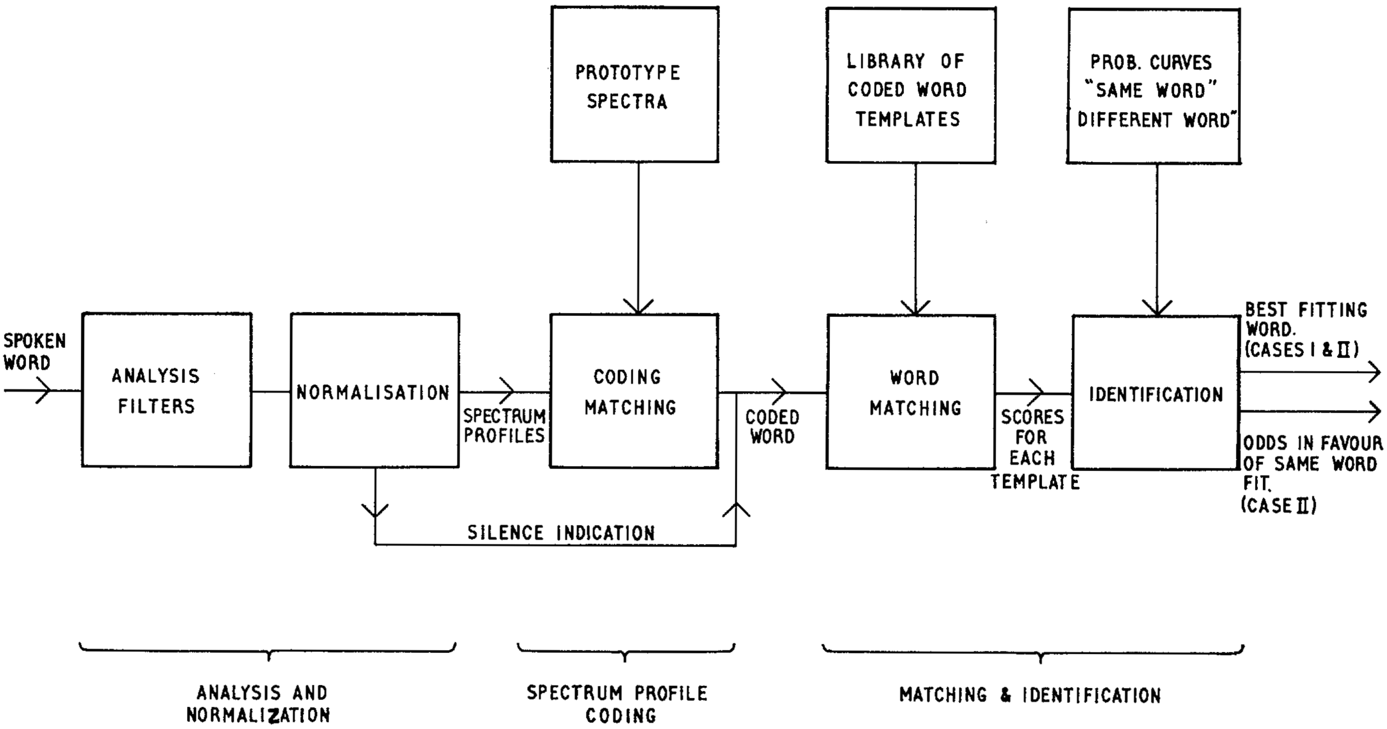
PROTOTYPE SPECTRA
LIBRARY OF CODED WORD TEMPLATES
PROB. CURVES "SAME WORD" DIFFERENT WORD"
SPOKEN WORD
ANALYSIS FILTERS
NORMALISATION
SPECTRUM PROFILES
CODING MATCHING
CODED WORD
WORD MATCHING
SCORES FOR EACH TEMPLATE
IDENTIFICATION
BEST FITTING WORD. (CASES I & II)
ODDS IN FAVOUR OF SAME WORD FIT. (CASE II)
SILENCE INDICATION
ANALYSIS AND NORMALIZATION
SPECTRUM PROFILE CODING
MATCHING & IDENTIFICATION

Fig. 2. Block diagram of recognition scheme.

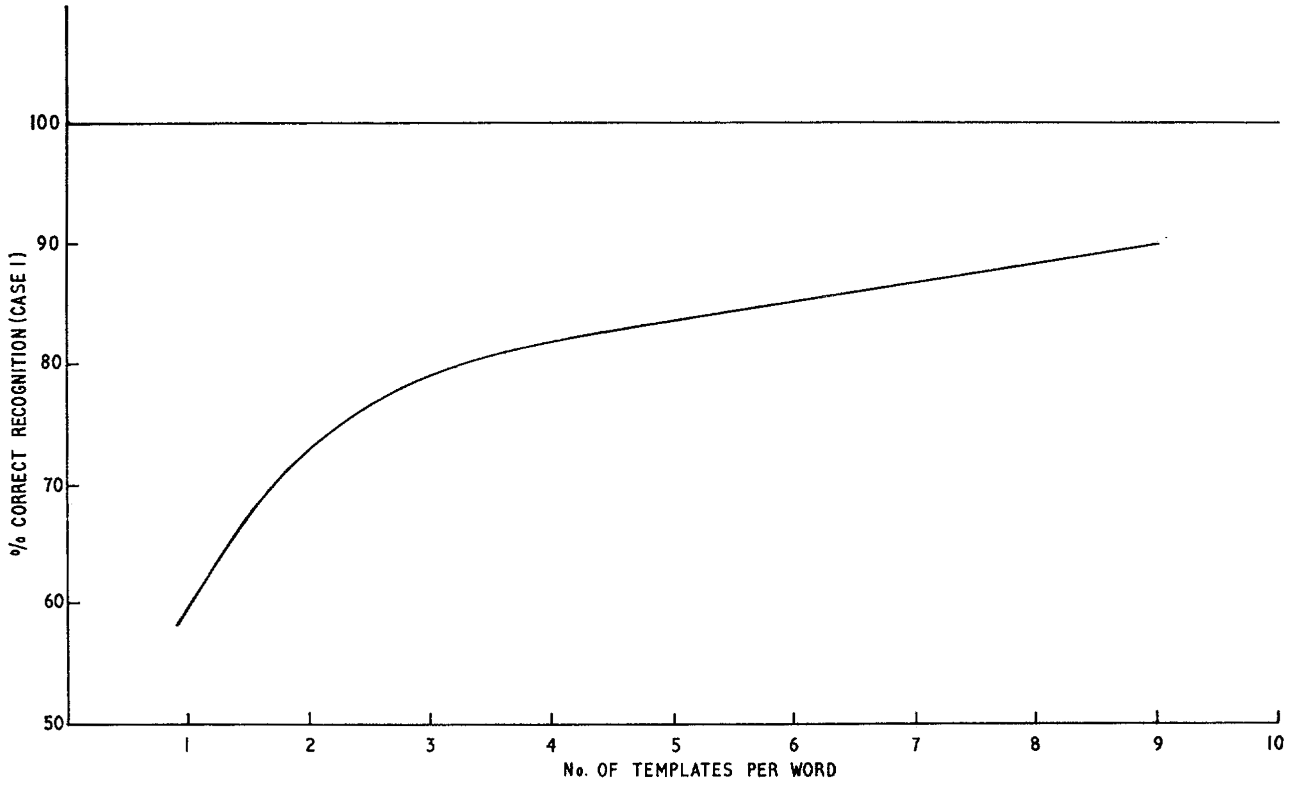

Fig. 3. Variation of percentage of correct recognition with number of templates per word.

VI. Conclusions

The type of pattern matching recognition scheme described in this paper turns out, on investigation, to present an almost endless variety of possible ways of spectrum coding, word matching, and scoring. A few of these configurations have been investigated experimentally, and some results are given, both for the case where the "unknown" word is restricted to the library vocabulary, and for the case where it is not.

The results suggest that a very coarse spectrum representation (a total of 21 classes including silence) is adequate, although it may be that the effects of poor spectral resolution are masked by other inadequacies.

Rather surprisingly, the technique of dealing with speaker variation by allowing each library word to be represented by several templates (derived from a population of talkers) shows signs of being very successful, with a reasonable number of templates per word. It seems possible, in fact, that the variation between members of a large population of speakers may be adequately dealt with by a fairly small number of templates per word; much further work is necessary to confirm this. Examination of the detailed results suggests that much of the deficiency of the system lies in its inability to cope with variation of the duration of spectral events constituting words. The use of a number of templates per word certainly helps to solve this problem, but it is unlikely to be effective enough where a word match depends on accurately matching a short-duration spectral event among other events of variable duration. Techniques for overcoming this problem are being studied. Meanwhile, for a vocabulary of 32 words from a population of 10 talkers, correct recognition of 90 percent of "unknown" words from the vocabulary is obtainable.

John N. Shearme was born in Warminster, England, in December, 1917.

He joined the British Post Office in 1937, and worked in the Dollis Hill Research Station until 1939, when he joined the Royal Corps of Signals. After service in the Middle and Far East, he returned to Dollis Hill in 1946, where he was concerned until 1953 with problems of speech voltage measurement and became interested in speech processing. During this period he attended the Northhampton Polytechnic Institute part-time and received the B.Sc. degree in engineering in 1951, from London University, England. In 1953, he joined the newly formed Joint Speech Research Unit (JSRU), Ruislip, Middlesex, England, to work on digital coding systems and speech processing. Later work included the development of improved formant and channel vocoders and the formulation, with co-workers, of a set of rules for computer-controlled speech synthesis. He is currently a Senior Principal Scientific Officer at JSRU, working on problems of analysis–synthesis telephony.

Mr. Shearme is a member of the Institution of Electrical Engineers (London).

Peter F. Leach was born in Abingdon, England, November, 1924.

He served with the Royal Air Force from 1943 to 1947, working on airborne radar. He joined the newly formed Joint Speech Research Unit in 1953, working on digital speech systems, and after studying part-time at the Northern Polytechnic Institute, obtained a Full Technological Certificate in telecommunications in 1956. He has since worked on analysis–synthesis telephony and pitch extraction problems, and on real-time continuous spectrum analysis. He is currently a Senior Experimental Officer working on automatic speech recognition.