BRITISH ACOUSTICAL SOCIETY.

"SPRING MEETING" at Chelsea College, London, S.W.3 on
Wednesday 25th April / Friday 27th April, 1973.

SPEECH AND HEARING: Session 'C': Speech Properties and Recognition.

Paper No:

73SHC4

Development of a System for the Automatic Recognition of Spoken Basic English.

P.D. Green and W.A. Ainsworth
Department of Communication, University of Keele, Keele, Staffordshire, England.

1. Introduction. During the last few years there have been a number of attempts to employ parameters derived by temporal analysis of speech waveforms as a basis for automatic recognition (1, 2, 3, 4). These experiments have demonstrated the usefulness of this approach, but all of this work has been directed towards the recognition of a small set of isolated words or syllables. The purpose of the present study is to investigate the automatic recognition of continuous speech.

This task is generally considered more difficult because (a) many phonemes are mispronounced or not pronounced, and (b) there are no reliable acoustic indicators of word boundaries. On the other hand, there are constraints of vocabulary and syntax which affect phoneme sequences and word order in English. Thus it might be possible to make use of this information to detect and correct errors introduced by mispronunciation. Dixon and Tappert (5) have reported work on this problem.

The scheme which is being developed consists of three sequential stages of processing. A parameter tracker takes the acoustic waveform as input and estimates the values of a set of formant frequencies and amplitudes by temporal analysis. An acoustic recogniser segments these continuous signals into phonemic units, and attempts to classify these units. Thirdly a linguistic processor segments this phoneme string into words, correcting the phonemic errors according to a predetermined error set, and generates a tree with the most likely spoken words at the nodes.

2. Speech parameter estimation. A hardware system has been developed which estimates the values of a set of speech parameters (6). The parameters are measured pitch-synchronously.

The first formant frequency, F1, is estimated from the duration of the first time-interval between zero-crossings after the start of each glottal cycle of the speech waveform low-pass filtered at 1000 Hz. The amplitude of the first formant, A1, is estimated on a logarithmic scale from the energy in the waveform in each single glottal cycle.

The frequency of the second formant, F2, is measured by counting the number of zero-crossings in the first 5 msec. of a waveform which has been band-pass filtered in the range 750-2500 Hz. and had the first formant removed by a programmable band-stop filter. The second formant amplitude, A2, is the energy in each glottal cycle of this wave.

The third formant frequency, F3, and amplitude, A3, are obtained in a similar manner. In this case the band-pass filter has

the range 1500-3500 Hz. The amplitude of the higher formants, A4, is the energy in the wave between 3,500 and 10,000 Hz.

3. Phoneme segmentation. The seven continuously varying parameters (F1, A1, F2, A2, F3, A3, A4) are sampled every 10 msec. by a 6 bit A-D converter and stored in the computer. An algorithm classifies each point in time as being part of a fricative, vowel, voiced consonant, or silent period.

A point is classified as fricative if A1 is less than a threshold and A2, A3 or A4 is greater than A1.

If the sum of the amplitudes is less than a threshold, the point is classified as silence. This test is made after the fricative test so that weak fricatives, provided they have an appropriate spectrum, will not be missed.

Vowels tend to have more energy than consonants, and this is most obvious in A2. Hence if A2 is greater than a threshold, the point is classified as vowel. Vowels which have a low F2, however, also tend to have a low A2, so a subsiduary rule lowers the threshold when F2 is low.

A second rule measures the differences between successive values of F1, A1, F2 and A2. A point is classified as vowel only if the sum of these differences is less than a threshold. This rule is necessary to split up long voiced sequences.

Any point which does not satisfy any of these tests is classified as voiced consonant.

The algorithm operates fairly successfully except that it does not segment two consecutive phonemes of the same type (e.g. /ts/or/ /ln/) and it occasionally causes part of a voiced consonant in a stressed syllable to be classified as vowel.

4. Phoneme recognition. Each sequence of similarly classified points constitutes a phoneme unit. These are treated differently depending on whether they are vowel, fricative or voiced consonant.

The F1-F2 space is divided into areas representing /i, I, ɛ, æ, ɑ, ɔ, ʊ, u, ɜ, ʌ/. If a unit is a vowel, the mean values of F1 and F2 are calculated for two segments near the beginning and end of the unit. Each of the two segments is then labelled as one of the above vowels according to the values of its F1 and F2. The pair of labels is then classified as one of the above vowels or one of the diphthongs /æi, æʊ, ɛə, ɛi, iə, ɔi, əʊ, uə/ according to a matrix stored in the computer.

The fricatives consist of the group /s, ʃ, z, k, t, d, tʃ, dʒ, f, θ, ts/. They are distinguished from each other chiefly on the basis of their duration and relative formant amplitudes, but the presence of a silent period immediately before the fricative helps to distinguish the stop consonants from the others. Different decision rules are employed depending upon whether the fricative occurs in initial, medial, or final position with respect to the vowels in the utterance.

The set of phonemes /b, d, g, p, ð, v, m, n, ŋ, w, r, l, j/ constitute the voiced consonants. The phoneme /p/ is included because it normally has its greatest concentration of energy in the first formant region of the spectrum. Measurement of mean values and slopes of the formant frequencies and amplitudes enables these consonants to be distinguished from each other.

Twenty five sentences were taken at random from newspapers and translated into Basic English (7, 8). The translations were read into the acoustic recognition system. Each phrase, P, produced a string of phonemes, T. A confusion matrix was constructed from the intended phoneme strings, P, and the received, T.

Only about 30% of the phonemes were correctly recognised, but this is about the same score as achieved by humans reading spectrograms of continuous speech (9).

5. Linguistic processor. The linguistic processor makes use of a phonetic dictionary of Basic English and an error set program. The dictionary contains about 3,000 separate entries and is stored on the disc of the computer. Searching the dictionary gives information about whether a string of phonemes is a valid word in the language, and whether it is a prefix of a longer word.

The error set program consists of:-

(a) A substitution operator which specifies the probability p (x/y) that a phoneme y in T was substituted for a phoneme x in P. Extra phoneme in T are dealt with by substituting a null phoneme.

(b) An insertion operator which specifies the probability p (z/x, y) that a phoneme z in P was deleted between phonemes x and y in T. (No insertion corresponds to the insertion of a null phoneme).

The combination of the dictionary, error set, and input string T implies a word tree W summarising the possible spoken phrases. Each node in W represents the different single words which could have been spoken starting from a given point in T. Each set of words which terminate at the same depth into T constitute a separate pathway from the node.

Off-line experiments using an assumed error set (10) have shown that for realistic error rates, W is far too complex to be fully evaluated. The task of the linguistic processor, therefore, is to develop and display a sub-tree of W, W', comprising the most likely spoken phrases.

Expansion of individual nodes in W' is performed by a breadth-first search procedure utilizing successive insertion and substitution operations. Successive pathways are formed by moving one place in T for every substitution operation. The prefix information from dictionary searches enables the lines of search to be curtailed as soon as possible. During each operation word candidates are examined in order of probability, and the operation is halted when this probability falls below an adjustable threshold, thus limiting the number of words found at a single node. Details of the pathways found are stored together with a measure of the likelihood of each pathway derived from the most probable words found in that pathway.

W' is developed in a depth-first manner, candidate nodes being accepted for expansion only if their associated likelihood measure compares favourably with those nodes already developed. Finally W' is 'pruned' to remove 'dead-end' pathways.

With an assumed error set in which three substitutions were allowed for each phoneme, together with naturally occurring sandhi and elision phenomena, readable word trees were evolved for phoneme error rates up to 60%.

With the error set generated by the Basic English sentences and the acoustic recognizer, the trees produced were more complex in many instances. An example of a word tree for a short phrase is shown in Figure 1. Work is in progress to reduce the complexity by syntactical constraints.

Acknowledgements. The work was supported by the Science Research Council.

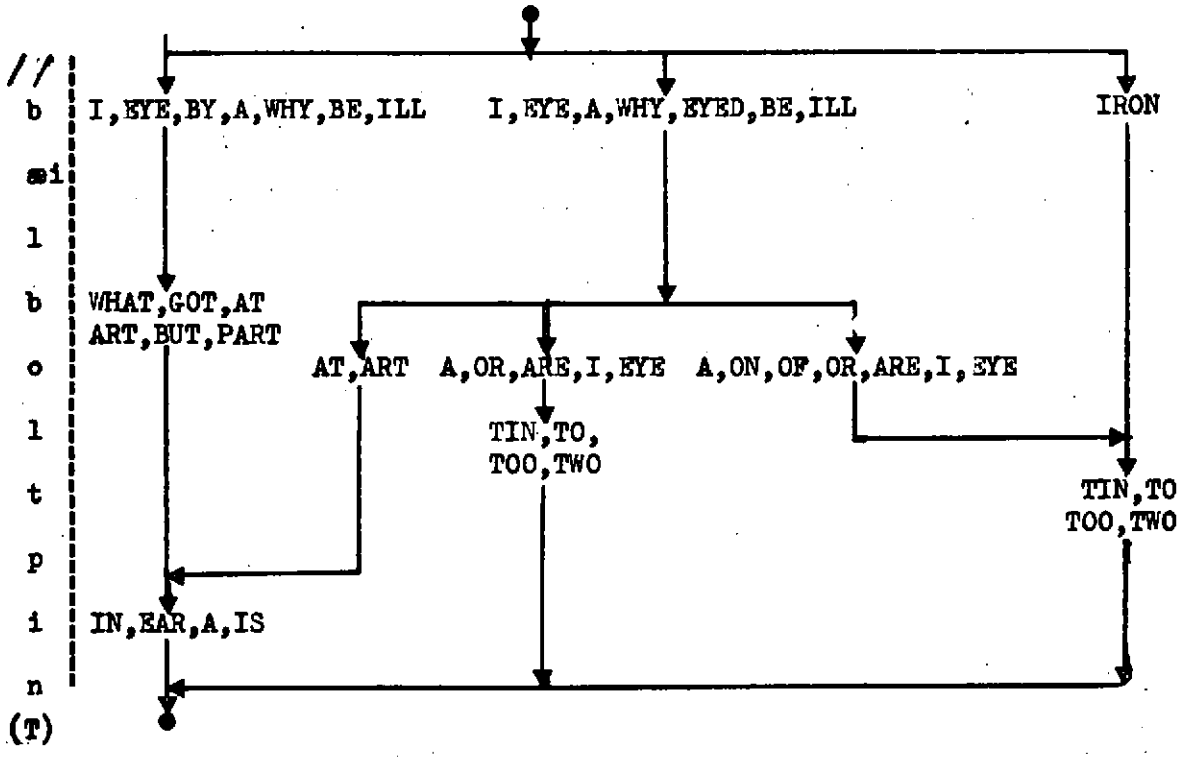

Figure 1: Word tree for the spoken phrase 'I got in'. Each group of words is ranked in order of probability. The input to the linguistic processor (T) is on the left.

References.

1. W. Bezdel and H.J. Chandler, Proc. Instn. Elec. Engrs., 112, p.2060, 1965.
2. S.H. Lavington and L.E. Rosenthal, Computer Journal, 9, p.330, 1967.
3. R.W.A. Scarr, IEEE Trans. Audio Electroacoustics, AU-16, p.247, 1968.
4. M.J. Underwood, T.R. Addis, and D.W. Boston, Machine Perception of Patterns and Pictures, Inst. Phys. Conf. Series No.13, p. 117, 1972.
5. N.R. Dixon and C.C. Tappert, Conf. on Speech Comm. and Processing, Boston, p.319, 1972.
6. W.A. Ainsworth, Int. J. Man-Machine Studies, 3, p.339, 1971.
7. C.K. Ogden, Basic English, Psyche Miniatures No.29, 1930.
8. C.K. Ogden, The Basic Words, Psyche Miniatures, No.44, 1932.
9. D.H. Klatt and K.N. Stevens, Conf. on Speech Comm. and Processing, Boston, p.315, 1972.
10. P.D. Green and W.A. Ainsworth, Machine Perception of Patterns and Pictures, Inst. Phys. Conf. Series No.13, p.161, 1972.