Proceedings of The Institute of Acoustics

TOWARDS SPEECH RECOGNITION BY MICROPROCESSOR

MRS M ABU EL-ATA, Dr. J SEYMOUR
SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING
THAMES POLYTECHNIC, LONDON SE18

Introduction

Although speech recognition systems are now available commercially their present cost is at a level where a substantial reduction is required before any widespread applications such as aids for the physically handicapped, can be realised. A reduction in computer costs can be achieved by using a microprocessor and minimal read-write memory is also indicated for the same reason. Due to the relatively slow speed of a microprocessor simplified analysis and recognition procedures are required, which must still be sufficiently reliable for practical purposes and yet allow speech processing in real time.

In the present work such a speech recognition system has been explored using a Minic minicomputer, which is an 8-bit machine operating at a similar speed to a microprocessor. Only those assembly code instructions likely to be found in microprocessors have been used and the system operates in real time.

Speech Processing

Many microprocessors have an 8-bit word length, so an initial decision was taken to limit the number of spectral coefficients to eight. One 8-bit word can then represent the frequency analysis for a particular time interval by indicating the presence or absence of energy for each coefficient. A word is derived from eight analogue levels which were obtained initially either from a Fast Fourier or a Fast Walsh Transform (1), and in order to account for the presence of formants each number is compared with its neighbour at a lower frequency (Fig. 1). The first number, corresponding to the lowest frequency, is compared with the mean value of a group of 16 input samples. A 1 is stored only if the first number is greater than the mean value, otherwise a 0 is stored. The second number is then compared with the first and a 1 is stored if the second number is the greater, otherwise a 0 is stored. The process is repeated up to the eighth coefficient and if any two numbers are equal the bit at the lower frequency is repeated. 16 such words are used to define a 1 second utterance, each one corresponding to a time slot of 64 ms. The bandwidth is from 200 Hz to 5 kHz so each coefficient covers a range of 600 Hz, as decided in a previous investigation (1). In this way a pattern of 128 cells is formed, which require only 16 storage locations for each utterance.

	Mean value	Frequency →								Computer word
Time ↓	015	020	013	027	052	023	031	037	026	10110110
	077	107	047	015	051	004	032	053	042	10010110
	061	077	041	003	002	000	023	032	021	10000110
	054	067	034	003	000	000	012	023	015	10000110

Fig. 1 Production of computer words for patterns.

Analysis Method

Walsh analysis reduces any given waveform to an orthogonal set of rectangular waves(2) and so produces a number of components for a single sine wave where Fourier analysis would produce only one. In an earlier investigation,

TOWARDS SPEECH RECOGNITION BY MICROPROCESSOR

IM Edwardes had used offline fast transform methods to compare Fourier and Walsh analysis of standard waveforms and speech (3). Over a bandwidth of 10 kHz Walsh speech analysis showed a greater number of formants than Fourier with some consonants better defined (4). This is also true when the bandwidth is reduced to 5 kHz and only 8 power coefficients are used as shown in Fig. 2 for the word 'seven'. Here sequency is half the total number of zero crossings in a given time interval and corresponds to frequency in the Fourier case.

Under these conditions Walsh analysis appears to give a more significant pattern than Fourier due to the larger number of 1's, representing formants. For a vocabulary of the numerals zero to nine uttered by two speakers there were about 50% more 1's in the Walsh than in the Fourier patterns and early recognition tests confirmed that better scores were obtained using the Walsh patterns (1).

Frequency	Sequency
00000010	00000111
00000010	00000110
00000010	00000110
10000000	10000011
10100000	10000010
10100000	10000010
10010000	10000100
10000000	10000110
10000000	10010000
10000000	10010011
10000000	10000100
00000000	10000000
00000000	00000000
00000000	00000000
00000000	00000000
00000000	00000000
Fourier	Walsh

Fig. 2 Pattern comparison for 'seven'

Walsh analysis also makes the use of a Fast Walsh Transform a practical possibility since only addition and subtraction is required. This can be implemented in hardware or software, while a Fast Fourier Transform requires complex multiplication which is expensive in hardware and time consuming in microprocessor software. At present a hardware Walsh analyser is being used, which produces coefficients in series for computer input and so requires no multiplexer.

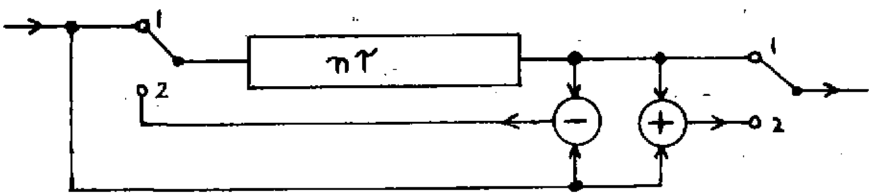
Walsh Analyser

Clark and Walker (5) have developed a logic cell for the transformation of television pictures, as shown in Fig. 3. The present analyser depends on a similar element, but uses analogue instead of digital signals. Four stages are required to produce 2^4 or 16 coefficients and they are connected in series with n = 8, 4, 2 and 1 proceeding from input to output. The delay $n\tau$ in Fig. 3 is implemented by a bucket-brigade analogue delay line, using discrete components, with a 2-phase 10 kHz clock which makes τ = 0.1 ms and the bandwidth 5 kHz. The adder and subtractor are normal opamp circuits and analogue switches are used at input and output.

The speech signal from a microphone or tape recorder is passed through a bandpass filter and when the filter output exceeds 0.1 V the analyser begins to sample it.

The way in which 16 input samples A to O, are processed by the first stage, n=8, is shown in Fig. 3. With the switches in position 1 the first 8 samples are delayed and added in order to the second 8 samples, with the pairs being passed to the second stage. At the same time ordered subtraction of the second from the first 8 samples is taking place and with the switch in position 2 these subtracted pairs are fed back into the delay line, to emerge immediately after the added pairs. In the second stage the samples are combined in fours, in the third in eights and in the final stage in sixteens to produce the 16 coefficients. At each stage the output is divided by 2, so that the final amplitude remains within the 2 V peak-to-peak limits of the input signal. Succeeding groups of 16 samples are analysed until the speech input has finished or the maximum analysis time of one second has elapsed.

TOWARDS SPEECH RECOGNITION BY MICROPROCESSOR




```
Switch              1                2               1
Input               A B C D E F G H  I J K L M N O P
Samples
                                     A B C D E F G H  A B C D E F G H
Output to 2nd stage                  + + + + + + + +  - - - - - - - -
                                     I J K L M N O P  I J K L M N O P
```

Fig. 3 Walsh-Hadamard Analyser Stage

The analyser is followed by a multiplier, which produces the square of each coefficient, and a logarithmic amplifier to emphasise those of low amplitude. The resulting levels are then passed into a simple 2-bit A-D converter for input to the computer. The analyser was tested with sine wave inputs, whose Walsh analysis had already been obtained by simulation on a large computer, and give an output showing good agreement with theory.

Computer Processing

The coefficients are classified as odd or even functions, called sal or cal respectively, which are analogous to sine and cosine in Fourier analysis. The coefficients are then a_o (corresponding to the mean value of the 16 samples), sal 1 to 8 and cal 1 to 7, the numbers denoting sequency. As coefficients are received from the anlyser they are routed to one of 9 locations, the first containing log $(a_o)^2$, the next log $(\text{sal } 1)^2$ + log $(\text{cal } 1)^2$ and so on up to the eighth location, with the ninth containing log $(\text{sal } 8)^2$. Since each complete transform lasts 1.6 ms 40 transforms are required to fill a 64 ms time slot, with each coefficient being added into the correct location. The computer inputs are binary representatives of the numbers 0, 1, 2 and 3 so the maximum number in any one location is 3 x 2 x 40 = 240, which is within the limit of 256 imposed by 8 bit words.

When the speech signal exceeds its threshold both the analyser clock and the computer input program are started. Any input during the first 1.6 ms after the threshold is exceeded is ignored since this time corresponds to the delay introduced by the analyser. The program runs for one second even if the speech concludes before that time; the last cells in the pattern being filled with zeros.

Master patterns with which incoming utterances can be compared are formed from five similar input utterances. At present an average analogue level is calculated from which 0's and 1's are produced by comparison of adjacent cells.

Recognition Method

In a learning phase the master pattern is produced for each utterance of the vocabulary, which comprises 10 utterances at present. The pattern of an incoming utterance is then compared with each master pattern in turn, using an exclusive - OR operation. This is carried out between each bit of corresponding 8-bit words in the input and each master pattern. The number of 1's produced by this operation then equals the number of differences between patterns, so the master pattern producing the lowest score is selected as the recognised utterance.

Recognition Test Results and Conclusions

Preliminary results have been obtained mainly with two speakers MAE and JS, but six other speakers have also been used. So far 19 trials have been made each with new master patterns of the numerals zero to nine. In four of these trials the utterances which were not recognised were repeated and the results are shown in Table 1. The initial score is given out of 10 with the number of times it occurred in each case. The final score refers to the repetition of unrecognised utterances and it may be seen that at the second attempt all the utterances were correctly recognised for three out of four speakers, with the initial score being increased in all cases.

Table 1 Recognition Test Results

	5	6	7	8	9	10	Initial score
MAE			1	3	2		Occurrence
				10			Final score
JS		1	3	1	1		Occurrence
			8				Final score
Six	2	1	2		1	1	Occurrence
others	10				10		Final score

At present successful operation of the system depends to some extent on the speaking skill of the user, which improves with experience. One reason for errors is the unequal duration of input and master patterns and this would be improved by time normalisation.

References

1. SEYMOUR J and GATWARD JF, A Microprocessor Approach to Speech Recognition, Acustica 37, 57-58, 1977.
2. BEAUCHAMP KG, Walsh Functions and their Applications, Academic Press, 1975.
3. EDWARDES IM, Feature Extraction for Speech Recognition using Fourier and Walsh Analysis, PhD thesis (CNAA), Thames Polytechnic (in preparation).
4. EDWARDES IM, and SEYMOUR J, Discrete Walsh Functions and Speech Recognition, IEE Colloquium, Hatfield.
5. CLARK CKP and WALKER R, Walsh-Hadamard Transformation of Television Pictures Applications of Walsh Function and Sequency Theory, IEEE Cat. No. 74CH0861-5EMS, 1974.