Proceedings of The Institute of Acoustics

COMPARATIVE ISOLATED WORD RECOGNITION EXPERIMENTS

Michael Robinson Taylor

Polytechnic of Central London (PCL), Signal Processing Group,
115 New Cavendish Street, London, U.K., in collaboration with
The Department of Health and Social Security (DHSS) U.K.

Now with: Smiths Industries Aerospace & Defence Systems,
Speech Technology Group, Cheltenham, Gloucestershire, U.K.

INTRODUCTION

This paper reports the comparative recognition performance of two isolated whole-word pattern matching recognition strategies.

The strategies were investigated as part of a project aimed at developing a low cost near real-time voice actuator, for use with environmental control equipment and other aids for the disabled, and have been described elsewhere by Taylor [1], [2] and [3].

Strategy A used a feature set consisting of the total number of speech frames J from an utterance, the first two major spectral energy peaks Q_1 and Q_2 per frame, and speech frame energy Q_3. Linear time normalisation was used in an attempt to overcome temporal variability.

Strategy B used a reduced feature set composed of the total number of speech frames J and the first two major spectral energy peaks Q_1 and Q_2 per frame. Time normalisation was based on a symmetric dynamic time warping (DTW) algorithm.

RECOGNISER ARCHITECTURE

Recognition Model

The recogniser was based on the well known isolated word recognition model shown in figure 1. The three major stages of this model include feature measurement, pattern similarity computation and the application of a decision rule.

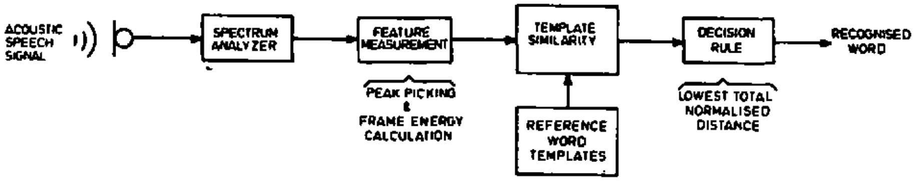

Figure 1. Isolated word recognition model

Concurrent Digital and Analogue Signal Processing

An 8 MHz 16/32 bit Motorola 68000 microprocessor was used to perform spectral energy calculation, normalisation, and distinctive feature extraction. This digital signal processing was performed concurrently with the analogue spectrum analyzer acquiring new frames of speech data.

ACOUSTIC ANALYSIS

Signal Conditioning

The acoustic speech signal was detected by a head-worn Shure Brothers SM-10 noise cancelling microphone and pre-emphasised at 6 dB per octave between 1 kHz and 7 kHz.

Spectrum Analyzer

The pre-emphasised signal was passed through a spectrum analyzer based on 16 contiguous 4th order analogue filters, covering the frequency range 200 Hz - 7 kHz. The centre frequencies and bandwidths of the filters were chosen in an attempt to resolve the first two formants of voiced speech and the two major spectral peaks applicable to voiceless fricatives, see [2] and [3].

Speech frames

The filter bank spectrum analyzer transformed time domain representations of the speech signal into time averaged frequency domain descriptions, known as frames. These consisted of 16 frequency related energy levels which approximated to short-time speech spectra averaged over a 10 ms time period T. The frames were read sequentially from the spectrum analyzer at a 100 Hz frame rate and transferred by the CPU to a circular memory speech buffer.

Utterance End-Point Detection

An adaptive backtracking utterance end-point detector based on the system proposed by Rabiner and Sambur [4] was used to locate utterance start and end frames. A complete description of the operation of the end-point detector may be found in reference [1].

TEMPLATE FEATURE SET

Utterance Length

The total number of speech frames J, located by the utterance end-point detector provided the coarsest parameter of the feature set for both recognition strategies.

Major Spectral Energy Peaks, Q_1 And Q_2

The 16 energy levels within each speech frame were reduced to a description consisting of only the two features Q_1 and Q_2. These approximated to either the first two formants of voiced speech or the frequencies of the two major spectral energy peaks of unvoiced speech. The voiced/unvoiced decision was based on a simple energy comparison between E_v, the average speech energy in the range 200 Hz - 2 kHz, and E_f, the average speech energy in the range 2.0 kHz - 7 kHz.

Each spectrum ordinate energy e_n was obtained by calculating the square value of each ordinate amplitude i_n, as suggested by Schafer and Rabiner [5].

The speech signal was treated as voiced if

$$\alpha(E_v) > E_f \tag{1}$$

where α was an empirically optimised scaling factor of 0.8.

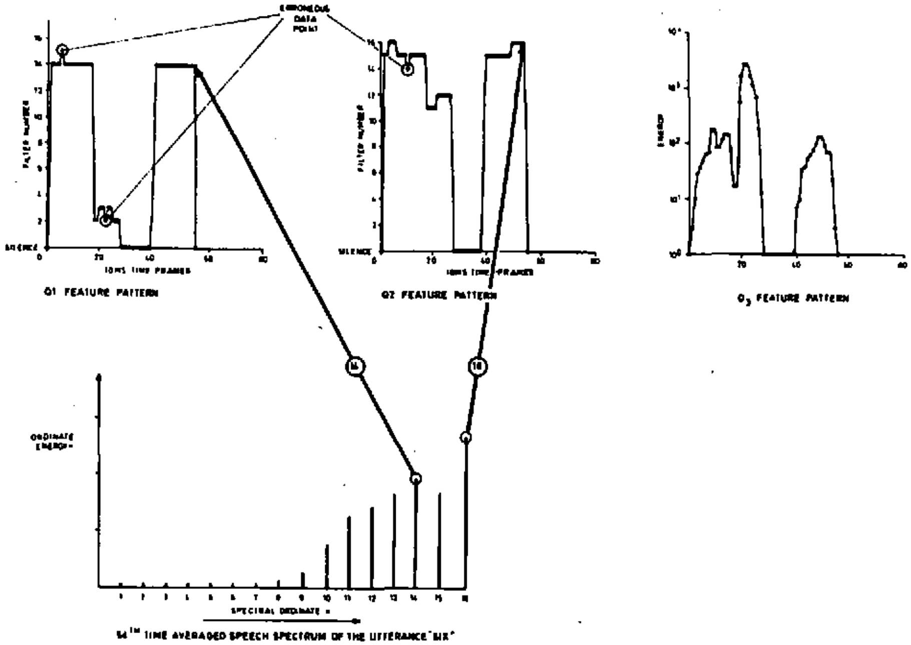

Figure 2. Construction of feature patterns

Utterance Frame Energy, Q_3

The normalised spectrum energy $\beta\varepsilon_j$ calculated for every frame within an utterance was used as feature Q_3 in recognition strategy A. $\beta\varepsilon_j$ was found by dividing each averaged spectrum frame energy level $\alpha\varepsilon_j$ by the mean energy value E of the complete utterance, as shown in equation (3).

The mean utterance energy value E was found by the summation of all frame energies $\alpha\varepsilon_j$, divided by the number of frames J, see equation (2).

$$E = \frac{1}{J} \sum_{j=1}^{J} (\alpha\varepsilon_j) = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{N} \sum_{n=1}^{N} e_{n_j} \qquad (2)$$

$1 > J > 130$
$N = 16$

$$\beta\varepsilon_j = \frac{\alpha\varepsilon_j}{E} = \frac{\frac{1}{N}\sum_{n=1}^{N} e_{n_j}}{\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{N}\sum_{n=1}^{N} e_{n_j}\right)} \qquad (3)$$

Reference Templates

Reference Templates for recognition strategy A were constructed from speech features so that a word of duration J X T seconds would be characterised by Q1 X J, Q2 X J and Q3 X J feature patterns.

Recognition strategy B used reference templates characterised by only Q1 X J and Q2 X J feature patterns.

Figure 2 shows the construction of typical feature patterns Q1 X J, Q2 X J and Q3 X J for the utterance 'SIX'. The figure also shows how the Q1 and Q2 features were extracted from an example frame of speech data.

Smoothing algorithm. The feature patterns Q1 and Q2 shown in figure 2 contain several erroneous data points subsequently removed by a median smoothing technique suggested by Rabiner et al [6].

TIME NORMALISATION

Linear Time Normalisation

A linear time normalisation algorithm was used in recognition strategy A as a compromise between robust normalisation and real time operation. The algorithm stretched the time axis of an utterance so that every feature template always contained J' data points, where J' was the maximum number of frames allowed for a valid utterance. The generation of additional data points was achieved by applying an interpolation technique between known frame values [1].

Non Linear Time Normalisation

A symmetric Dynamic Time Warping (DTW) algorithm was used in recognition strategy B which has been well described by Sakoe and Chiba [7]. The formulation for the dynamic programming calculation is the recursive expression reproduced here as equation (4). Further descriptions may also be found in [2] and [3]. Although improved recognition performance was expected with this algorithm it was also anticipated to be at the expense of real-time operation.

$$g(i,j) = \min \begin{bmatrix} g(i-1,j-2) + 2d(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-2,j-1) + 2d(i-1,j) + d(i,j) \end{bmatrix} \qquad (4)$$

RECOGNITION EXPERIMENTS

Two sets of recognition trials were conducted with four different vocabularies spoken by five adult male and five adult female naive speakers, in an acoustically noisy laboratory. None of the ten subjects chosen for the tests was allowed any exposure to the recognition system prior to the recognition tests, except for the time necessary for input amplifier gain adjustments. The background noise sound pressure level was 66 dBA with short-term acoustic noise transients of 79 dBA.

Lexicons

The vocabularies used in the trials consisted of three 10 word lexicons and a 30 word lexicon consisting of the other three lexicons, see Table 1.

Table 1. Word and syllable contents of the four tested lexicons

Lexicon 1		Lexicon 2		Lexicon 3	
Word	Syllable Content	Word	Syllable Content	Word	Syllable Content
Zero	2	Birmingham	3	Airconditioning	5
One	1	Manchester	3	Central Heating	4
Two	1	Aberdeen	3	Television	4
Three	1	London	2	Telephone	3
Four	1	Edinburgh	3	Down	1
Five	1	Newcastle	3	Off	1
Six	1	Leeds	1	Radio	3
Seven	2	Belfast	2	Lighting	2
Eight	1	Cardiff	2	On	1
Nine	1	Glasgow	2	Up	1
Average Syllable Content = 1.2		Average Syllable Content = 2.4		Average Syllable Content = 2.5	
Lexicon 4 Combines Lexicons 1, 2 & 3 and has an Average Syllable Content = 2.03					

Structure of Experiments

Training. The recogniser was trained by every subject in order to generate two templates for every word in each lexicon under test.

Recognition. The recognition procedure was conducted with the assistance of an observer who recorded the words recognised by the machine and ensured the tests were carried out in a consistent manner. During the performance tests the confidence and rejection thresholds within the recogniser were disabled in order to force a recognition decision.

Recognition Decision

The recognition decision rule was simply based on the Reference Template which produced the lowest normalised total distance score during the similarity computation.

Microphone position. The microphone used by the subjects was initially set to a fixed distance from the side of each subject's mouth using an expanded polystyrene slip gauge. During the performance tests the correct positioning of the microphone was regularly checked.

RESULTS

Confusion matrices

Tables 2 and 3 are the confusion matrices for the combined male and female subject groups speaking the isolated digits contained in lexicon 1, using recognition strategies A and B respectively.

Table 2. Recognition Strategy A. Confusion matrix obtained from 5 naive female and 5 naive male speakers using lexicon 1 (digits 0-9)

RECOGNISED WORD

SPOKEN WORD	ZERO	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	NUMBER OF OCCASIONS EACH SPOKEN WORD WAS CORRECTLY RECOGNISED (MAX POSSIBLE = 50)	PERCENTAGE RECOGNITION ACCURACY FOR EACH WORD
ZERO	41			.2		2	2			3	41	82
ONE		31	3		6		1	2		7	31	62
TWO		5	36	1	1		1	4		1	36	72
THREE	2	1		36	1	4				6	36	72
FOUR	1	9	4	1	29			2		4	29	58
FIVE	3	1		1	4	30	1			10	30	60
SIX	2		2		1	4	32	2	6	1	32	64
SEVEN	1	1	7		3			36		2	36	72
EIGHT	5		7		2	1			35		35	70
NINE	2	6			1	5	2			35	35	70

Table 3. Recognition Strategy B. Confusion matrix obtained from 5 naive female and 5 naive male speakers using lexicon 1 (digits 0-9)

RECOGNISED WORD

SPOKEN WORD	ZERO	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	NUMBER OF OCCASIONS EACH SPOKEN WORD WAS CORRECTLY RECOGNISED (MAX POSSIBLE = 50)	PERCENTAGE RECOGNITION ACCURACY FOR EACH WORD
ZERO	48							2			48	96
ONE		44			4					2	44	88
TWO	2		44	2	2						44	88
THREE				46			1	1		2	46	92
FOUR					48					2	48	96
FIVE						44	1			5	44	88
SIX							47		3		47	94
SEVEN	2			1			3	44			44	88
EIGHT		1						1	47	1	47	94
NINE				1		1				48	48	96

Quantitative Recognition Performance

From a comparative examination of matrices obtained using the two recognition strategies, it is clear that strategy B gives superior performance.

The wide spread of confusions shown in Table 2 are so scattered it is virtually impossible to estimate how such misrecognitions could have occurred. Some of the confusions shown in Table 3, however, are far more predictable, particularly where the 'Five' 'Nine' confusions occurred. For these digit confusions it may be assumed that the initial fricative /f/ from the digit 'Five' was lost in background noise; subsequent voiced frames, however, were found to match against those from the reference template 'Nine'. This template then returned the lowest overall distance penalty and was therefore recognised as the spoken word.

The differences in overall recognition accuracy may be seen by comparing the recognition results obtained using strategy A, tabulated in Table 4, with the results using strategy B, tabulated in Table 5. The differences in recognition accuracy for all speakers and lexicons are represented in Figure 3. It is shown that, for every speaker and lexicon, strategy B gave superior performance, the minimum difference being 10% and the maximum 46%. The average improvement in accuracy for all ten speakers and all four lexicons using strategy B was 26%.

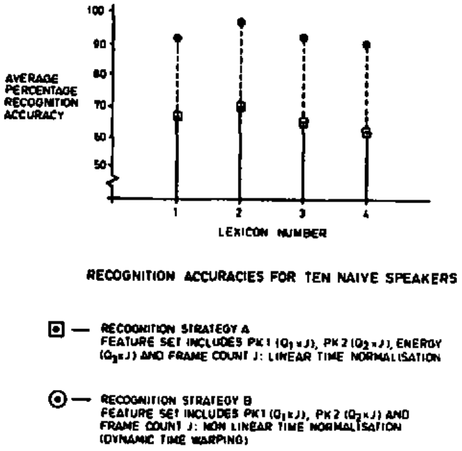

Figure 3. Comparative recognition accuracies

The increase in digit recognition using the DTW strategy, is worthy of further comment, as it is contrary to the findings of White and Neely [8]. In their experiments they found no difference in digit recognition accuracy when using either linear or DTW based recognition algorithms.

Post Utterance Processing Delays

The recognition strategy which used linear time normalisation gave a post utterance processing delay of no greater than 200 ms. The DTW based strategy, although more accurate, had a considerably extended post utterance delay of up to 3 s for the same 30 active word vocabulary.

Table 4. Strategy A Recognition scores for naive female and naive male speakers, recorded for every lexicon

Recognition accuracy scores in percentage (%)													
Speaker Female	Age Years	Lexicon Number 1	2	3	4	Average Recognition Per Speaker	Speaker Male	Age Years	Lexicon Number 1	2	3	4	Average Recognition Per Speaker
A	25	68	72	74	56	67.5	A	36	86	76	88	83	83.2
B	31	50	70	42	40	50.5	B	33	68	68	54	57	61.7
C	38	78	76	82	64	75	C	21	62	76	62	66	66.5
D	28	76	56	62	60	63.5	D	37	62	74	76	66	69.5
E	35	56	66	52	61	58.7	E	32	84	70	64	68	71.5
Average recognition accuracy = 65 68 62 56 per lexicon for female speakers							Average recognition accuracy = 72.4 72.8 68.8 68 per lexicon for male speakers						
Average recognition accuracy for female speakers = 63.04%							Average recognition accuracy for male speakers = 70.48%						

Table 5. Strategy B Recognition scores for naive female and naive male speakers, recorded for every lexicon

Recognition accuracy scores in percentage (%)													
Speaker Female	Age Years	Lexicon Number 1	2	3	4	Average Recognition Per Speaker	Speaker Male	Age Years	Lexicon Number 1	2	3	4	Average Recognition Per Speaker
A	26	96	96	88	91	92.7	A	37	96	100	98	93.3	96.8
B	32	90	94	99	67.3	84.8	B	34	88	98	96	88	92.5
C	39	94	96	92	92	93.5	C	22	94	98	90	94.6	94.1
D	29	92	92	92	92	92.0	D	38	90	100	94	93.3	94.3
E	35	86	96	90	92.6	91.1	E	33	94	100	94	96	96.0
Average recognition accuracy = 91.6,94.8,90, 87 per lexicon for female speakers							Average recognition accuracy = 92.4,99.2,94.4,93 per lexicon for male speakers						
Average recognition accuracy for female speakers = 90.8%							Average recognition accuracy for male speakers = 94.7%						

CONCLUSIONS

The results show that for recognition systems based on minimal distinctive features, dynamic time warping makes a far more significant contribution to recognition accuracy than an additional descriptive feature such as utterance energy. So significant is the increase in recognition performance that it may be concluded that even low cost systems should be designed with dynamic time warping, or similar techniques, before attempting to increase the number of descriptive features employed in the speech pattern. Although the extra computation load of a DTW algorithm will cause a significant increase in recognition transaction time, this may be reduced using variable frame rate data acquisition techniques.

ACKNOWLEDGEMENTS

Gratitude is expressed to the DHSS for supporting this research programme under Research Contracts RDV/88/60/02 - 80/82 and RDV/88/60/03 - 84/84; also to The Royal Jubilee Trust Fund and to Laryngograph Ltd., London.

Gratitude is also expressed to the subject groups who took part in the recognition assessment trials.

REFERENCES

[1] Taylor, M.R., 1984, "A microprocessor implementation of a speaker trained isolated word recognition system", MPhil thesis, published by University Microfilms International (UMI), Ann Arbor, Michigan, USA. Represented in Europe, Africa, the Middle East and Australasia by Information Publications International Limited, White Swan House, Godstone, Surrey, UK.

[2] Taylor, M.R., 1985, "Performance recognition tests on an improved voice actuator", Smiths Industries UK., Report No.RID.1903, prepared for the DHSS under Contract No. RDV/88/60/03-84/84.

[3] Taylor, M.R., 1986, "Isolated word recognition based on distinctive feature extraction and dynamic time warping". Proc. IEE Int. Conf. on Speech Input/Output; Techniques and Applications, London, UK., 24th - 26th March 1986.

[4] Rabiner, L.R., and Sambur M.R., 1975, "An algorithm for determining the endpoints of isolated utterances", The Bell System Technical Journal, Vol.54, No.297-315.

[5] Schaefer, R.W., and Rabiner, L.R., 1975, "Digital representations of speech signals", Proceedings of the IEEE, Vol.63, No.47, 662-676.

[6] Rabiner, L.R., Sambur, M.R., and Schmidt, C.E., 1975, "Application of a non-linear smoothing algorithm to speech processing". IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-23, 552-558.

[7] Sakoe, H. and Chiba, S., 1978, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-23, 67-72.

[8] White, G.M. and Neely, R.G., 1976, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering and Dynamic Programming", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-24, 183-188.