A COMPARISON OF THE PERFORMANCE OF WHOLE WORD HMM RECOGNISERS ON AN ALPHANUMERIC VOCABULARY.

Lynn C Wood, David J B Pearce

GEC-Marconi Limited, Hirst Research Centre,
East Lane, Wembley, Middlesex, HA9 7PP.

1. INTRODUCTION

This paper details an extensive set of isolated word recognition experiments using a hidden Markov model (HMM) recogniser with continuous probability distributions. The aim of this work has been to study and optimise the performance of the recogniser for both speaker-dependent and speaker-independent applications, and it brings together a number of successful techniques which have been reported elsewhere in the literature. The experiments using a common set of speech databases permit effective cross comparisons of the algorithms to be made. The main aspects that have been investigated are the choice of front-end parameters, the use of mixture densities, variance pooling schemes, the choice of model topology, the use of full covariance distributions and linear discriminant analysis (LDA). A number of novel comparisons, combinations and extensions to the reported methods have also been implemented which have lead to a greater understanding and improved performance. In particular the combination of LDA with mixture densities in an HMM framework has given the best performance for multi-speaker and speaker-dependent recognition.

The format of the paper is a follows: In section 2 a description of the two databases used in the performance comparison tests is given. The HMM training and recognition algorithms employed are described in section 3. Four acoustic front-ends were chosen for evaluation. A brief summary of each front-end and a comparison of its performance is given in section 4. Sections 5 to 8 describe the many extensions to the HMM which have been evaluated. Finally a summary of the report and conclusions are provided in section 9.

2. DATABASES

Two databases were used in the performance evaluations:

i) Alphanumeric database

> This is high quality speech data which is a subset of the APLAWD database [11] and consists of 10 repetitions of the alphanumeric vocabulary spoken by 10 speakers (5 male and 5 female). The speech data was first recorded in an anechoic room on a Sony Betamax PCM recorder, and subsequently digitised at 20 kHz with 12 bit resolution and down-sampled to 10 kHz. The alphanumeric vocabulary was chosen as it contains several highly confusable subsets e.g. the E-set as well as a set of phonetically distinct words, the digits, which are used in many applications. In the recognition evaluations, 5 repetitions of each word were used for training and five for evaluation. Two types of recognition performance are reported in this paper: (i) average speaker-dependent recognition [SD] (ii) multi-speaker recognition; most of the recognition runs are based on the 5 male speakers [MS-5] because a smaller number of speakers allowed a greater throughput of performance comparisons. To confirm these finding for a larger speaker inventory, a small number of 10 speaker runs were subsequently performed [MS-10].

ii) **Multi-Speaker telephone quality database**

This speech data forms a subset of a database provided by Marconi Speech and Information Systems (MSIS). The data was collected over the UK telephone network for a variety of channels and telephone handsets. The data used for the evaluations consisted of 2 repetitions of a 14 word vocabulary (the digits + "oh","cancel","stop","help") spoken by 64 male speakers. Closed speaker recognition was performed [MS-64] using one repetition of each word for training the multi-speaker models and one for evaluation. Subsequent recognition evaluations for an open speaker test, using a further 32 male speakers (not included in the training set) gave comparable performance with the closed speaker results. This suggests that the results reported here are indicative of speaker-independent performance.

The use of the two databases enables the behaviour of the recogniser to be observed over a range of operating conditions: i) speaker-dependent recognition of high quality speech on a difficult vocabulary, ii) multi-speaker operation on the same material as "i)" and iii) Multi-speaker recognition of telephone quality speech with an easy vocabulary.

The recognition evaluations on the two databases are summarized in tables 1-5. The recognition rate (%) and standard deviation across the speakers are given for each result.

3. HMM TRAINING AND RECOGNITION ALGORITHMS

The HMM model was a continuous probability emission model with 10 states and a left to right topology allowing skips over a single state. Unless otherwise stated a single Gaussian probability density function with a diagonal covariance matrix at each state is assumed. The HMM training method comprised of initial estimation of the HMM model parameters followed by reestimation of those parameters using the Baum-Welch algorithm. Isolated word recognition was performed using a Viterbi algorithm employing beam-clipping and a log Gaussian distance metric.

4. FRONT-END PARAMETERISATION

Four acoustic front-ends were used in the performance comparisons:

i) LPC Cepstrum: The cepstral coefficients were derived from an 8th order linear predictive analysis of the short-time windowed speech signal using the autocorrelation method [4].

ii) MEL-Cepstrum: This was obtained by a cosine transformation of the real logarithm of the short-term energy FFT spectrum expressed on a mel scale [5] using a bank of 20 triangular filters.

iii) Filterbank: The filterbank front-end was based on the RSRE standard equations [6]. At 5 kHz bandwidth the number of filter channels (N) was 23 arranged on a non-linear frequency scale. The first filter was a total energy measure above 60 Hz. Filters 1 to N-1 were 4th order Butterworth chosen to be non-overlapping and the top filter was a high-pass filter.

iv) Bark Warped Cepstrum: This method combines the all-pole modelling of the LPC analysis with the critical band spacing [7]. The technique has practical advantages over some other perceptually based processing techniques. Computationally it approaches the efficiency of the standard linear predictive analysis and it can be directly substituted for LP analysis in speech recognition systems. The Bark cepstral parameters were generated by first obtaining predictor coefficients using the Burg Lattice method [8] which was altered to provide Bark frequency warping. The lattice method differs from the autocorrelation and covariance methods in that the predictor coefficients are obtained directly from the speech samples without an intermediate calculation of a correlation function.

A pre-emphasis filter $(1\text{-}0.95z^{-1})$ was applied to the speech data prior to performing the front-end analysis. For the cepstral front-ends a 20 ms Hamming window was applied every 10 ms. 12 cepstral coefficients were obtained plus an energy term (C_0). Each acoustic front-end vector was augmented with its time derivative computed as the difference between two frames spaced 40 ms apart.

Automatic word end-point detection was performed to remove silence at the start and end of each word using an energy thresholding method; no hand-labelling of the data was performed.

Table 1 summarises the performance of the four acoustic front-ends for the alphanumeric and telephone quality databases. A comparison of the three cepstral front-ends shows that the linear frequency spacing technique (LPC cepstrum) obtains the best performance in the speaker-dependent tests, however, the non-linear techniques perform better for the multi-speaker tests. This result suggests that the non-linear techniques are more robust in modelling speaker variation whereas the linear technique provides improved acoustic discrimination on a per-speaker basis. The poor performance of the filterbank front-end compared to the cepstral front-ends may be due to the larger number of free parameters (46 filter channels including time derivatives, compared to 26 cepstral parameters) which could lead to undertraining on limited data. Other factors which may account for the difference are the validity of the assumption of diagonal variances and the spectral smoothing inherent in the reduced parameter cepstral representation.

The subsequent recognition experiments were performed using the LPC and MEL cepstral front-ends since they achieved the best performance for the speaker-dependent and multi-speaker tests respectively.

5. VARIANCE POOLING

The use of fixed or pooled variances over all states and word models have been found to provide superior results over the use of individual nodal variances. Examples of such schemes include the computation of a "grand" feature vector [2] and the application of empirical weighting functions (e.g. quefrency weighting) which attempt to approximate the statistically derived within-class weights [9]. The superior results suggest that the variations in the training set are not sufficiently great to adequately cover the variations in the recognition set. The use of variance pooling reduces the number of free parameters in the system and therefore reduces the problem of undertraining. The disadvantage of pooling is that states which correspond to sounds which may have quite different second order statistics are averaged together. There is thus a trade-off between the amount of training and the type of pooling which results in the best performance.

Variance pooling was performed within the Baum-Welch reestimation procedures by accumulating the partial scores across states. Two pooled variance estimates were obtained:

(i) A "grand" variance pooled over all states and all words [2].

(ii) A word-dependent pooled variance.

Table 2 summarises the results of the two variance pooling methods compared to the case where nodal variances are allowed. These experiments clearly demonstrate the trade-off between the number of free parameters which can be adequately trained and the size of the available training data. For the SD tests, the "grand" variance obtained the best performance (97.1% grand c.f. 95.5% nodal for the mel-cepstrum front-end). For the MS-5 tests, where a larger number of tokens were used to train each word HMM, the "grand" variance gave the lowest performance (94.3% for the "grand" variance compared to 94.7% with nodal variances). The superiority of the word-dependent variance in this case is probably because the distributions are more sound specific since many of the confusable words (e.g. E-set) have only a small number of phones per word. Finally in the MS-64 case, where 64 repetitions of each word were used for training the HMMs, the use of nodal variances obtained the best performance (97.0% grand c.f. 98.8% nodal).

It is interesting to note that for the speaker-dependent runs, the advantage of the LPC-cepstrum over the Mel cepstrum reported in section 4 with nodal variances is not maintained when using variance pooling.

6. STATE DISTRIBUTIONS - MIXTURE DENSITY HMM

Mixture densities have been applied to HMM speech recognition by Bell Laboratories [1]. (Another type of mixture, called the Richter mixture, has also been used by IBM [10] but is not considered further here). In a mixture model a single gaussian probability distribution at each state is replaced by a set of gaussians and the output pdf at each state becomes a weighted summation of the gaussian mixtures. The advantage of the mixture approach is that i) it attempts to improve the modelling of outliers in a distribution since the outliers are more likely to be closer to a mixture distribution than a single gaussian, ii) it improves the modelling of multi-modal distributions which typically occur over a range of speakers, and iii) The mixture model can approximate other (non-gaussian) pdfs and covariations.

A mixture density HMM similar to the Bell Laboratories approach was implemented where the mixture centroids were allowed to be different and the covariance matrix for each mixture was constrained to be nodal and diagonal. The initial estimates of the gaussian mixtures were obtained by boot-strapping from a single Gaussian model using a k-means training procedure [1].

Table 3 summarises the performance of the mixture density modelling. For the MS-5 runs, the mixture model consistently obtains better recognition performance than the single Gaussian model. This was also observed for the MS-10 runs. Comparison between the mixture HMM and a single gaussian full covariance model indicates that the mixture model is more effective at modelling the state distributions which occur for a number of speakers. The use of mixtures in the speaker-dependent tests, however, deteriorated performance; the likely causes are the increase in the number of free parameters to be trained on limited data and the fragmentation of the state distributions.

7. TRANSFORMATIONS

Linear transformations are used to convert speech front-end parameters to a reduced representation while preserving much of the information in the original spectrum. The motivation is three-fold: i) to obtain transformed speech feature vectors which are uncorrelated and have a unit variance so that a Euclidean metric is valid in the transformed space. ii) a reduction in the number of parameters can be achieved by eliminating the less reliable features. On limited training data the performance can actually be better with the relevant set of dimensions since only the directions where statistical noise has least effect are considered. iii) The storage and computational requirements of the HMM recogniser are reduced due to the reduced front-end representation.

The linear discriminant transform uses the directions in parameter space which maximise the ratio of the between-class to the within-class statistics. An example of this approach is the IMELDA transform [3][12] which combines linear discriminant analysis (LDA) and a MEL-scale representation and has been shown to improve the robustness of speech recognisers for a wide range of distortions.

Table 4 shows the effect of applying a LDA transformation for the two databases. After the LDA transform was applied to the front-end parameters the HMM models were trained using diagonal covariances pooled over all words and states. Attempts were made to optimize the transformation for the number of feature elements remaining at each stage of the transformation computation. This optimization proved inconclusive since the best performance was obtained for a variety of combinations.

Comparison between the LDA and a full covariance HMM model is a useful indicator of the effectiveness of feature selection. Considering the multi-speaker results with a mel-cepstrum front-end, the results show that the LDA transform was generally 0.5-1.0% better that the pooled covariance model. Some of the improvement using the LDA transform therefore arises from the modelling of the pooled covariance and a further improvement from the use of discriminant analysis to remove detrimental vector directions.

Table 4 also shows a comparison between a "grand" full covariance HMM and a "grand" diagonal covariance HMM. The improvement obtained from the use of a full covariance model is much greater in the multi-speaker experiments than for the SD case. This result indicates that the full covariance model can to some extent compensate for the multi-modality in the distributions.

Although the LDA transformation obtained improved performance, the mixture density approach proved to be more effective in modelling multi-speaker distributions. A combined mixture and LDA model was therefore proposed with the advantages of improved modelling provided by the mixture model, and the modelling of correlation, and parameter reduction achieved by the LDA transform. The results obtained for the combined model, shown in table 4, were the best multi-speaker recognition results obtained for the 10 state HMM model (98% for the alphanumeric database and 99.5% for the telephone quality database).

The LDA transform has also been used with the filterbank front-end. The improvement in performance obtained with the transform was much greater than with the cepstral front-end because the features are more correlated (For the MS-64 case: 90.0% for filterbank only c.f. 98.5% filterbank with transform). With the transform applied, the filterbank front-end obtains a similar performance to the cepstral front-end.

8. MODEL TOPOLOGY

Research has shown that recognition performance can be improved when the number of states is related to the duration of the word [13]. Table 5 summarises a set of performance comparisons between a 10 state HMM and a variable state model where the number of states is set equal to half the average frame duration of each word minus one standard deviation (the average number of states in this case was 17). The larger average number of states per word results in greater model "resolution". The results show that without a LDA transform applied the variable state HMM performs significantly better than a 10 state model (98.7% compared to 97.7% for three mixtures). When a LDA transform is applied however, the performance is worse (95.1% compared to 96.1%). A possible explanation for this result is that when the number of states is large, the long duration sounds (e.g. the vowels) will contribute more to the computation of the pooled transform since they will be mapped onto a larger number of states. As a result the transform is biased towards the distribution of the longer duration sounds which may not be pertinent to the discrimination between the words of the vocabulary.

9.DISCUSSION AND CONCLUSIONS

A set of isolated word recognition experiments using a hidden Markov model recogniser have been detailed. From all these experiments two general principles have been found to be important in achieving good performance from the maximum likelihood recogniser: (i) the use of appropriate statistical models to match the distributions that occur in the data (ii) the ability to train the model parameters from a limited amount of training data. As can be observed in the results presented here there is often a trade-off between these two requirements. The best performance is achieved with the right combination of choice of distribution to match the form of the data and the choice of constraint on the number of free parameters which can be adequately trained on the data available. For a fixed number of states the best performance was obtained with a novel combination of Mel cepstrum front-end, a transform based on linear discriminant analysis and mixture density distributions. An analysis of the errors remaining for the alphanumeric database shows that half are due to significant end-point errors. While it is difficult to make comparisons with results reported elsewhere on different databases, those presented here are among the best reported on comparable vocabularies and conditions e.g. [14][15].

ACKNOWLEDGEMENTS

This work was conducted for Marconi Speech and Information Systems (MSIS) on the Alvey Integrated Speech Technology Demonstrator Project (ISTD) with CSTR and HUSAT. We would like to thank Melvyn Hunt and other MSIS staff for discussions on the use of linear discriminant analysis.

REFERENCES

[1] L.R Rabiner et Al, 'Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities', AT&T Technical Journal, Vol 64, No.6, July 1985.

[2] D. B Paul, 'A Speaker Stress Resistant HMM Isolated Word Recogniser', IEEE ICASSP 1987, pp713-716

[3] M J Hunt & C Lefebvre,'A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech', IEEE ICASSP 1989,pp262-265.

[4] J Makhoul,'Linear Prediction: A Tutorial Review', Proc IEEE, Vol 63, pp 561-580, 1975.

[5] S B Davis & P Mermelstein, 'Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences',IEEE Trans., ASSP, August 1980.

[6] D Knox, 'SRUBANK Filterbank Design Program - Users manual (V1.03)', 1987 by ENSIGMA LTD for SRU at RSRE, Malvern.

[7] K Frimpong-Ansah, 'A Stochastic Feature Based Recogniser and its Training Algorithm', IEEE ICASSP 1989, pp 401-404.

[8] J P Burg,"A New Technique for time series data", NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics", 1968.

[9] K K Paliwal,"On the Performance of the Quefrency Weighted Cepstral Coefficients in Vowel Recognition", Speech Communication, pp151-154, May 1982.

[10] P F Brown,"The Acoustic-Modelling Problem in Automatic Speech Recognition", Phd Thesis, Carnegie-Mellon University, 1987.

[11] G Lindsey et Al, "SPAR'S Archivable Actual-Word Databases", Alvey SPAR Project Internal Document, June 1987.

[12] M J Hunt, S M Richardson, M G Abbott,"Use of Linear Discriminant Analysis in Isolated and Continuous Speech Recognition Experiments", in these proceedings.

[13] J Picone,"On modelling Duration in Context in Speech Recognition", IEEE ICASSP 90, Vol 1

[14] L R Rabiner & J G Wilpon, "Some performance benchmarks for isolated word speech recognition systems", Computer Speech & Language, Vol 2 No 3/4 Sept/Dec 1987

[15] C-H Lee, "On the use of some robust modeling techniques for speech recognition", Computer Speech and Language Vol 3 No 1 Jan 89.

TABLE 1 - FRONT END PARAMETERIZATIONS

	HIGH QUALITY ALPHANUMERICS		TELEPHONE QUALITY 14 WORDS
FRONT-END	SD %	MS-5 %	MS-64 %
LPC-CEPSTRUM	96.5(1.6)	92.4(3.1)	-
MEL-CEPSTRUM	95.5(1.7)	94.7(3.2)	98.8(2.5)
FILTER BANK	90.7(4.6)	92.2(2.6)	-
BARK WARPED CEPSTRUM	95.6(1.5)	93.2(2.9)	-

TABLE 2 - VARIANCE POOLING

		HIGH QUALITY ALPHANUMERICS		TELEPHONE QUALITY 14 WORDS
FRONT-END	VARIANCE POOLING	SD %	MS-5 %	MS-64 %
LPC-CEPSTRUM	NODAL	96.5(1.6)	92.4(3.1)	
	WORD DEPENDENT	96.2(1.8)	92.9(3.5)	
	"GRAND" VARIANCES	96.5(2.0)	87.7(3.3)	
	"GRAND" COVARIANCES	96.4(1.7)	93.7(3.9)	
MEL-CEPSTRUM	NODAL	95.5(1.7)	94.7(3.2)	98.8(2.5)
	WORD-DEPENDENT	96.8(1.3)	95.1(1.5)	97.4(3.5)
	"GRAND" VARIANCES	97.1(1.0)	94.3(3.4)	97.0(4.1)
	"GRAND" COVARIANCES	97.2(1.3)	94.9(2.6)	97.9(3.6)

TABLE 3 - STATE DISTRIBUTIONS

FRONT-END	STATE DISTRIBUTION	HIGH QUALITY ALPHANUMERICS			TELEPHONE QUALITY 14 WORDS
		SD %	MS-5 %	MS-10 %	MS-64 %
LPC-CEPSTRUM	SINGLE GAUSSIAN (nodal diagonal covariance)	96.5(1.6)	92.4(3.1)	88.3(7.0)	
	2 MIXTURES	95.2(1.5)	94.3(3.6)	93.7(3.5)	
	3 MIXTURES	-	95.0(3.4)	94.8(2.7)	
	4 MIXTURES	-	95.5(2.8)	95.1(2.6)	
	SINGLE GAUSSIAN (pooled full covariance)	96.4(1.7)	93.7(3.2)	-	
MEL-CEPSTRUM	SINGLE GAUSSIAN (nodal diagonal covariance)	95.5(1.6)	94.7(3.2)	92.1(5.6)	98.9(1.3)
	2 MIXTURES	95.0(2.4)	96.9(1.4)	95.6(2.3)	-
	3 MIXTURES	-	97.7(1.3)	96.8(1.1)	99.3(1.7)
	4 MIXTURES	-	97.3(2.5)	97.0(1.0)	-
	SINGLE GAUSSIAN (pooled full covariance)	97.2(1.3)	94.9(2.6)	-	97.9(3.6)

TABLE 4 - TRANSFORMATIONS

FRONT-END	COMMENT	HIGH QUALITY ALPHANUMERICS		TELEPHONE QUALITY 14 WORDS
		SD %	MS-5 %	MS-64 %
LPC-CEPSTRUM	"GRAND" VARIANCE	96.5(2.0)	87.7(3.3)	
	"GRAND" COVARIANCE	96.4(1.7)	93.7(3.9)	
	LDA TRANSFORM	96.5(1.7)	94.3(3.5)	
	3 GAUSSIAN MIXTURES	-	95.0(3.4)	
	LDA + 3 MIXTURES	-	96.0(3.0)	
MEL-CEPSTRUM	"GRAND" VARIANCE	97.1(1.0)	94.3(3.4)	97.0(4.1)
	"GRAND" COVARIANCE	97.2(1.3)	94.9(2.6)	98.0(3.6)
	LDA TRANSFORM	97.4(1.2)	96.1(2.6)	98.7(3.3)
	3 GAUSSIAN MIXTURES	-	97.7(1.3)	99.3(1.7)
	LDA + 3 MIXTURES	-	98.0(1.2)	99.5(1.9)

TABLE 5 - MODEL TOPOLOGY

FRONT-END	MODEL TOPOLOGY	HIGH QUALITY ALPHANUMERICS		
		SD %	MS-5 %	MS-10 %
MEL-CEPSTRUM	10 STATES (10S)	96.5(1.7)	94.7(3.2)	92.1(5.6)
	VARIABLE STATES (VS)	95.1(2.1)	96.3(2.1)	93.2(4.3)
	10S + 3 MIXTURES	-	97.7(1.3)	96.8(1.1)
	VS + 3 MIXTURES	-	98.7(0.9)	97.9(1.1)
	10S + IMELDA	97.4(1.2)	96.1(2.6)	
	VS + IMELDA	-	95.1(2.7)	