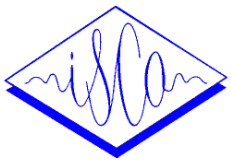

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# ESTIMATING LINEAR DISCRIMINANT PARAMETERS FOR CONTINUOUS DENSITY HIDDEN MARKOV MODELS.

*Eluned S. Parris and Michael J. Carey*

Ensigma Ltd.
Turing House, Station Road, Chepstow, Gwent, U.K.

## ABSTRACT

This paper describes a new technique for performing Linear Discriminant Analysis (LDA) on Hidden Markov Models(HMMs) incorporating state specific mixture densities. Previous work with LDA in speech recognition has used models comprising a unimodal multivariate Gaussian density per state or semi-continuous models using tied densities across states. As the number of models in the mixture densities in a HMM are increased and speech frames are mapped into the new feature space the LDA does not produce the expected unit variance distributions. By treating each state as a single class the assumption that a state can be described as a single multivariate Gaussian is violated. A new technique has been developed which maintains the mapping into the LDA feature space and constructs new HMMs directly from the transformed speech frames, considerably reducing the computation required. An 11% reduction in error rate has been achieved over the standard LDA technique for monophone HMMs and 25% of the system parameters can be discarded without loss in performance.

## 1. INTRODUCTION

Linear Discriminant Analysis (LDA) first proposed by Fisher [1] is a technique frequently used in pattern classification to provide an improved feature set. LDA is a method of transforming and scaling variables to improve the performance of a classification system. It shows which linear combinations of features are most useful in classification and can also reduce the amount of computation and storage required by reducing the number of parameters in the system. LDA was first successfully applied to speech recognition by Hunt [2] in Independent Mel-scale Linear Discriminant analysis (IMELDA.) Each state in a set of HMMs is treated as a separate class and the transform is derived from measures of within class and between class separability. The IMELDA transform produces a set of features ordered with respect to their discriminative ability. The least discriminative features are often noisy and can be discarded without loss in performance. The transform gives rise to a grand covariance matrix equal to the identity matrix producing unit variances for each parameter within a state. A simple Euclidean distance metric can then be used giving reductions in both computation and storage. This technique has given substantial improvements in sub-word recognition [3,4,5] and has also been successfully adapted by Ayer [6] to do discriminative whole-word recognition.

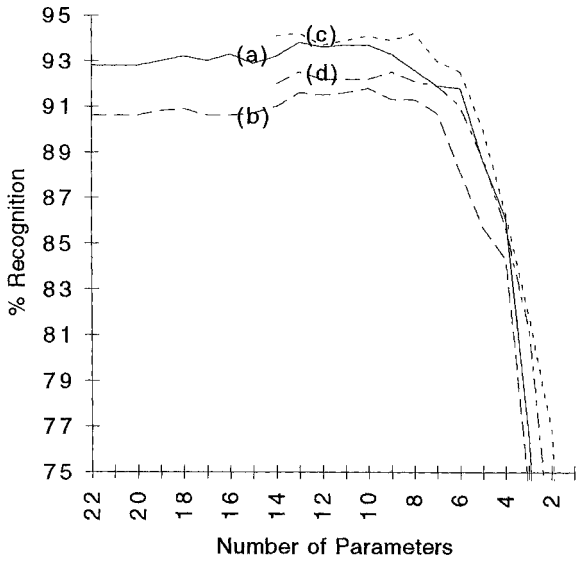


**Figure 1** Results of a Digit Recognition Experiment.
a) Filterbank Input Parameters - State Variances
b) Filterbank Input Parameters - Unit Variances
c) Mel Cepstra Input Parameters -State Variances
d) Mel Cepstra Input Parameters - Unit Variances

A large number of current speech recognition systems use a discrete cosine transform to compute cepstral coefficients. This approximately orthogonalises the feature space and a diagonal covariance matrix is then assumed. This has been shown to give better results than using a feature set comprising of filter bank outputs or linear predictive coefficients. Although the off diagonal elements of the covariance matrix are close to zero for the cepstra indicating that the cosine transform is producing an orthogonal set of features. When the feature set is expanded to include transitional cepstra, correlations between the features are introduced and the off diagonal elements of the covariance matrix become more important. The use of full covariance matrices is computationally intensive and state specific full covariance matrices require large amount of training data. They are therefore rarely used. In contrast the use of LDA to produce a discriminative set of features with a diagonal covariance matrix is simple to implement and gives worthwhile reductions in storage.

As an example an experiment was carried out using LDA on an isolated digit recognition task using LDA firstly on filterbank outputs and then on cepstra. In each case the classes for LDA were the states of the HMM digit models consisting of continuous density multivariate Gaussians. Figure 1 shows the recognition results for both feature sets as the number of parameters used for recognition was reduced. We observe that the recognition rate initially improves as noisy parameters are discarded as the parameter set decreases. While further parameter reduction degrades the performance it maintains a high level with as few as eight parameters.

Repooling of the training data onto the LDA classes, HMM states, was needed for the filterbank feature set to achieve the best results whilst there was no gain in performance for the cepstra when the data was repooled. The consistent mapping of frames to states prior to the calculation of the transform is a crucial factor in LDA. The initial HMMs derived from the cepstra are already good and therefore there is little movement between the frames and states when repooling is done. The HMMs derived from filterbank outputs gave recognition errors three times greater than the cepstral HMMs and therefore the mapping of frames to states is less consistent. Assuming unit variances reduces the performance as Figure 1 shows. Examination of the state variances shows that they vary over the range 0.6-1.5. In contrast the state variances for the cepstra varied by almost two orders of magnitude.

For these digit experiments and the other work on LDA already described, the speech models either comprise one Gaussian density per state or are semi-continuous models using tied densities across states. Recent results indicate that superior recognition results are achieved by systems using multimodal continuous density HMMs [7,8]. This paper describes a method of incorporating mixture densities into LDA. Section 2 discusses these problems relating to the theory of LDA and describes the new technique using mixture based pooling. Section 3 describes the databases used and experiments performed using the technique. Section 4 presents the conclusions of this work.

## 2. THEORY

### 2.1 Continuous Density Hidden Markov Models

Continuous density HMMs have been successfully used in speech processing for a number of years. The output distribution of an observation $\mathbf{y}_t$ in state j at time t is given by

$$b_j(\mathbf{y}_t) = \sum_{m=1}^{M} c_{jm} N(\mathbf{y}_t; \mu_{jm}, \sigma_{jm}^2)$$

where N denotes a Gaussian mixture with mean $\mu_{jm}$ and variance $\sigma_{jm}^2$ i.e. diagonal covariance matrix is assumed, and $c_{jm}$ are the mixture weights.

It has been shown that the recognition performance increases monotonically as more mixture densities are used per state, provided that there is enough training data [9].

### 2.2 Linear Discriminant Analysis

LDA involves finding a linear transformation which combines features together to optimise class separability. In speech recognition, IMELDA is implemented by taking each state of each HMM as a separate class. The training data is mapped to the closest class using Viterbi alignment. Each class then has an associated set of speech frames called a pool. LDA makes the assumption that the set of samples from any class have the same multivariate Gaussian distribution about their own class centroid. A measure of class separability

$$J = tr(W^{-1}B)$$

is optimised, where W is the pooled within class covariance matrix and B is the between class covariance matrix calculated from the class centroids. The first stage of LDA performs a linear transformation on the parameters describing the samples to turn W into an identity matrix. This transformation consists of a rotation and scaling which averages the within-class variance to unity in all directions. The F-ratio is a measure of discrimination between classes and is the ratio of the means of the different classes to the average variance of samples within their classes. After the first stage of LDA, the F-ratio is equal to the variance of the class centroids projected onto that direction. The second stage of LDA searches the space to find the direction with maximum F-ratio i.e. the direction for which the variance of class centroids is maximum. The subspace orthogonal to this is then searched to find the next largest F-ratio and so on until the last direction has the minimum F-ratio. This operation can be achieved by a single rotation. The two stages of LDA can be represented as a single matrix operation. The transform produced is used to transform the training data to the new feature space and HMMs rebuilt using Baum-Welch or Viterbi training.

### 2.3 Mixture Based Pooling

When mixture densities are used to describe each HMM state, the LDA assumption that each class has the same unimodal multivariate Gaussian distribution is invalid. As the number of mixture densities used in each state is increased, the effects of invalidating the assumption will increase. If each mixture density within a state is used as a separate class then the LDA assumption still holds. The training data can then be mapped to the closest mixture density using Viterbi alignment to create the data pools.

When the LDA transform is used to rebuild the HMMs, the frames of speech data are allowed to align to different mixtures and states from the original mapping to pools. When a single mixture density is used the frames of data can only move between states, but as the number of mixtures is increased the frames of data have more freedom in alignment. In particular, when the Baum-Welch training algorithm is used the frames of data can be shared between mixtures and states. The pooling used to create the LDA transform is unlikely to remain the same during training and the variances will no longer be close to unity. The problem will increase as more mixtures are added.

If the original data pools were used to reconstruct the models then the LDA transform would not be invalidated and the unit variance assumption holds. This is simply

achieved by transforming each pool of data to the new feature space. The means $\mu_{jk}$ and variances $\sigma^2_{jk}$ of each parameter $j$ within a pool $k$ are found using

$$\mu_{jk} = \sum_{i=1}^{I_k} y_{ijk} / I_k$$

and

$$\sigma^2_{jk} = \sum_{i=1}^{I_k} (y_{ijk} - \mu_{jk})^2 / I_k$$

where $y_{ijk}$ is the observation of frame $i$, parameter $j$ in pool $k$,
$I_k$ is the total number of frames in pool $k$.

The mixture coefficients $C_k$ of a new model state are given by

$$C_k = \frac{I_k}{\sum_{k=1}^{K} I_k}$$

where $K$ is the total number of pools (mixtures) in the state.

## 3. EXPERIMENTS

### 3.1 Databases and Analysis

Speaker-independent sub-word models were built from the British-English SCRIBE and SCRIBE-SRU databases [10]. The speech was downsampled to 8 kHz and analysed using nineteen mel-spaced filters. The log-power outputs of the filterbank were transformed using a discrete cosine transform to give the mel cepstrum of the speech at a frame rate of 10 ms. The feature vector consisted of twenty six elements; energy, twelve cepstra and the time derivatives of energy and the cepstra calculated over a 50 ms window. Forty-four sub-word HMMs were built, each HMM having 3 states with left to right topology. The sub-word recognition performance was tested on a portion of the BBC Radio 4 database, a fifty hour database collected in-house for use in word and topic spotting. The material consists of news broadcasts, correspondent's reports, interviews and discussions. The average speaking rate of the data is very high at 160 words/minute.

### 3.2 State and Mixture Pooling

A series of experiments were carried out using the state based and mixture based pooling techniques described in Section 2.3. Table 1 shows the sub-word recognition results achieved for the two techniques as the number of mixture densities per state was increased from one to five. All 26 parameters were used for recognition in each experiment. The performance of the models built using state based pooling degraded for large number of mixtures. The variances of the parameters in the five mixture models were not close to unity, indicating that the LDA assumption that the classes are single Gaussians was indeed being violated. In contrast, the variances of the five mixture models built using mixture based pooling were close to unity, varying between 0.5 and 1.5. The recognition result of 37.9% correct for five mixture models is significantly better than five mixture non LDA models, which gave 35.7% correct on the same test.

| Mixtures | State Based | Mixture Based |
|---|---|---|
| 1 | 28.8 | 32.4 |
| 5 | 30.5 | 37.9 |

**Table 1** Subword Recognition Performance as a Function of the Number of Mixture Modes.

### 3.3 Parameter Reduction and Unit Variances

A further set of experiments were carried out using the five mixture models constructed using mixture-based pooling. Figure 2 shows the sub-word recognition performance as the number of LDA parameters used was reduced. Results are shown for two series of recognition experiments, one when the variances in the HMMs were used and the other when unit variances were assumed and a simple Euclidean distance metric used. The recognition results using the variances outperform the unit variance models by about 2% for all numbers of LDA parameters. This reduction in performance may be tolerated for systems where storage is an important factor, since the variances no longer need to be stored. The performance of the system is maintained as the least discriminative LDA parameters are discarded, allowing 25% of the parameters to be lost without degradation.

### 3.4 Parameter Weighting

The weighting given to each of the input parameters gives an insight into the function of the transform. In Figure 3 the root mean square of the transform coefficients taking the first thirteen, twenty and all the dimensions is shown. This is a measure of the contribution of each input parameter to the LDA coefficients. When all the dimensions are included the weighting approximates to a root power sum weighting [11] since the transform is equalising the variances of all the input parameters to achieve unit variances within the classes. Retaining twenty parameters, which gives the best overall performance, the weighting given to the higher order cepstra and delta cepstra is reduced. This effect is more noticeable when only thirteen parameters are retained when the weighting resembles the raised sinusoidal weighting described in [12]. This indicates that the higher order cepstra and their derivatives are noisy and contribute less to discrimination between the states. The energy and its derivatives have considerable importance in distinguishing between the states.

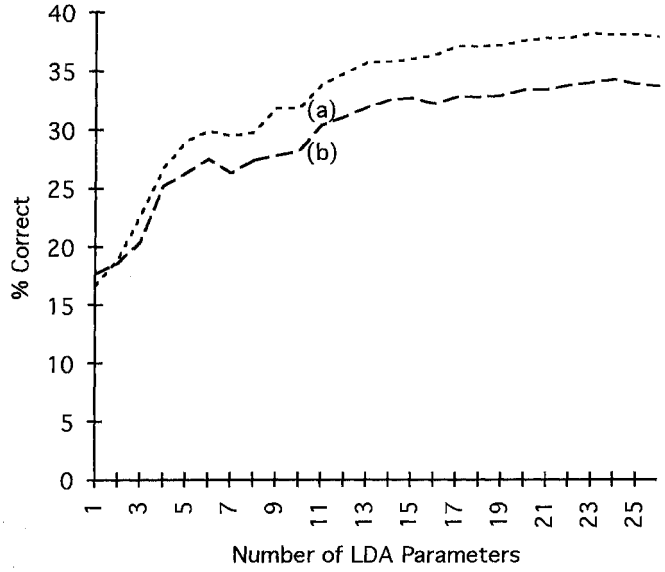


**Figure 2** Subword Recognition Performance as a Function of the Number of LDA Parameters. a) State Variances b) Unit Variances

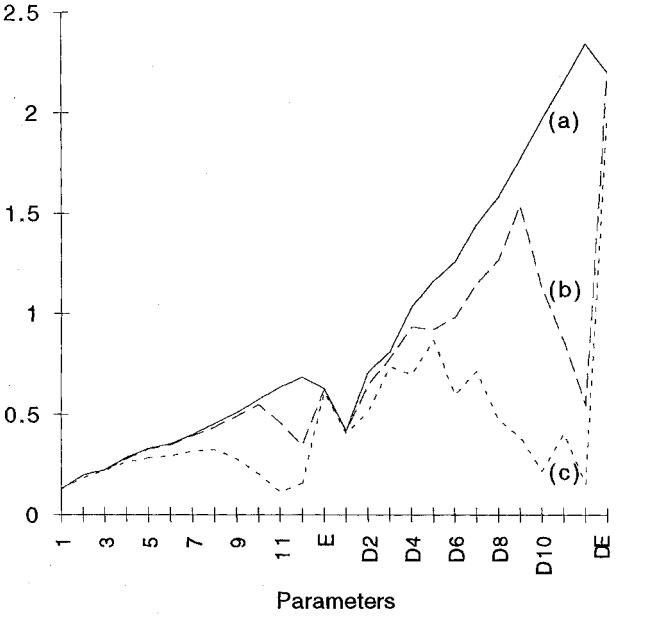


**Figure 3** Relative Contribution of the Input Parameters for

a) all b) 20 and c) 13 LDA coefficients.

## 4. CONCLUSIONS

We have successfully implemented a new technique incorporating continuous density HMMs with mixtures and LDA. A reduction in error rate of 11% has been achieved over the standard LDA technique with 25% of the system parameters being discarded without loss in performance. Constructing the new LDA models directly from the transformed pools reduces the training time considerably compared with techniques such as Baum-Welch or Viterbi.

## 5. ACKNOWLEDGEMENTS

This work was supported by HMG. The authors would like to thank Dr. Eddy Andrews for providing the results on the digit experiments and the Speech Research Unit at DRA, Malvern for providing the SCRIBE-SRU database.

## 6. REFERENCES

[1] R. A. Fisher 'The Use of Multiple Measures in Taxonomic Problems', in Contributions to Mathematical Statistics, Wiley, New York, 1950, pp 32.179-32.188.

[2] M. J. Hunt et al., 'An investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination.', Proc. ICASSP 1991, Toronto.

[3] R. Haeb-Umbach and H. Ney, 'Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition.', Proc. ICASSP 1992, San Francisco.

[4] X. Aubert et al., 'Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models.' Proc. ICASSP 1993, Minneapolis.

[5] R. Roth et al., 'Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data.', Proc. ICASSP 1993, Minneapolis.

[6] C. M. Ayer et al., 'A Discriminatively Derived Linear Transform for Improved Speech Recognition.', Proc. Eurospeech 1993, Berlin.

[7] P. C. Woodland et al., 'Large Vocabulary Continuous Speech Recognition Using HTK.', Proc. ICASSP 1994, Adelaide.

[8] J. L. Gauvain et al., 'The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task.', Proc. ICASSP 1994, Adelaide.

[9] S. J. Young, 'The General Use of Tying in Phoneme-Based Hmm Speech Recognisers', Proc. ICASSP 1992, San Francisco.

[10] J. Laver et al., 'ATR /CSTR Sentence Design.', SCRIBE Documentation 1988.

[11] B. A. Hanson and J. Wakita, 'Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise', IEEE Trans ASSP July 1987 Vol. ASSP-35 No. 7.

[12] Y. Tohkura, 'A Weighted Cepstral Distance Measure for Speech Recognition', IEEE Trans ASSP Oct 1987 Vol.ASSP-35 No. 10.