FEATURE CONCATENATION FOR SPEAKER IDENTIFICATION

R. D. Zilca
Research and Development Division
Amdocs Israel
8 Hapnina St. Raanana, ISRAEL
ranz@amdocs.com

Y. Bistritz
Department of Electrical Engineering
Tel Aviv University
Tel Aviv 69978, ISRAEL
bistritz@eng.tau.ac.il

ABSTRACT

The use of feature vectors obtained by concatenation of different features for text independent speaker identification from clean and telephone speech is studied. The composite feature vectors are examined with GMM and VQ models used to classify speakers. Linear discriminant analysis (LDA), a statistical tool designed to select a reduced set of features for best classification, is applied to enhance performance. The use of LDA for reducing the size of composite feature vector was found satisfactory for clean speech but not for telephone speech. On the other hand, using LDA in the not conventional manner – as a nonsingular transformation (i.e. without size reduction) - improved the performance of composite features in both clean and the telephone speaker identification experiments.

1. INTRODUCTION

The performance of an automatic speaker recognizer in distinguishing between different speakers depends on the combined power of the classifier and feature vectors that are used. An important property of feature vectors in this task is its ability to discriminate between speakers.

Different speech features relate to different aspects of the speech signal. For example, features obtained using linear prediction pertain to spectral envelope of the signal, while filter-banks follow the energy distribution in specific spectral bands. Some features are extracted independently from each frame while others also contain inter-frame information. Speaker recognizers admit a tradeoff between using longer feature vectors with simpler models or using shorter feature vectors with models of richer structure. The need for speaker recognizers with short training and testing periods limits the size of the feature vector that the model may possess, and increases the requirement on the ability of the selected feature vector to discriminate between speakers.

One reasonable approach to improve speaker identification, based on little training and testing data, is to extract from the data several feature vectors (assumedly independent) and then extract from the concatenated features a reduced size vector of enhanced separability. Linear discriminant analysis (LDA) presents a statistical tool designed for this kind of approach. It aims to derive from a set of features a set of reduced size that best meets a separability criterion (e.g. [1]). LDA has been introduced to speech processing by Hunt and was since then applied mostly to speech recognition [2],[3]. Openshaw *et al* [4] used LDA also for speaker identification in experiments that combined a few composite vectors to a new reduced size vector used in VQ classifier of speakers from speech with varying levels of additive noise.

In the currently reported study we apply LDA transformation to concatenation of a selection of feature vectors that includes MFC and LPC based cepstra and LSP (e.g. [5],[6]). The concatenated vectors are used in speaker recognition experiments based on simple VQ and Gaussian Mixture Model (GMM) [7] classifiers. The experiments were performed using short training and testing data of clean and telephone speech (taken from TIMIT and NTIMIT). We show that LDA may be used without size reduction to improve speaker identification rates for composite and even for single features. In fact, in our experiments, using of LDA in the conventional manner of reducing dimension of concatenated features worked satisfactory for only clean speech.

2. LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis (LDA) considers measures for class separability, and linear transformations of feature vectors to optimize these measures for classifying feature vectors $\mathbf{x}$ (say of length N) into one of L classes, ω_i, $i = 1,...,L$ [1]. There exists more than a single way to define a criterion to measure class separability and to obtain the linear transformation. We selected the method most commonly used in the speech recognition literature. It is based on defining a *within-class* scatter matrix and a *between-class* scatter matrix as follows. The within-class

Proc. of the 10-th European Signal Processing Conference,
EUSIPCO-2000, 5 - 8 September, 2000, Tampere, Finland.

matrix expresses the scatter of sample vectors around their respective class expected vectors and is defined by

$$S_w = \sum_{i=1}^{L} P(\omega_i) E\{(\mathbf{x} - M_i)(\mathbf{x} - M_i)^t \mid \omega_i\} = \sum_{i=1}^{L} P(\omega_i)\Sigma_i$$

where $P(\omega_i)$, M_i and Σ_i are the a-priori probability, expected vector and covariance matrix of the i-th class, respectively. The *between-class* scatter matrix is defined as the scatter of expected vectors around the global mean

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i - M_0)(M_i - M_0)^t \ ,$$

where M_0 is the global mean vector defined by the global average of all expected vectors,

$$M_0 = E\{\mathbf{x}\} = \sum_{i=1}^{L} P(\omega_i) M_i \ .$$

These two matrices are used to define the *discrimination matrix* $S_w^{-1} S_b$, which may be used to define a discrimination measure

$$J = tr(S_w^{-1} S_b)$$

This measure increases when the between-class scatter is larger or the within-class scatter is smaller. Thus higher value of J corresponds to better discrimination ability of the feature vector, and hence lower error rate in the classification task. It may be regarded as a multidimensional extension of the F-ratio [8]. Consider now the application of a linear transformation to the classified feature vectors $x \to Ax$. A nonsingular linear transformation can not improve the discrimination power of the feature vector because this measure is invariant to such transformation. Linear discriminant analysis becomes effective when used to select a linear transformation that reduces dimension of feature vectors. If A is a $R \times N$, $R < N$ matrix, then different matrices get different separability measure scores. Let Φ_i and λ_i $(i = 1, 2, ..., N)$ denote the eigenvectors and eigenvalues of the discrimination matrix $S_w^{-1} S_b$ and assume that the eigenvalues are distinct and ordered $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Then the best discriminant score that a reduced vector of size R, obtained by linear combination of entries of the original vector, may achieve is

$$J_R = \sum_{i=1}^{R} \lambda_i$$

This value is attained by choosing the rows of A to be the eigenvectors that correspond to the R largest eigenvalues, viz.,

$$A^t = [\Phi_1 \Phi_2 \cdots \Phi_R]$$

In the following we use the term LDA transformation to a so chosen $R \times N$ matrix A.

In the current application each class (i.e. speaker) was given equal a-priori probability. In other words all $P(\omega_i)$ assumed the value 1/L. We examined in our experiments also nonsingular LDA transformation. The reason for considering also this case, is that even though a nonsingular A does not improve separability, it may still improve classification performance by rotating the decision hyperplanes of the classifier operating to positions perpendicular to the feature axes. This possibility is supported by the fact that the LDA transformation performs *simultaneous diagonalization* of both S_w and S_b [1]. Consequently the transformation affect recognition performance also via “diagonalization on average” of the covariance of each speaker and the covariance of the speakers means.

3. EXPERIMENTAL EVALUATION

The speech database was extracted from the TIMIT (clean speech) and NTIMIT (telephone speech) databases. We selected a subset of the TIMIT/NTIMIT corpora, that includes 32 male speakers, 4 from each dialect. The 2 “sa” sentences, 3 “si” and 2 of the “sx” sentences were used for training. This experimental setup results in a training session with a length of approximately 12 to 20 seconds, depending upon the speaker’s typical rate of speech. The remaining 3 “sx” sentences were used for testing. This experimental setup includes 96 test utterances (3 for each speaker); each of them 1 to 3 seconds long, resulting in a binomial significance interval of 4.75%. The same preprocessing procedure was used for training and for testing. The speech samples are downsampled to 8KHz sampling rate, segmented into 25ms frames with 50% overlap, and multiplied by a Hamming window. Only voiced frames are selected for further use. Three types of features were used: Line Spectra Pairs (LSP), Linear Predictive Cepstra (LPCep) and Mel Frequency Cepstra (MFC). For LSP and LPCep computation, 10th order LPC analysis was performed. The resulting LPC filter was then subject to a 15Hz bandwidth expansion. For computation of the MFC, 18th order Mel Cepstra Coefficients were derived, using the triangle filter banks suggested in [9]. LPCep and LSP that originate from a common linear prediction analysis were not combined but each of them was paired separately with MFC to form a new set of features. The resulting 28th order feature vector was used both in its original form and after applying LDA. To compare performance of LDA with composite features to its performance when applied to a single feature we also bring the results of LDA transformation applied to the MFC feature vector alone.

The VQ classifier was trained using the K-means algorithm [10] for codebook construction with random initial values. On each training iteration, “empty” clusters for which no training vectors were allocated were assigned half the training vectors originally assigned to the largest cluster. This procedure prevents the creation of unchanged codebook vectors. The VQ codebook was also used to initialize the EM iterations for training the GMM model. The initialization involved setting the means of the Gaussian components equal to the codebook vectors, and setting the initial variances and weights to be the sample variances and relative number of training vectors in each cluster, respectively. Each Gaussian component was assumed to include a diagonal covariance matrix, and different variance vectors were assumed for different components. The experiments were performed with VQ codebook and GMM componenets of sizes ranging from 2 to 64.

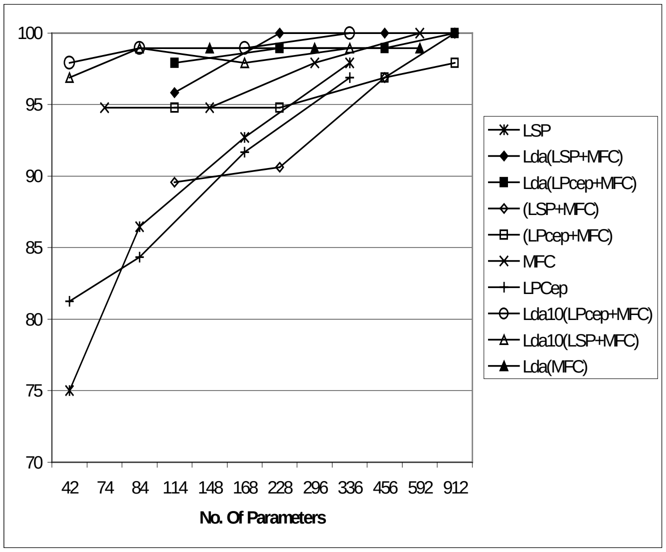

Figure 1. Results for Clean Speech, GMM

The identification rates that were achieved for clean speech and telephone speech, using GMM and VQ classifiers are presented in Figures 1–4. The abscissae in these figures are the number of parameters that participate in the classifier. Fig. 1 and Fig. 2 depict the performance of GMM and VQ speaker modeling in clean speech for the indicated set of feature vectors that included raw, concatenated and LDA transformed features. Fig. 3 and Fig. 4 bring corresponding results using telephone speech.

The labels in the figures use Lda(·) to denote LDA transformation without size reduction, and Lda10(·) to denote LDA transform with reduction to a vector of length 10. The first 10 eigenvalues accounted for 85-90% of the discrimination measure of the full length vectors.

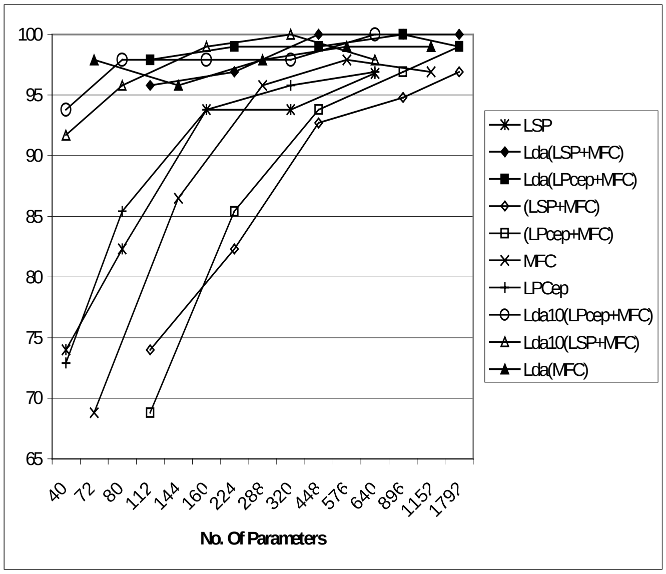

Figure 2. Results for Clean Speech, VQ

Examining the results for clean speech reveals that LDA substantially boosts performance in the low parameter range compared to using the same feature vectors without this discrimination transformation. The differences in performance in the high range of parameters are too small to be counted significant. The use of concatenated features seems to contribute for speaker models only in the low range of parameters. In the higher range of parameters, LDA transformed MFC performed better alone than in combination with additional features. Another important observation concerns dimensionality reduction. It is seen that for both GMM and VQ, LDA admitted feature vector reduction (from 28 to 10) without performance degradation.

The results for telephone speech (Figures 3 and 4) are more widely spread and thus allow more significant conclusions. Again, feature concatenation by itself (i.e. without LDA) did not improve performance. In fact it even may degrade performance, an observation that was in particular evident for VQ. Non-singular LDA transformation improved identification rates significantly for both concatenated features as well as for MFC alone and provided the best attained scores for both the VQ and the GMM classifiers. However, in remarkable departure from the results for clean speech, the use of LDA to reduce dimension of feature vectors provided very poor performance for telephone speech. This is a surprising outcome because the reduced vectors achieved the same portion of the full separability measure as in the case of clean speech. The usefulness of the nonsingular LDA transformation on single and concatenated features may be explained by the tendency of this transformation to also diagonalize the covariance of the transformed features.

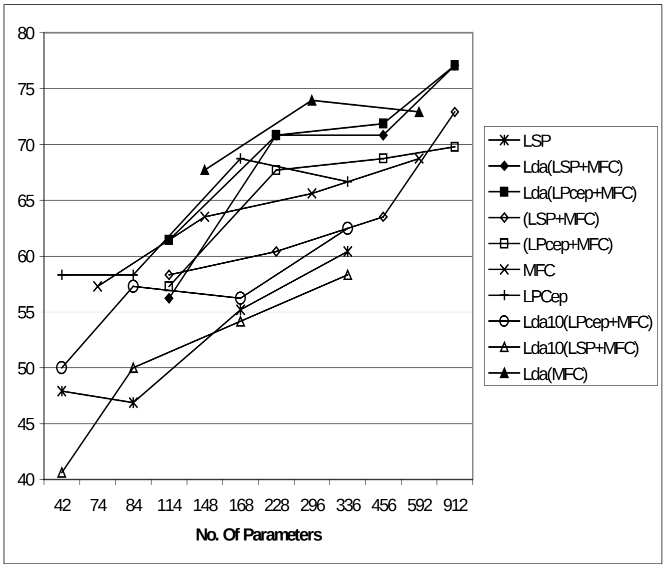

Figure 3. Results for Telephone Speech, GMM

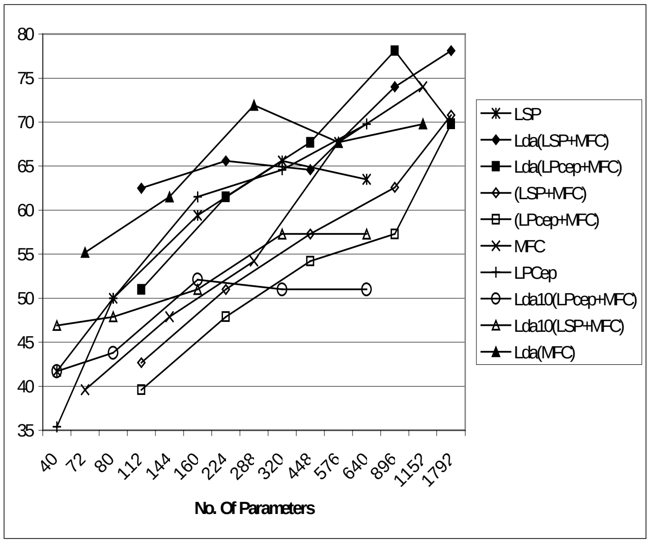

Figure 4. Results for Telephone Speech, VQ

4. CONCLUSIONS

We have carried out speaker identification experiments with VQ and GMM based classifiers to examine the performance of compositions of a few pairs of feature vectors, combined by the use of the LDA transformation. The experiments were held on clean and telephone speech. They demonstrate LDA as a viable technique for improving identification rates for concatenated features. For clean speech LDA was also successful in obtaining reduced size feature vectors from concatenation of features to improve performance of models with relatively low number of parameters. For telephone speech LDA kept boosting performance of transformed feature vectors when used without size reduction but performed poorly as a tool for reducing the size of composite vectors.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

[2] M. J. Hunt, "A statistical approach to metrics for word ans syllable recognition", J. Acoust. Soc. America, Vol. 26, pp. 535-536, 1979.

[3] M. J. Hunt, "Spectral signal processing for ASR" *Acoustic Speech Recognition and Understanding Worshop*, Keystone, Colorado, 12-15 December 1999.

[4] J. P. Openshaw, Z. P. Sun and J.S. Mason, "A comparison of composite features under degraded speech in speaker Recognition," *Proc. of the IEEE Int. Conf. Acoust. Speech Sig. Proc*. 1993.

[5] L. Rabiner and B. H. Huang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[6] D. O'shaughnessy, *Speech Communications Human and Machine*, 2nd edition. IEEE PRESS, 2000.

[7] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, pp. 91-108, 1995.

[8] M. Sambur, "Selection of feature vectors for speaker identification" *IEEE Trans. Acoust., Speech Signal Processing*, ASSP-23, pp.176-182, 1975.

[9] S. B. Davis and P. Marmelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, Aug. 1980.

[10] A. Gersho, R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.