

# Improving the Efficiency of Automatic Speech Recognition by Feature Transformation And Dimensionality Reduction

Xuechuan Wang, Douglas O'Shaughnessy

INRS-Telecommunications

University of Quebec  
800 de la Gauchetiere West  
Montreal, Quebec, H5A 1K6, Canada  
{wwang, dougo}@inrs-telecom.quebec.ca

## Abstract

In speech recognition systems, feature extraction can be achieved in two steps: parameter extraction and feature transformation. Feature transformation is an important step. It can concentrate the energy distributions of a speech signal onto fewer dimensions than those of parameter extraction and thus reduce the dimensionality of the system. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are the two popular feature transformation methods. This paper investigates their performances in dimensionality reduction tasks in continuous speech recognition systems. A new type of feature transformation, LP transformation, is proposed and its performance is compared to those of LDA and PCA transformations.

## 1. Introduction

Feature extraction is a key component in continuous speech recognition (CSR) systems. A major objective of feature extraction is to compress the speech signal so as to overcome what Waibel and Lee called "dimensions of difficulty" [1, 2]. Feature extraction is achieved in two steps: parameter extraction and/or feature transformation. In the parameter extraction step, information relevant for speech recognition is extracted from speech data in the form of  $p$ -dimension parameter vectors. In the feature transformation step, parameter vectors are transformed to feature vectors, which have a dimensionality  $m$  ( $m \leq p$ ). If the parameter extractor is properly designed so that the parameter vectors are precise representations of the information contained in the speech and their dimensionality is low, then there is no necessity for feature transformation. In practice, however, parameter vectors are not satisfactory. For example, redundancy [3] and correlation exist between parameter vectors. Furthermore, the dimensionality of parameter vectors is normally very high ( $\geq 35$ ) and needs to be reduced for the sake of less computational cost and system complexity. Due to these reasons, feature transformation is necessary for robust speech recognition.

Feature transformation is normally implemented by projecting the parameter vectors onto a feature space through a linear transformation matrix. Feature transformation methods vary with the criteria of optimizing the transformation matrix. The two fundamental feature transformation optimization criteria are PCA and LDA. PCA optimizes the transformation matrix by finding the largest variations in the original feature space

[6]. LDA pursues the largest ratio of *between*-class variation and *within*-class variation when projecting the original feature to a subspace [7]. Both of them have been explored in CSR by a number of researchers [8, 9, 10] in recent years. In this work, we investigate the use of LDA and PCA for feature dimensionality reduction. Then we propose a new type of transformation, LP transformation, aiming at improving the robustness and efficiency of CSR systems.

The rest of this paper is organized as follows: Section 2 introduces the framework of LDA and PCA. Section 3 describes the use of LDA and PCA transformations for feature dimensionality reduction and the proposed LP transformation. In Section 4 we give a description of the database and the features used in our experiments. The results of the experiments are given in Section 5 and finally in Section 6 we compare and conclude with the results obtained.

## 2. Feature Transformations

### 2.1. Linear Discriminant Analysis

The goal of linear discriminant analysis is to separate the classes by projecting classes' samples from  $p$ -dimensional space onto a finely oriented line. For a  $K$ -class problem,  $m = \min(K - 1, p)$  different lines will be involved. Thus the projection is from a  $p$ -dimensional space to a  $c$ -dimensional space [7].

Suppose we have  $K$  classes,  $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_K$ . Let the  $i$ th observation vector from the  $\mathcal{X}_j$  be  $x_{ji}$ , where  $j = 1, \dots, J$  and  $i = 1, \dots, N_j$ .  $J$  is the number of classes and  $N_j$  is the number of observations from class  $j$ . The *within*-class covariance matrix  $S_W$  and *between*-class covariance matrix  $S_B$  are defined as:

$$\begin{aligned} S_W &= \sum_{j=1}^K S_j = \sum_{j=1}^K \frac{1}{N_j} \sum_{i=1}^{N_j} (x_{ji} - \mu_j)(x_{ji} - \mu_j)^T \\ S_B &= \sum_{j=1}^K N_j (\mu_j - \mu)(\mu_j - \mu)^T \end{aligned} \quad (1)$$

where  $\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{ji}$  is the mean of class  $j$ ,  $\mu = \frac{1}{N} \sum_{i=1}^N x_i$  is the global mean and  $N = \sum_{j=1}^K N_j$ .

The projection from observation space to feature space is accomplished by a linear transformation matrix  $T_L$ :

$$y = T_L^T x \quad (2)$$

The corresponding *within*-class and *between*-class covariance

matrices in the feature space are:

$$\begin{aligned}\tilde{S}_W &= \sum_{j=1}^K \sum_{i=1}^{N_j} (y_{ji} - \tilde{\mu}_j)(y_{ji} - \tilde{\mu}_j)^T \\ \tilde{S}_B &= \sum_{j=1}^K N_j (\tilde{\mu}_j - \tilde{\mu})(\tilde{\mu}_j - \tilde{\mu})^T\end{aligned}\quad (3)$$

where  $\tilde{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{ji}$  and  $\tilde{\mu} = \frac{1}{N} \sum_{i=1}^N \tilde{y}_i$ . It is straightforward to show that:

$$\begin{aligned}\tilde{S}_W &= T_L^T S_W T_L \\ \tilde{S}_B &= T_L^T S_B T_L\end{aligned}\quad (4)$$

A *linear discriminant* is then defined as the linear function  $T_L^T x$  for which the objective function

$$J(T) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|T_L^T S_B T_L|}{|T_L^T S_W T_L|}\quad (5)$$

is maximum. It can be shown that the solution of (5) is that the  $i$ th column of an optimal  $T_L$  is the generalized eigenvector corresponding to the  $i$ th largest eigenvalue of matrix  $S_W^{-1} S_B$ .

## 2.2. Principal Component Analysis

PCA is a well-established technique for feature extraction and dimensionality reduction. It is based on the assumption that most information about classes is contained in the directions along which the variations are the largest. The most common derivation of PCA is in terms of a standardised linear projection which maximises the variance in the projected space [6]. For a given  $p$ -dimensional data set  $\mathcal{X}$ , the  $m$  principal axes  $T_{P1}, T_{P2}, \dots, T_{Pm}$ , where  $1 \leq m \leq p$ , are orthonormal axes onto which the retained variance is maximum in the projected space. Generally,  $T_{P1}, T_{P2}, \dots, T_{Pm}$  can be given by the  $m$  leading eigenvectors of the sample covariance matrix  $S_G = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)(x_i - \mu)^T$ , where  $x_i \in \mathcal{X}$ ,  $\mu$  is the sample mean and  $N$  is the number of samples, so that:

$$S_G T_{Pi} = \lambda_i T_{Pi} \quad i \in 1, \dots, m \quad (6)$$

where  $\lambda_i$  is the  $i$ th largest eigenvalue of  $S_G$ . The  $m$  principal components of a given observation vector  $x \in \mathcal{X}$  are given by:

$$y = [y_1, \dots, y_m] = [T_{P1}^T x, \dots, T_{Pm}^T x] = T_P^T x \quad (7)$$

The  $m$  principal components of  $x$  are decorrelated in the projected space. In multi-class problems, the variations of data are determined on a global basis, that is, the principal axes are derived from a global covariance matrix:

$$\hat{S}_G = \frac{1}{N} \sum_{j=1}^K \sum_{i=1}^{N_j} (x_{ji} - \hat{\mu})(x_{ji} - \hat{\mu})^T \quad (8)$$

where  $\hat{\mu}$  is the global mean of all the samples,  $K$  is the number of classes,  $N_j$  is the number of samples in class  $j$ ,  $N = \sum_{j=1}^K N_j$  and  $x_{ji}$  represents the  $i$ th observation from class  $j$ . The principal axes  $T_{P1}, T_{P2}, \dots, T_{Pm}$  are therefore the  $m$  leading eigenvectors of  $\hat{S}$ :

$$\hat{S}_G T_{Pi} = \hat{\lambda}_i T_{Pi} \quad i \in 1, \dots, m \quad (9)$$

where  $\hat{\lambda}_i$  is the  $i$ th largest eigenvalue of  $\hat{S}$ .

## 3. Feature Dimensionality Reduction and LP Transformation

### 3.1. Feature Dimensionality Reduction

Computational efficiency is an important problem for real-time CSR systems. The amount of computations required for pattern recognition and the amount of data required for training systems grows exponentially with the increase of the dimensionality of the feature vectors. This is what Bellman called “the curse of dimensionality” [11].

Reducing the dimensionality of parameter vectors is the most direct way to solve the problems caused by high dimensionalities. However, directly reducing the number of parameters will cause unpredictable information loss and consequently make the system performance unstable. This problem can be overcome by using LDA and PCA transformations for feature dimensionality reduction.

Both LDA and PCA transformations,  $T_L$  and  $T_P$ , are composed by the eigenvectors of  $S_W^{-1} S_B$  or the global covariance  $S_G$ . It is usually the case some eigenvalues of matrices  $S_W^{-1} S_B$  or  $S_G$  are zero [12] or very close to zero, as shown in Figure 1. This fact indicates that the energy of a speech signal is mainly distributed along a subset of coordinates corresponding to non-zero eigenvalues. It, therefore, will not cause heavy or unpredictable information loss to discard the coordinates corresponding to zero or zero-proximate eigenvalues. Based on this idea, the dimensionality of parameter vectors is reduced by a reduced-rank  $T_{p \times m}$ :

$$y = T^T x = \{T_1, T_2, \dots, T_M\} x \quad (10)$$

where  $m < p$  and  $T_1, T_2, \dots, T_M$  are first  $m$  leading eigenvectors of  $T_L$  or  $T_P$ .

### 3.2. LP Transformation

LDA and PCA transform parameter vectors with different initiations. LDA transformed feature vectors represent the energy of a speech signal distributed along the eigenvector-spanned coordinates on which the classes have the largest discriminants, while PCA transformed features represent the energy distributions along the directions with largest variates. Given the facts that most information concentrates on the directions spanned by the leading eigenvectors of the two transformations, the combinations of the two sorts of leading directions will give a more detailed description of speech signal energy distributions. This leads to our definition of an LP transformation  $T_{LP}^1$ :

$$T_{LP} = \{T_1, T_2, \dots, T_m, T_{m+1}, T_{m+2}, \dots, T_{m+n}\} \quad (11)$$

where  $T_1, T_2, \dots, T_m$  are the first  $m$  leading eigenvectors of LDA transformation and  $T_{m+1}, T_{m+2}, \dots, T_{m+n}$  are the first  $n$  leading eigenvectors of PCA transformation.

## 4. Experiments

### 4.1. Database

The TIMIT database [13] is used in the experiments of this work. The TIMIT corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United states.

<sup>1</sup>LP = LDA + PCA, for short.

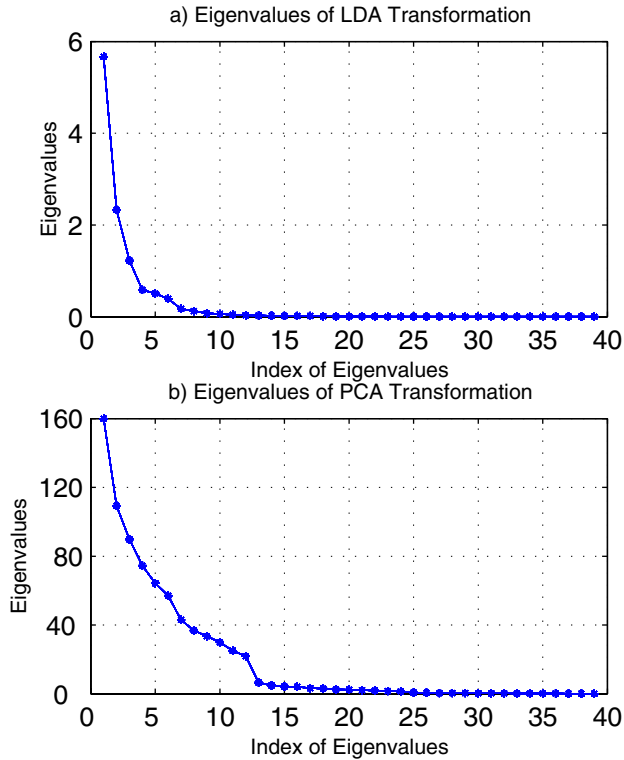


Figure 1: Eigenvalues obtained when computing LDA and PCA transformations.

#### 4.2. CSR System

The HTK-based speech recognition system [14] is used throughout all our experiments. HTK is a Hidden Markov Model (HMM)-based speech recognition system and designed for both isolated and continuous speech recognition. A continuous whole-word-based speech recognition system is built in our experiments. The system uses 5-state HMMs for each of 46 selected monophones. Each state of HMMs has a single Gaussian pdf. model (GMM).

#### 4.3. Experiment Configuration

Our experiments include baseline, LDA transformation, PCA transformation and LP transformation experiments. In baseline experiments, MFCC features are used. The dimensionality of MFCC vectors varies from 3 to 39 in each sub-experiment. In LDA, PCA and LP experiments, 39-dimensional MFCC vectors are obtained first. Then they are transformed to feature vectors in each sub-experiment with the dimensionality corresponding to that of baseline experiments. Features and the dimensionality of feature vectors used in these experiments are shown in Table 1.

### 5. Results

The results of the four experiments are shown in Figure 2. Observations from these results can be summarized as follows:

- The word recognition rate curves of PCA, LDA and LP transformation experiments are flatter than that of the baseline experiment. This indicates that the reduction of

Dimension	Baseline	LDA	PCA	LP
3	3-MFCC	3-LDA	3-PCA	2-PCA 1-LDA
9	3-MFCC 3-Delta 3-Acce	9-LDA	9-PCA	5-PCA 4-LDA
15	5-MFCC 5-Delta 5-Acce	15-LDA	15-PCA	8-PCA 7-LDA
21	7-MFCC 7-Delta 7-Acce	21-LDA	21-PCA	11-PCA 10-LDA
27	9-MFCC 9-Delta 9-Acce	27-LDA	27-PCA	14-PCA 13-LDA
33	11-MFCC 11-Delta 11-Acce	33-LDA	33-PCA	17-PCA 16-LDA
39	13-MFCC 13-Delta 13-Acce	39-LDA	39-PCA	20-PCA 19-LDA

Table 1: Features and dimensionality of feature vectors used in the experiments.

feature dimensionality by a PCA, LDA or LP transformation does not have significant negative influence on the performance of CSR systems.

- The recognition rates are improved after PCA, LDA and LP transformations.
- The word recognition rate after a PCA, LDA or LP transformation degrades steeply only when the feature dimensionality is reduced below 9. This indicates that heavy information loss after PCA, LDA or LP transformations starts only when feature dimensionality is reduced to a very small number.
- LP transformation has a better performance than those of PCA and LDA transformations.

### 6. Conclusions

In this paper, we investigate the performance of feature transformation methods for dimensionality reduction in continuous speech recognition tasks. A new type of transformation, LP, which is based on PCA and LDA is proposed. Experimental results show the following patterns of feature transformation methods:

- The word recognition rate is improved after the feature transformation step.
- The performance of CSR systems does not have significant degradation before feature dimensions are reduced to 15 after feature transformation by PCA, LDA or LP.
- The performance of PCA transformation is not stable in high dimensional feature spaces.
- The overall performance of LP transformation is better than that of PCA and LDA. However, the PCA transformation has better performances on certain dimensions (21 and 3) than LP transformation.

This implies that it is applicable to reduce speech model storage and computation expenses by employing feature transformation

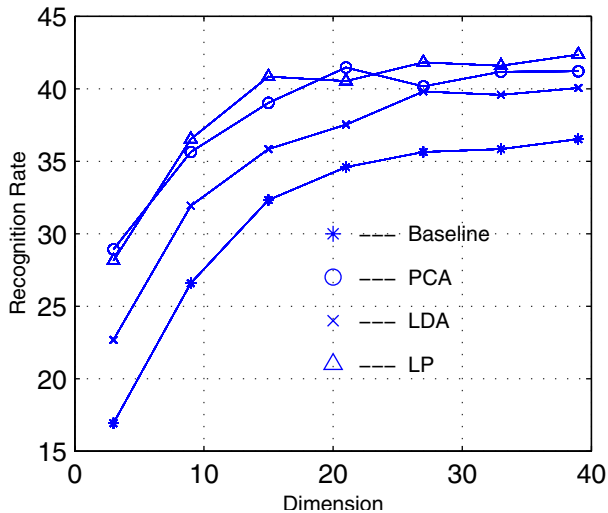


Figure 2: Comparison of the word recognition rate of the baseline, PCA, LDA and LP transformation experiments.

methods. The results also show two directions for future work. One is the improvement of the CSR system performance on very low dimensions ( $< 10$ ). This would be extremely useful in some commercial CSR systems. The other is to find a better optimized criterion for feature transformation, since the LP transformation, a simple concatenation of PCA and LDA transformation, shows improved performance.

## 7. References

- [1] Waibel, A. and Lee, K.-F., eds. Readings in Speech Recognition, Palo Alto, Calif.: Morgan-Kaufmann, 1990.
- [2] Deller, J.R., Jr., Hansen, J.H.L. and Proakis, J.G., Discrete-Time Processing of Speech Signals, Macmillan Publishing Co., New York, 1993.
- [3] Barlow, H.B., "Possible principles underlying the transformation of sensory messages", In Rosenbluth, W.A., Editor, Sensory Communication, Cambridge:MIT Press, pp. 217-234, 1961.
- [4] Sharma, S., Ellis, D., Kajarekar, S., Jain, P. and Hermansky, H., "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database", ICASSP'00, Istanbul, 2000, pp. 1117-1120.
- [5] Shire, M.L. and Chen, B.Y., "On data-derived temporal processing in speech feature extraction", ICSLP'00, Beijing, 2000, pp. 71-74.
- [6] I.T. Jolliffe, Principal component analysis, Springer-Verlag, New York, 1986.
- [7] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons Press, New York, 1973.
- [8] Selouani, S. and O'Shaughnessy, D., "Noise-robust speech recognition in car environments using genetic algorithms and a Mel-cepstral subspace approach", ICSLP'02, Denver, USA, 2002, pp. 2173-2176.
- [9] Somervuo, P., "Experiments with linear and nonlinear feature transformations in HMM based phone recognition", ICASSP'03, Hongkong, 2003.

- [10] Tolba, H., Selouani, S. and O'Shaughnessy, D., "Comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for automatic speech recognition using a multi-stream paradigm", ICSLP'02, Denver, USA, 2002, pp. 2113-2116.
- [11] R.E. Bellman, *Dynamic Programming*, Princeton University Press, 1957.
- [12] Ephraim, Y. and Van Trees, H.L., "A signal subspace approach for speech enhancement", IEEE Trans. on Speech and Audio Processing, 3(4):251-266, 1995.
- [13] Fisher, W., Doddington, G. and Goudie-Mashall, K., The DARPA speech recognition research database: specifications and status, Proceedings DARPA of Speech Recognition Workshop, 1986, pp. 93-99.
- [14] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., The HTK Book, Cambridge University Speech Group, July 2000.