



# Clustering-based Two-Dimensional Linear Discriminant Analysis for Speech Recognition

Xiao-Bing Li, Douglas O'Shaughnessy

INRS-Energy, Materials and Telecommunications  
 800 de la Gauchetiere Ouest, Montreal, H5A 1K6, Canada  
 {xiaobing, dougo}@emt.inrs.ca

## Abstract

In this paper, a new, Clustering-based Two-Dimensional Linear Discriminant Analysis (Clustering-based 2DLDA) method is proposed for extracting discriminant features in Automatic Speech Recognition (ASR). Based on Two-Dimensional Linear Discriminant Analysis (2DLDA), which works with data represented in matrix space and is adopted to extract discriminant information in a joint spectral-temporal domain, Clustering-based 2DLDA integrates the cluster information in each class by redefining the between-class scatter matrix to tackle the fact that many clusters exist in each state in Hidden Markov Model (HMM)-based ASR. The method was evaluated in the TiDigits connected-digit string recognition and the TIMIT continuous phoneme recognition. Experimental results show that 2DLDA yields a slight improvement on the recognition performance over classical LDA, and our proposed Clustering-based 2DLDA outperforms 2DLDA.

**Index Terms:** Speech recognition, LDA, 2DLDA, cluster information, Clustering-based 2DLDA,  $K$ -means.

## 1. Introduction

Linear Discriminant Analysis [1], which finds a linear transform by maximizing the Fisher ratio, is a popular technique for feature transformation and dimensionality reduction with good discriminative properties in the pattern classification field. It and its extensions (e.g., Heteroscedastic Linear Discriminant Analysis (HLDA) [2], weighted LDA [3], etc.) also have been successfully applied to extract discriminative features in current state-of-the-art, HMM-based ASR [4][5][6][7].

The discriminant information can be extracted from the spectral [8], temporal [9], or joint spectral-temporal domain [4][10][11]. For joint spectral-temporal LDA, the necessary scatter matrices will be large as multiple temporal context speech frames are used to form a high-dimensional feature. Then the computation of eigen-decomposition of a large matrix becomes expensive and the small sample size problem may be encountered because of the high dimensionality. Thus, Two-Dimensional Linear Discriminant Analysis, which was first proposed to overcome the singularity problem in

face recognition [12], was adopted to speech recognition [10]. Different from classical LDA, which extracts discriminant information in vector space, the extraction of 2DLDA is implemented directly by matrix represented data to reduce the computational cost and to overcome the small sample size problem.

In general, each state is identified as a class to get the LDA transformation for HMM-based ASR. However, due to variations of speaker, environment, context, emotion, etc., it is often the case that multiple clusters exist in one state, as, in continuous density HMM-based ASR, a state is commonly modelled by a mixture of Gaussian components. So using only the class mean to represent the whole data structure of a state, as does the conventional LDA method, may not be appropriate. To solve this problem, in [13], we proposed a Modified Linear Discriminant Analysis (MLDA) method, in which the cluster information in each state is integrated through redefining the between-class scatter matrix. In this paper, based on 2DLDA, the Clustering-based 2DLDA method is presented as an extension of our proposed MLDA method to exploit the data structure inside each state.

The rest of the paper is organized as follows. In Section 2, the 2DLDA method and its iterative optimization procedure are reviewed. In Section 3, the proposed clustering-based 2DLDA method and its implementing procedure are presented. Databases, experimental setups and results are presented in Section 4. We summarize our work in Section 5.

## 2. Two-Dimensional Linear Discriminant Analysis

Different from classical LDA, in which the data is represented in vector space, matrix representation is adopted by 2DLDA. For a set of  $N$  matrix represented samples  $\mathbf{X}_n \in \mathbb{R}^{t \times f}, 1 \leq n \leq N^1$ , which belongs to  $C$  different classes, 2DLDA aims at finding two transforms,  $\mathbf{T} \in \mathbb{R}^{t \times t'}$  and  $\mathbf{F} \in \mathbb{R}^{f \times f'}$ , to project  $\mathbf{X}_n$  to  $\mathbf{Y}_n \in \mathbb{R}^{t' \times f'}$  by  $\mathbf{Y} = \mathbf{T}^T \mathbf{X} \mathbf{F}$ . Similar to classical LDA, these two transforms can be obtained by maximizing the Fisher ratio [1] of between-class and within-class scatter matrices after

<sup>1</sup>for speech recognition,  $\mathbf{X}_n$  represents the concatenated acoustic vectors computed on successive speech frames.

$$\mathbf{S}_W = \sum_{i=1}^C \sum_{n=1}^{N_i} \|\mathbf{X}_{in} - \mathbf{M}_i\|_F^2 = \text{trace} \left( \sum_{i=1}^C \sum_{n=1}^{N_i} (\mathbf{X}_{in} - \mathbf{M}_i)(\mathbf{X}_{in} - \mathbf{M}_i)^T \right) \quad (1)$$

$$\mathbf{S}_B = \sum_{i=1}^C N_i \|\mathbf{M}_i - \mathbf{M}\|_F^2 = \text{trace} \left( \sum_{i=1}^C N_i (\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^T \right) \quad (2)$$

$$\tilde{\mathbf{S}}_W = \text{trace} \left( \sum_{i=1}^C \sum_{n=1}^{N_i} \mathbf{T}^T (\mathbf{X}_{in} - \mathbf{M}_i) \mathbf{F} \mathbf{F}^T (\mathbf{X}_{in} - \mathbf{M}_i)^T \mathbf{T} \right) \quad (3)$$

$$\tilde{\mathbf{S}}_B = \text{trace} \left( \sum_{i=1}^C N_i \mathbf{T}^T (\mathbf{M}_i - \mathbf{M}) \mathbf{F} \mathbf{F}^T (\mathbf{M}_i - \mathbf{M})^T \mathbf{T} \right) \quad (4)$$

$$\mathbf{S}_B = \text{trace} \left( \sum_{i=1}^{C-1} \sum_{j=i+1}^C \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} N_{ik} N_{jl} (\mathbf{M}_{ik} - \mathbf{M}_{jl})(\mathbf{M}_{ik} - \mathbf{M}_{jl})^T \right) \quad (9)$$

$$\mathbf{S}_B^{\mathbf{F}} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} N_{ik} N_{jl} (\mathbf{M}_{ik} - \mathbf{M}_{jl}) \mathbf{F} \mathbf{F}^T (\mathbf{M}_{ik} - \mathbf{M}_{jl})^T \quad (10)$$

$$\mathbf{S}_B^{\mathbf{T}} = \sum_{i=1}^{C-1} \sum_{j=i+1}^C \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} N_{ik} N_{jl} (\mathbf{M}_{ik} - \mathbf{M}_{jl})^T \mathbf{T} \mathbf{T}^T (\mathbf{M}_{ik} - \mathbf{M}_{jl}) \quad (11)$$

the projection.

Using the Frobenius norm as the similarity metric between matrices, the within-class scatter matrix and between-class scatter matrix are defined as Eqs. (1) and (2), where  $\mathbf{X}_{in}, 1 \leq n \leq N_i$  are the data samples of the  $i$ -th class,  $\mathbf{M}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathbf{X}_{in}$  is the mean of the  $i$ -th class, and  $\mathbf{M} = \frac{1}{N} \sum_{n=1}^N \mathbf{X}_n$  is the global mean. Thus in the low-dimensional space, the within-class scatter matrix and between-class scatter matrix can be computed as Eqs. (3) and (4).

It is difficult to get the optimal transforms  $\mathbf{T}$  and  $\mathbf{F}$  simultaneously by maximizing  $\mathcal{J}(\mathbf{T}, \mathbf{F}) = \tilde{\mathbf{S}}_B / \tilde{\mathbf{S}}_W$ . In [12] an algorithm was proposed to find each of them by iteratively fixing another one. This iterative procedure is summarized as follows: 1) Fix  $\mathbf{F}$ , and compute:

$$\mathbf{S}_W^{\mathbf{F}} = \sum_{i=1}^C \sum_{n=1}^{N_i} (\mathbf{X}_{in} - \mathbf{M}_i) \mathbf{F} \mathbf{F}^T (\mathbf{X}_{in} - \mathbf{M}_i)^T, \quad (5)$$

$$\mathbf{S}_B^{\mathbf{F}} = \sum_{i=1}^C N_i (\mathbf{M}_i - \mathbf{M}) \mathbf{F} \mathbf{F}^T (\mathbf{M}_i - \mathbf{M})^T; \quad (6)$$

- 2) Eigen-decompose  $(\mathbf{S}_W^{\mathbf{F}})^{-1} \mathbf{S}_B^{\mathbf{F}}$  to obtain the current  $\mathbf{T}$ ;  
3) Fix  $\mathbf{T}$ , and compute:

$$\mathbf{S}_W^{\mathbf{T}} = \sum_{i=1}^C \sum_{n=1}^{N_i} (\mathbf{X}_{in} - \mathbf{M}_i)^T \mathbf{T} \mathbf{T}^T (\mathbf{X}_{in} - \mathbf{M}_i), \quad (7)$$

$$\mathbf{S}_B^{\mathbf{T}} = \sum_{i=1}^C N_i (\mathbf{M}_i - \mathbf{M})^T \mathbf{T} \mathbf{T}^T (\mathbf{M}_i - \mathbf{M}); \quad (8)$$

- 4) Eigen-decompose  $(\mathbf{S}_W^{\mathbf{T}})^{-1} \mathbf{S}_B^{\mathbf{T}}$  to obtain the current  $\mathbf{F}$ ;  
5) repeat the above procedure by using the obtained transforms until a reasonable number of iterations is reached.

### 3. Clustering-based Two-Dimensional Linear Discriminant Analysis

It is often the case that multiple clusters exist in each state of HMMs due to variations of speaker, environment, context, emotion, etc. So it may not be appropriate to use only a mean to represent the whole data structure of a state. In order to tackle this possible underlying weakness, we proposed a new, Modified Linear Discriminant Analysis method in [13]. In this paper, we extend this idea to 2DLDA to adopt the cluster information in each state by rewriting the between-class scatter matrix, and term it as Clustering-based 2DLDA.

Let  $K_i$  denote the number of clusters in class  $i$ ,  $N_{ik}$  denote the number of samples belonging to the  $k$ -th cluster in class  $i$ , and  $\mathbf{M}_{ik}$  denote the mean of the  $k$ -th cluster in class  $i$ ; the between-class scatter matrix  $\mathbf{S}_B$  is defined as Eq. (9).

In speech recognition, the difference between the clusters of a state is usually smaller than the difference between different states, so we should separate different classes/states while putting constraints on the clusters of each class, i.e., each class/state should be kept compactly. Thus the definition of the within-class scatter matrix is maintained as Eq. (1).

The transforms of Clustering-based 2DLDA can be obtained by following the same iterative procedure shown in Section 2, only Eqs. (6) and (8) become Eqs. (10) and (11).

In our study, the clusters in each state are obtained by  $K$ -means clustering on the center frame (the center row of each matrix), and the same number of clusters is used in each state. Then, the transforms can be obtained by the iterative procedure. In our experiments, only one iteration (as suggested by [12]) was used to get the transform matrices. Below is the procedure:

1. Use a well-trained model to align the training data (or a randomly selected subset of the full training set) to the state-level with the correct transcription;
2. For each state, use  $K$ -means clustering to get its clusters with the frames aligned to it;
3. Calculate the necessary statistics to get the optimal 2DLDA transforms.

## 4. Experimental Results

### 4.1. Connected-digit String Recognition

TiDigits, a speaker independent, connected digit utterances database, was used to test our method. The speech signal was recorded from various regions of the United States. It contains utterances from 326 speakers (50 boys, 51 girls, 111 men, and 114 women). The digit string has a random length from 1 to 7. We only used the adult portion of this database in our experiments. There are 8,623 strings for training and 8,700 strings for testing.

Each digit (“ONE” — “NINE”, “ZERO”, and “OH”) was modelled by a 12-state, whole-word based HMM. A 3-state silence model and a 1-state short pause model were added. All of the 136 states were identified as the classes to get the classical LDA/2DLDA/Clustering-based 2DLDA transforms. Nine frames with the 23-dimensional log Mel-scale Filterbank coefficients form the 207-dimensional super-frame. Then it is represented as vector or matrix to obtain the corresponding transforms. For 2DLDA and Clustering-based 2DLDA, 3 temporal discriminants and 13 spectral discriminants are extracted from the original 9 by 23 matrix to form the final 39-dimensional features, and 4 temporal discriminants and 10 spectral discriminants are extracted to form the 40-dimensional features.

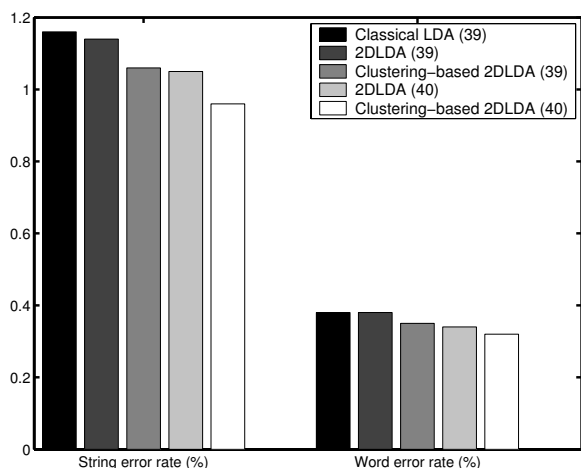


Figure 1: Recognition performance in string error rate and word error rate on the TiDigits database with different methods. The number inside parentheses is the number of dimensions of the feature.

Fig. 1 gives the recognition performance in string er-

ror rate and word error rate for Classical LDA, 2DLDA, and Clustering-based 2DLDA, respectively. We can find that 2DLDA gives similar results as classical LDA, and our proposed Clustering-based 2DLDA gives the best recognition performance. In addition, using 40-dimensional features shows clear improvement of the recognition performance, compared with the systems with 39-dimensional features.

### 4.2. Continuous Phoneme Recognition

We also applied Clustering-based 2DLDA for continuous phoneme recognition on the TIMIT database. The standard training set (3,696 utterances) and coreTest set (192 utterances) were used excluding the “sa” sentences. 39 phones, which were mapped from the original 61 phonetic labels as in [14], were used to create the context-dependent HMMs. Table 1 gives the recognition performance in phone error rate for four different systems with different features:

*Baseline:* the features are the conventional 39-dimensional MFCCs (12 static MFCCs, log energy, and their first- and second-order time derivatives);

*Classical LDA:* the 234-dimensional super-frame, which is formed from 9 concatenated feature vectors (each vector is the 26-dimensional log Mel-scale Filterbank coefficients), is reduced to form the final 39/40-dimensional features by the classical LDA transformation;

*2DLDA:* similar to the above, but the super-frame is represented by a  $9 \times 26$  matrix. The matrix represented feature is then transformed to final dimension  $3 \times 13$  (i.e., 39-dimensional features) and  $4 \times 10$  (i.e., 40-dimensional features) by the 2DLDA method;

*Clustering-based 2DLDA:* the features are similar to the above, and 2DLDA is replaced by the Clustering-based 2DLDA.

For each system, about 930~950 tied states are used. We can find that, compared with the baseline, applying classical LDA brings clear improvement of recognition performance. With the same number of feature dimensions and comparable number of model parameters, 2DLDA gives slightly better results than classical LDA does, while the computation cost is lower. As expected, Clustering-based 2DLDA outperforms 2DLDA as it exploits the data structure inside a class/state. We also find that clear performance improvement is obtained when 40-dimensional features are used.

## 5. Conclusions

A new method, termed Clustering-based 2DLDA, was presented to extract discriminant features from a joint spectral-temporal domain in HMM-based ASR. Based on 2DLDA, it is designed to tackle the multiple clusters

# Dimensions	39		40	
# Mixtures	10	12	10	12
Baseline	30.23%	30.19%	—	—
Classical LDA	28.94%	28.96%	28.47%	28.41%
2DLDA	29.16%	28.75%	28.25%	27.89%
Clustering-based 2DLDA	28.93%	28.60%	28.09%	27.57%

Table 1: Recognition performance in phone error rate for four different systems with different features on the TIMIT coreTest set.

problem in each state by integrating the cluster information in each class into the definition of the between-class scatter matrix.  $K$ -means clustering is used to get the clusters of each state. Clustering-based 2DLDA was successfully applied to two standard ASR tasks: the connected-digit string recognition with the TiDigits database and the continuous phoneme recognition with the TIMIT database. Experimental results show that Clustering-based 2DLDA performs better than 2DLDA and classical LDA in both cases. Further, the recognition error rate is reduced by using it.

## 6. References

- [1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, 2001.
- [2] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, December 1998.
- [3] Y. Li, Y. Gao, and H. Erdogan, "Weighted pairwise scatter to improve linear discriminant analysis," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, September 2000.
- [4] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, San Francisco, CA, March 1992, pp. 13–16.
- [5] E. Schukat-Talamazzini, J. Hornegger, and H. Niemann, "Optimal linear feature transformations for semi-continuous hidden Markov models," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Detroit, MI, May 1995, pp. 369–372.
- [6] A. Ljolje, "Optimization of class weights for LDA feature transformations," in *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, September 2006, pp. 385–388.
- [7] R. Schluter, A. Zolnay, and H. Ney, "Feature combination using linear discriminant analysis and its pitfalls," in *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, September 2006, pp. 345–348.
- [8] H. Hermansky and N. Malayath, "Spectral basis functions from discriminant analysis," in *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 1379–1383.
- [9] C. Avendano, S. van Vuren, and H. Hermansky, "Data-based RASTA-like filter design for channel normalization in ASR," in *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, PA, October 1996, pp. 2087–2090.
- [10] S. Kajarekar, B. Yegnanarayana, and H. Hermansky, "A study of two dimensional linear discriminants for ASR," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, UT, May 2001, pp. 137–140.
- [11] F. Valente and H. Hermansky, "Discriminant linear processing of time-frequency plane," in *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, September 2006, pp. 349–352.
- [12] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proceedings of Advances in Neural Information Processing Systems (NIPS 2004)*, vol. 17, 2004, pp. 1569–1576.
- [13] X.-B. Li and D. O'Shaughnessy, "Modified linear discriminant analysis for speech recognition," in *Proceedings of the 20th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Vancouver, Canada, April 2007.
- [14] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, November 1989.