DIMENSIONALITY REDUCTION USING MCE-OPTIMIZED LDA TRANSFORMATION

Xiao-Bing Li, Jin-Yu Li, Ren-Hua Wang

USTC iFly Speech Lab, University of Science and Technology of China, Hefei, Anhui, China
{lixiaobing, jinyuli}@ustc.edu, rhw@ustc.edu.cn

ABSTRACT

In this paper, Minimum Classification Error (MCE) method is extended to optimize both Linear Discriminant Analysis (LDA) transformation and the classification parameters for dimensionality reduction. Firstly, under the HMM-based Continuous Speech Recognition (CSR) framework, we use MCE criterion to optimize the conventional dimensionality reduction method, which uses LDA to transform standard MFCCs. Then, a new dimensionality reduction method is proposed. In the new method, the combination of Discrete Cosine Transform (DCT) and LDA, as used in the conventional method, is replaced by a single LDA transformation, which is optimized according to MCE criterion along with the classification parameters. Experimental results on TiDigits show that even when the feature dimension is reduced to 14, the performance of this new method is as good as that of the MCE-trained system using 39 dimension MFCCs. It also outperforms our MCE-optimized conventional dimensionality reduction method.

1. INTRODUCTION

In order to implement speech recognition on a resource-limited platform, we always try to reduce the model size as much as possible. One choice is to use a small number of model units, states or Gaussian mixtures. Another choice is to reduce the feature dimension, which is the focus in our work. LDA transformation [1] is usually chosen to perform dimensionality reduction.

Figure 1 shows the Conventional dimensionality reduction feature extractor (Conventional-DRFE), in which LDA transformation is used to transform the standard MFCCs to a new, lower dimension feature vector. LDA attempts to separate classes through maximizing the ratio of between-class scatter matrix and within-class scatter matrix, however, this has little direct relation with the final classifier's target of minimum recognition error rate. In contrast, MCE [2] can adjust the classification parameters to achieve minimum recognition error. And its extension, Discriminative Feature Extraction (DFE), has been employed in various speech recognition tasks, such as filterbank design [3], feature transformation [4], and dynamic feature design [5]. In our work, DFE is extended to carry out dimensionality reduction. We adjust the LDA transformation parameters and the classification parameters simultaneously with the MCE criterion in the DFE framework. A similar idea was reported in [6] to solve the Mahalanobis distance based vowel recognition, while not under HMM framework. Since HMM is the mainstream algorithm in speech recognition. So we develop the MCE-optimized conventional dimensionality reduction method into the HMM-based CSR framework.

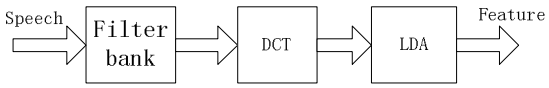

Figure 1. Block diagram of Conventional-DRFE

As we know, both DCT and LDA can be used for feature decorrelation. In Conventional-DRFE, DCT is used for this purpose. However, it has been reported that LDA is a better choice than DCT for feature decorrelation [7]. Moreover, LDA can also be used for dimensionality reduction besides feature decorrelation. So we use a single LDA transformation to replace the combination of DCT and LDA to perform feature decorrelation and dimensionality reduction simultaneously. This dimensionality reduction method (New-DRFE) is shown in figure 2. It is similar to the method reported in [7] that using LDA to replace DCT. But their system was trained by MLE. In contrast, we propose to use MCE criterion to optimize the LDA transformation and the classification parameters in the DFE framework. Three versions of our method are derived.

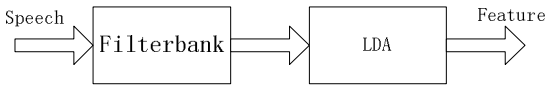

Figure 2. Block diagram of New-DRFE

The rest of the paper is organized as follows. Section 2 describes the DFE in the HMM-based CSR framework, and provides the derivation of the updating formulas for the LDA transformations both in Conventional-DRFE and New-DRFE under the MCE criterion. In section 3, we show our experimental results on TiDigits. The choice of initial transformation in New-DRFE is discussed in section 4. Finally, we summarize our work in section 5.

2. DFE-BASED LDA TRANSFORMATION OPTIMIZATION

2.1. DFE in HMM-Based continuous speech recognition framework

DFE is the extension of MCE for joint optimization of models and features. Let $\Phi = (\Lambda, \Gamma)$ denote the parameter set, where Λ denotes the model parameters, and Γ denotes the parameter set of the feature extraction module.

In CSR, string-model-based discriminant function [2] is used. For an input speech utterance, the final feature is $O = \{\vec{o}_1, \cdots, \vec{o}_T\}$. Let $S_i, i = 1, \cdots, N$ denote the top N best competing strings, the corresponding discriminant function is given by:

$$g(O, S_i, \Phi) = \log f(O, Q_{S_i}, S_i \mid \Phi) \tag{1}$$

And for the correct string S_{lex}, the discriminant function is:

$$g(O, S_{lex}, \Phi) = \log f(O, Q_{S_{lex}}, S_{lex} \mid \Phi) \tag{2}$$

where Q_{S_i} ($Q_{S_{lex}}$) is the optimal state sequence of the word string S_i (S_{lex}). Then the misclassification measure is defined as:

$$\begin{aligned} d(O,\Phi) = &- g(O, S_{lex}, \Phi) \\ &+ \log\left\{\frac{1}{N-1}\sum_{k=1, S_k \neq S_{lex}}^{N} \exp\big(g(O, S_k, \Phi)\eta\big)\right\}^{1/\eta} \end{aligned} \tag{3}$$

It is embedded into the sigmoid function: $l(O, \Phi) = (1 + e^{-\gamma d(O,\Phi)})^{-1}$.

The goal of DFE is to minimize the expected loss $L(\Phi) = E_X\big[l(O, \Phi)\big]$. This can be solved by the Generalized Probabilistic Descent (GPD) algorithm as:

$$\Lambda_{n+1} = \Lambda_n - \varepsilon_n U_1 \frac{\partial l(O_n, \Phi)}{\partial \Lambda}\bigg|_{\Lambda = \Lambda_n} \tag{4}$$

$$\Gamma_{n+1} = \Gamma_n - \tau_n U_2 \frac{\partial l(O_n, \Phi)}{\partial \Gamma}\bigg|_{\Gamma = \Gamma_n} \tag{5}$$

where U_1 and U_2 are positive definite matrices, ε_n and τ_n are the learning step size for Λ and Γ.

The chain rule of differential calculus is used to adjust Λ and Γ. When $\tau_n = 0$, this training is the classical MCE, and when $\varepsilon_n = 0$, it is only to optimize the feature extractor's parameters. The complete updating formula for Λ can be found in [2]. The updating formula for Γ is described in detail as follows.

2.2. Gradient calculation of LDA transformation

LDA transformation W transforms the original n dimension feature vector $\vec{x}$ into a new d ($d \le n$) dimension vector $\vec{y}$. It is formulated as: $\vec{y}_t = W\vec{x}_t$.

To use DFE to adjust LDA transformation, we have $\Gamma = W$. The gradient calculation is given as follows:

$$\frac{\partial l(O, \Phi)}{\partial W} = \frac{\partial l(O, \Phi)}{\partial d(O, \Phi)} \frac{\partial d(O, \Phi)}{\partial W} \tag{6}$$

$$\frac{\partial l(O, \Phi)}{\partial d(O, \Phi)} = \gamma l(O, \Phi)\big[1 - l(O, \Phi)\big] \tag{7}$$

$$\begin{aligned} \frac{\partial d(O, \Phi)}{\partial W} = &- \frac{\partial g(O, S_{lex}, \Phi)}{\partial W} + \\ &\sum_{i=1, S_i \neq S_{lex}}^{N} \left[\frac{\exp\big[g(O, S_i, \Phi)\eta\big]}{\sum_{j=1, S_j \neq S_{lex}}^{N} \exp\big[g(O, S_j, \Phi)\eta\big]} \frac{\partial g(O, S_i, \Phi)}{\partial W}\right] \end{aligned} \tag{8}$$

For LDA transformation in Conventional-DRFE, let $\vec{x}_t$ denote the input feature vector as MFCCs, the output feature vector is $\vec{o}_t = W\vec{x}_t$. We have:

$$\begin{aligned} \frac{\partial g(O, S, \Phi)}{\partial W} &= \sum_{t=1}^{T} \delta(q_t - j) b_j^{-1}(\vec{o}_t) \frac{\partial b_j(\vec{o}_t)}{\partial W} \\ &= -\sum_{t=1}^{T} \delta(q_t - j) \sum_{m=1}^{M} \gamma_{jm}(\vec{o}_t)\big[C_{jm}^{-1}(\vec{o}_t - \mu_{jm})\vec{x}_t^T\big] \end{aligned} \tag{9}$$

where, S is S_i or S_{lex} , $\delta(\cdot)$ denotes the Kronecker delta function,

$$\begin{aligned} b_j(\vec{o}_t) &= \sum_{m=1}^{M} c_{jm} b_{jm}(\vec{o}_t) \\ &= \sum_{m=1}^{M} \frac{c_{jm}}{(2\pi)^{d/2}|C_{jm}|^{1/2}} \exp\left(-\frac{1}{2}(\vec{o}_t - \mu_{jm})^T C_{jm}^{-1}(\vec{o}_t - \mu_{jm})\right) \end{aligned} \tag{10}$$

is the state output probability with diagonal covariance matrix, and $\gamma_{jm}(\vec{o}_t) = c_{jm} b_{jm}(\vec{o}_t) b_j^{-1}(\vec{o}_t)$.

For LDA transformation in New-DRFE, we give three versions. Version 1 is similar to the condition in Conventional-DRFE. $\vec{x}_t$ denotes the input static and dynamic log filterbank energies. Then $\vec{o}_t = W\vec{x}_t$. The updating formula is the same as formula (9).

In version 2 and version 3, we consider the static features and the dynamic features separately. Let $\vec{x}_t$ denote the static log filterbank energies, $\Delta\vec{x}_t$ and $\Delta\Delta\vec{x}_t$ denote the first and second order derivatives of $\vec{x}_t$. In version 2, we use the same transformation to transform them. Then the new, transformed static feature vector is given by: $\vec{y}_t = W\vec{x}_t$, and the dynamic features of $\vec{y}_t$ are: $\Delta\vec{y}_t = W\Delta\vec{x}_t$ and $\Delta\Delta\vec{y}_t = W\Delta\Delta\vec{x}_t$. The final feature $\vec{o}_t$ is composed of $\vec{y}_t$, $\Delta\vec{y}_t$, $\Delta\Delta\vec{y}_t$, log energy and its derivatives. Then we get:

$$\begin{aligned} \frac{\partial g(O, S, \Phi)}{\partial W} &= \sum_{t=1}^{T} \delta(q_t - j) b_j^{-1}(\vec{o}_t) \frac{\partial b_j(\vec{o}_t)}{\partial W} \\ &= -\sum_{t=1}^{T} \delta(q_t - j) \sum_{m=1}^{M} \gamma_{jm}(\vec{o}_t) \left\{\begin{array}{l} C_{jm}^{-1}(\vec{y}_t - \mu_{jm})\vec{x}_t^T \\ + \Delta C_{jm}^{-1}(\Delta\vec{y}_t - \Delta\mu_{jm})\Delta\vec{x}_t^T \\ + \Delta\Delta C_{jm}^{-1}(\Delta\Delta\vec{y}_t - \Delta\Delta\mu_m)\Delta\Delta\vec{x}_t^T \end{array}\right\} \end{aligned} \tag{11}$$

In version 3, we use different transformations to transform $\vec{x}_t$, $\Delta\vec{x}_t$ and $\Delta\Delta\vec{x}_t$, So: $\vec{y}_t = W\vec{x}_t$, $\Delta\vec{y}_t = \Delta W\Delta\vec{x}_t$ and $\Delta\Delta\vec{y}_t = \Delta\Delta W\Delta\Delta\vec{x}_t$. We get:

$$\frac{\partial g(O,S,\Phi)}{\partial W} = \sum_{i=1}^{T} \delta(q_t - j) b_j^{-1}(\vec{o}_t) \frac{\partial b_j(\vec{o}_t)}{\partial W} \quad (12)$$
$$= -\sum_{t=1}^{T} \delta(q_t - j) \sum_{m=1}^{M} \gamma_{jm}(\vec{o}_t) \left[C_{jm}^{-1} \left(\vec{y}_t - \mu_{jm} \right) \vec{x}_t^T \right]$$

$$\frac{\partial g(O,S,\Phi)}{\partial \Delta W} = \sum_{i=1}^{T} \delta(q_t - j) b_j^{-1}(\vec{o}_t) \frac{\partial b_j(\vec{o}_t)}{\partial \Delta W} \quad (13)$$
$$= -\sum_{t=1}^{T} \delta(q_t - j) \sum_{m=1}^{M} \gamma_{jm}(\vec{o}_t) \left[\Delta C_{jm}^{-1} \left(\Delta \vec{y}_t - \Delta \mu_{jm} \right) \Delta \vec{x}_t^T \right]$$

Similar derivations for $\Delta\Delta W$ can be easily accomplished.

3. EXPERIMENTAL RESULTS

We test our methods on TiDigits, a speaker independent, connected digit utterances database. The speech signal was recorded from various regions of the United States. The database contains 12549 strings for training and 12547 strings for testing. The digits string has a random length from 1 to 7. The model we used is a 10-state, whole-word based HMM model. A 3-state silence and a 1-state short pause models were added. Each HMM-state was chosen as a class used to get the LDA transformation.

Since DFE has two kinds of trainable parameters: the parameters of HMM models and the parameters of feature extractor, the following training schemes were investigated:

- DFE-M: MCE training of the HMM model parameters only;
- DFE-F: MCE training of the transformation parameters only;
- DFE-FM: MCE training of the transformation parameters and the HMM model parameters simultaneously;

For the three versions of New-DRFE, we currently only test version 2 in our experiments. The experiments on version 1 and version 3 of New-DRFE will be tested in the future.

All the experimental results given in the following are represented by Word Error Rate (WER).

3.1. Dimensionality reduction of Conventional-DRFE

We use LDA transformation to transform the original 39 dimension features (12 for static MFCCs, 1 for log energy and their first and second order derivatives) to the new 13 and 26 dimension features. The results of 2 mixtures (where 39-MLE and 39-DFE-M indicate the ML and MCE estimation results of the original 39 dimension features) is shown in table 1. We can see that the MLE-trained system performance is degraded severely after dimensionality reduction. But by using DFE, as we see, there is a significant WER reduction compared with the MLE-trained system, even though only the LDA transformation is adjusted. Furthermore updating models and transformation simultaneously gives a much better result than updating models only. When using DFE-FM, we get slightly better performance in the 26 dimension system than that of the original MCE-trained 39 dimension MFCCs system.

Table 1. % WER of dimensionality reduction of Conventional-DRFE

Dimension	13	26	39
MLE	2.96	2.40	1.81
DFE-F	2.11	1.30	---
DFE-M	1.16	0.86	0.72
DFE-FM	1.00	0.70	---

3.2. Comparison between DCT and LDA in New-DRFE

Here DCT in New-DRFE denotes using DCT to replace LDA in figure 2. We compared the results of by using DCT and by using LDA to transform the log filterbank coefficients in New-DRFE.

Table 2 shows the comparison in different mixtures per state using standard ML estimation. 26 dimension features (12 for transformed static features, 1 for log energy and their first order derivatives) were used. We can see clearly that LDA is better than DCT in MLE-based system.

Table 2. % WER of DCT and LDA in MLE-trained New-DRFE

Transformation	DCT	LDA
1mix	3.00	2.28
2mix	1.80	1.75
4mix	1.37	1.22

In table 3 we can see the comparison results with different training algorithms in 2 mixtures. It is obvious that LDA outperforms DCT, especially in DFE-based system. A WER reduction of 16% is obtained from updating LDA only as DFE-F, in comparison with the DCT-based MLE-training system. WER reduction gets close to 70% when DFE-M or DFE-FM is used. DFE-FM is a little better than DFE-M, with WER reduction at about 5%.

Table 3. % WER of DCT and LDA initialized New-DRFE with different training algorithm

Training algorithm	DCT	LDA
MLE	1.80	1.75
DFE-F	1.74	1.50
DFE-M	0.81	0.57
DFE-FM	0.79	0.54

3.3. Dimensionality reduction of New-DRFE

The results of dimensionality reduction using LDA initialized New-DRFE are shown in table 4. As we see, the performance is significantly improved. Even when we reduce the dimension to 14, the WER is 0.71%, which is comparable to the performance of 0.72% in WER of the MCE-trained 39 dimension MFCCs system. The WER of the 26 dimension system is 0.54%, which has a 25% WER reduction to the MCE-trained 39 dimension MFCCs system.

Compared with the results of DFE-FM in Conventional-DRFE as shown in table 1, we can see that the 26 dimension system of New-DRFE is much better than that of MCE-

optimized Conventional-DRFE. And the 14 dimension system of New-DRFE is as good as the 26 dimension system of MCE-optimized Conventional-DRFE.

Table 4. % WER of dimensionality reduction of New-DRFE

Dimension	10	14	18	22	26
MLE	2.60	1.80	1.77	1.74	1.75
DFE-FM	0.92	0.71	0.65	0.64	0.54

4. DISCUSSION

Instead of using LDA, DCT can also be used in New-DRFE. A method using state-dependent DCT initialized transformations was reported in [4]. Though focusing on feature decorrelation, it can also be used for dimensionality reduction. However, its dimensionality reduction performance is not satisfying. According to [4], the static feature dimension can only be reduced to 12, in order to get an acceptable performance. 12 static features with log energy and their derivatives added, the feature dimension is not reduced (the same as the conventional 39 dimension features: 12 for static MFCCs, 1 for log energy and their first and second order derivatives). In contrast, our method using LDA can reduce the feature dimension much more. The results in section 3.2 also show that the dimensionality reduction performance using DCT is worse than using LDA. That is to say, the effect of MCE-training is sensitive to the initial parameters. The choice of DCT limits its dimensionality reduction performance. Our choice of LDA is more effective.

5. CONCLUSION

In this paper, we use MCE criterion to reduce feature dimension in both Conventional-DRFE and New-DRFE.

In Conventional-DRFE, we apply MCE-optimized dimensionality reduction method to the HMM-based CSR framework. Using this method, we get slightly better performance in the 26 dimension system than that of the MCE-trained 39 dimension MFCCs system.

In our proposed New-DRFE method, a single LDA transformation is used to perform feature decorrelation and dimensionality reduction simultaneously. This LDA transformation together with the classification parameters can be optimized by MCE criterion. This New-DRFE method can get significant performance improvement on TiDigits. Compared with the original MCE-trained 39 dimension MFCCs system, 25% WER reduction is achieved in the new 26 dimension system and comparable performance can even be got in the new 14 dimension system. This method also outperforms our MCE-optimized conventional dimensionality reduction method. In addition, our experimental results show LDA as the initial transformation is a reasonable choice.

We don't use big number of model mixtures in our experiment because of the heavy computation load; future work will be done to get the improvement results of the best possible models. Another future work is to compare our result with other improved projection methods.

REFERENCES

[1] R.O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis,* John Wiley and Sons, New York, 1973.

[2] W. Chou, "Discriminant-Function-Based Minimum Recognition Error Rate Pattern-Recognition Approach to Speech Recognition," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1201-1223, August 2000.

[3] A. Biem, S. Katagiri, E. McDermott, and B.H. Juang, "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No.2, pp.96-110, February 2001.

[4] R. Chengalvarayan, and L. Deng, "HMM-Based Speech Recognition Using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, pp.243-256, May 1997.

[5] R. Chengalvarayan, and L. Deng, "Use of Generalized Dynamic Feature Parameters for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, pp.232-242, May 1997.

[6] X.C. Wang, K.K. Paliwal, "Feature Extraction and Dimensionality Reduction Algorithms and their Applications in Vowel Recognition," *Pattern Recognition*, 36, pp. 2429-2439, 2003.

[7] E. Batlle, C. Nadeu, and J.A.R.Fonollosa, "Feature Decorrelation Methods in Speech Recognition: A Comparative Study," *Proc. ICSLP*, Vol. 3, pp. 951-954, 1998.