# EMPIRICAL ERROR RATE MINIMIZATION BASED LINEAR DISCRIMINANT ANALYSIS

*Hung-Shin Lee and Berlin Chen*

Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taiwan

## ABSTRACT

Linear discriminant analysis (LDA) is designed to seek a linear transformation that projects a data set into a lower-dimensional feature space while retaining geometrical class separability. However, LDA cannot always guarantee better classification accuracy. One of the possible reasons lies in that its formulation is not directly associated with the classification error rate, so that it is not necessarily suited for the allocation rule governed by a given classifier, such as that employed in automatic speech recognition (ASR). In this paper, we extend the classical LDA by leveraging the relationship between the empirical classification error rate and the Mahalanobis distance for each respective class pair, and modify the original between-class scatter from a measure of the squared Euclidean distance to the pairwise empirical classification accuracy for each class pair, while preserving the lightweight solvability and taking no distributional assumption, just as what LDA does. Experimental results seem to demonstrate that our approach yields moderate improvements over LDA on the large vocabulary continuous speech recognition (LVCSR) task.

***Index Terms***— Feature extraction, Pattern classification, Speech recognition

## 1. INTRODUCTION

There are two primary reasons why linear discriminant analysis (LDA) has been widely used in replace of (or in conjunction with) the conventional MFCC-based feature extraction method. First, to reduce the model complexity for lower time and space consumption, LDA can be used to project a higher-dimensional speech feature vector, usually formed by splicing several consecutive frames for capturing the contextual information of speech signals, into a lower-dimensional subspace with a minimal loss in discrimination. Second, it has a desirable characteristic – its derivation is succint and fast without requiring any iterative optimization techniques.

The basic idea behind LDA is to find a transformation matrix that maximizes the ratio of the between-class scatter of a given data set, which represents the class separability in a geometrical sense [1], to the within-class scatter, which can be also taken as a constraint for metric scaling [2], in a projected feature space. As can be proved, LDA in a sense may be thought of as a procedure that maximizes the average squared Mahalanobis distance between each class-mean pair after dimensionality reduction, rather than as an *ad hoc* design for maximizing the classification accuracy, while assuming that all classes share the same within-class covariance. In other words, it implies that LDA does not *directly* relate itself to the classification error rate, the figure of merit that we are interested in most pattern classification tasks. More precisely speaking, in the $C$-class homoscedastic case, LDA can indeed extract statistically optimal features in the $(C-1)$ -dimensional subspace for the Bayesian classifier [3], whereas in the $p$-dimensional subspace $(p < C-1)$, the so-called "overemphasis" problem, where the derivation of the LDA transformation is mostly dominated by well-separated class pairs, will arise and contrarily make the classification performace deteriorated [4].

To alleviate the aforementioned problem and retain the lightweight solvability simultaneously, a considerable amount of research effort has been devoted to integrating various kinds of weighting functions into the between-class scatter to adjust the contribution of each class pair made to the derivation of LDA. The weighting function can be determined on the basis of the theoretical pairwise Bayes error rate occurring between any two Gaussian-distributed classes (Loog's method) [4], or the empirical classification error rate resulting from a given classifier (Lee's method) [5]. Although Loog has skillfully converted the problem of maximum class separation to that of minimum Bayes errors, it still has its limitation that the allocation rules, governed by some complicated classifiers, such as that adopted in hidden Markov model (HMM) based automatic speech recognition (ASR), are not always set in a strict Bayesian sense. To get rid of the deficiency latent within Loog's method, Lee incorporated the empirical classification information gathered from the training data into the derivation of LDA to make its objective function more classifier-related. Nevertheless, in his approach, the contributions from the empirical classification error rates and the distances between class pairs cannot be traded off in an analytical way.

In this paper, we first provide a practical investigation of the relationship between the empirical classification error rates and the Mahalanobis distances of the respective class pairs for ASR. Moreover, we propose a novel reformulation of LDA, called the *approximate pairwise empirical accuracy criterion* (aPEAC), which attempts to approximate the average empirical classification accuracy between each class pair, but not merely the geometrical class separation. It is worthwhile to highlight three key aspects of aPEAC here:

1. The measurement of class separation used in aPEAC is a pairwise classification accuracy function.
2. aPEAC can be well-conducted on some classification systems, which adopt not merely the Bayesian-based classifier but instead the other special or more complicated ones. Phrased another way, as a mediator between the front-end feature extractor and the back-end recognizer, it can mitigate their inconsistency or mismatch.
3. Inheriting from LDA, aPEAC makes no distributional assumption and retains the computational simplicity

A detailed account on the theoretical property of aPEAC will be given in the rest of this paper.

## 2. LDA AND RELATED WORK

### 2.1. Classical LDA

Let $\mathbf{S}_B$ and $\mathbf{S}_W \in \Re^{n\times n}$, respectively, denote the between-class and within-class scatter matrices for a data set of $C$ classes and are defined as follows [6]:

$$\mathbf{S}_B = \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T,\ \mathbf{S}_W = \sum_{i=1}^{C} p_i \mathbf{S}_i, \qquad (1)$$

Here, $p_i$, $\mathbf{m}_i$, and $\mathbf{S}_i$ denote the prior probability, mean, and covariance matrix of class $i$, respectively. The goal of LDA is to seek a linear transformation $\mathbf{\Theta} \in \Re^{n\times p}$ that reduces the dimensionality of a given $n$-dimensional feature vector to $p$ $(p < n)$ by maximizing the following discrimination criterion [7]:

$$J_{LDA}(\mathbf{\Theta}) = \mathrm{trace}\left((\mathbf{\Theta}\mathbf{S}_W\mathbf{\Theta})^{-1}(\mathbf{\Theta}\mathbf{S}_B\mathbf{\Theta})\right), \qquad (2)$$

where the column vector $\boldsymbol{\theta}_i$ of $\mathbf{\Theta}$ can be solved as a generalized eigen-analysis problem $\mathbf{S}_B\boldsymbol{\theta}_i = \lambda_i\mathbf{S}_W\boldsymbol{\theta}_i$ with $\boldsymbol{\theta}_i$ being the eigenvector associated with the $i$-th largest eigenvalue $\lambda_i$ of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

There are two notes that need be mentioned here. First, the optimization of LDA can also be viewed as a two-stage procedure [5]. The first stage conducts a whitening transformation $\mathbf{S}_W^{-1/2}$ on the feature vectors, and the second stage involves a principal component analysis (PCA) on the whitened class means. In the whitened space, the within-class scatter matrix turns out to be an identity matrix, and all discrimination information resides only in the following whitened between-class scatter matrix:

$$\widehat{\mathbf{S}}_B = \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j (\widehat{\mathbf{m}}_i - \widehat{\mathbf{m}}_j)(\widehat{\mathbf{m}}_i - \widehat{\mathbf{m}}_j)^T, \qquad (3)$$

where $\widehat{\mathbf{m}}_i = \mathbf{S}_W^{-1/2}\mathbf{m}_i$. Then it can be proved that the matrix $\widehat{\mathbf{\Theta}}$, which is made up of the eigenvectors corresponding to the $p$ largest eigenvalues of $\mathbf{S}_B$, maximizes the new criterion $\mathrm{trace}(\mathbf{\Theta}^T\widehat{\mathbf{S}}_B\mathbf{\Theta})$. Finally by combining these two stages, the matrix $\mathbf{S}_W^{-1/2}\widehat{\mathbf{\Theta}}$ can be shown to maximize the original LDA criterion in (2).

Second, from a geometrical viewpoint, what LDA attempts to maximize can be interpreted as the average squared Euclidean distance between each whitened class-mean pair in the transformed subspace, since it can be proved that

$$J'_{LDA}(\widehat{\mathbf{\Theta}}) = \mathrm{trace}(\widehat{\mathbf{\Theta}}^T\widehat{\mathbf{S}}_B\widehat{\mathbf{\Theta}}) = \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j \left\|\mathbf{\Theta}^T\widehat{\mathbf{m}}_i - \mathbf{\Theta}^T\widehat{\mathbf{m}}_j\right\|^2. \qquad (4)$$

### 2.2. Weighting-based LDA

As mentioned in Section 1, LDA is not always optimal for a $C$-class classification task due to the "overemphasis" problem. One of the possible solutions is to modify the LDA criterion in the whitened space by replacing $\widehat{\mathbf{S}}_B$ with the following weighted form:

$$\widehat{\mathbf{S}}'_B = \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j w(i,j)(\widehat{\mathbf{m}}_i - \widehat{\mathbf{m}}_j)(\widehat{\mathbf{m}}_i - \widehat{\mathbf{m}}_j)^T, \qquad (5)$$

where $w(i,j)$ is a weighting factor used to control the contribution of a class pair $(i,j)$ made to the derivation of LDA.

Apparently, if the value of $w(i,j)$ is invariant to any nonsingular transformation, *e.g.*, $\mathbf{S}_W^{-1/2}$, then analogous to LDA, the transformation matrix $\mathbf{\Theta}' \in \Re^{n\times p}$ of the weighted LDA can be easily derived by $\mathbf{S}_W^{-1/2}\widehat{\mathbf{\Theta}}'$, where $\widehat{\mathbf{\Theta}}'$ is constituted by the eigenvectors corresponding to the $p$ largest eigenvalues of $\widehat{\mathbf{S}}'_B$.

In Loog's method [4], based on the minimization of the theoretical pairwise Bayes error rate occurring between any two Gaussian-distributed classes $i$ and $j$, $w(i,j)$ is defined by

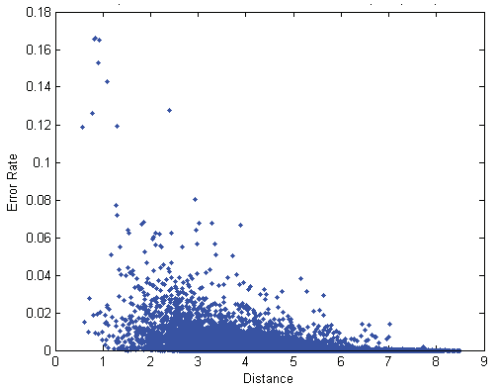


**Figure 1**: *A dot plot of the empirical classification error rates versus the corresponding Mahalanobis distances of feature class-mean pairs for the features of speech training data.*

$$w(i,j) = \frac{1}{2\Delta_{ij}^2}\mathrm{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right),\ \Delta_{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{S}_W^{-1}(\mathbf{m}_i - \mathbf{m}_j)}, \qquad (6)$$

where erf(·) denotes the error function that is twice the integral of the Gaussian distribution with zero-mean and variance of 1/2. It can be shown that, for a given pair of classes $i$ and $j$, $\Delta_{ij}$ represents their Mahalanobis distance in the original space, or the Euclidean distance in the whitened space, *i.e.*, $\Delta_{ij} = \left\|\widehat{\mathbf{m}}_i - \widehat{\mathbf{m}}_j\right\|$. Also, $w(i,j)$ in (6) in essence is a monotonically decreasing function of $\Delta_{ij}$ such that those class pairs with large $\Delta_{ij}$ will not be overemphasized. Furthermore, with this weighting function, the new criterion, named as *approximate pairwise accuracy criterion* (aPAC), can approximate the average *theoretical* accuracy among all class pairs.

Yet another modification (Lee's method) [5], advocating the concept of pairwise class confusion information, takes account of the *practical* classification error rates resulting from a given classifier, rather than operating merely in the Bayesian sense (*cf.* (6)). The corresponding weighting factor is expressed as a function of the pairwise confusion information $CI_{ij}$:

$$w(i,j) = \alpha + (1-\alpha)\times CI_{ij},\ 0 \le \alpha \le 1, \qquad (7)$$

where $CI_{ij}$ represents the empirical classification error rate for a sample item that belongs to class $i$ but is misclassified to class $j$, and $\alpha$ denotes an adjustable factor trading off between the empirical classification error rates and the class-mean distances, which can only be set heuristically.

## 3. THE PROPOSED APPROACH

Inspired by the work mentioned above, we attempt to introduce a new weighting factor such that the distance information and confusion information are more tightly coupled. Moreover, we expect that the whole new criterion can approximate the average empirical classification accuracy among all class pairs.

### 3.1. Observations

First, we define the empirical pairwise classification error rate as

$$ER_{ij} = \frac{e_{ij} + e_{ji}}{n_i + n_j},\ 1 \le i < j \le C, \qquad (8)$$

where $n_i$ denote the number of sample items of class $i$, and $e_{ij}$ denotes the number of sample items that originally belong to class $i$ but are misallocated to class $j$ by the ASR. $ER_{ij}$, in a sense, can be used to measure the confusability between any class pair $i$ and $j$.

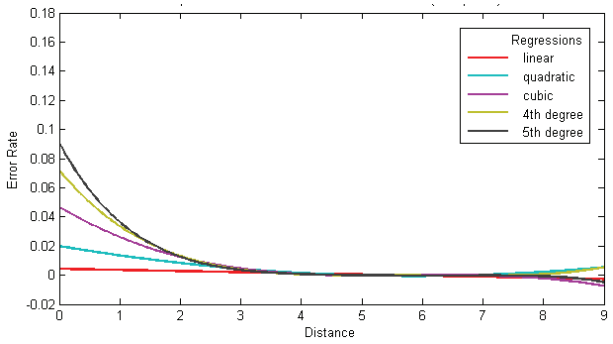


**Figure 2**: *Plots of the polynomial regressions of various degrees based on the data points in Figure 1.*

That is, for class *i*, the higher the value of $ER_{ij}$ , the more confusable it would be with class *j*.

Figure 1 shows a dot plot of the empirical pairwise (phone) classification error rates (*cf.* Eq. (8)), which are obtained through the use of the LDA-transformed speech features on the speech recognition task, versus the corresponding Mahalanobis distances of class-mean pairs, which are measured in the original feature space. Refer to Figure 1, we can roughly characterize the relationship between these two variables: class pairs with shorter distances (*e.g.* $\Delta_{ij} < 4$ ) tend to have higher error rates (*e.g.* $ER_{ij} > 0.01$ ); similarly, class pairs with larger distances (*e.g.* $\Delta_{ij} > 4$ ) are likely to have lower error rates (*e.g.* $ER_{ij} < 0.01$ ). Therefore, such a phenomenon, to some extent, not only gives us researchers, who stand behind the "veil of ignorance" with respect to the back-end classification results, some tractable clues for prediction, but also confirms our expectation: the statistics contributed by the class pairs with shorter distances need to be emphasized, while those of the class pairs with larger distances should be deemphasized instead when deriving the LDA matrix.

### 3.2. The proposed criterion (aPEAC)

To appropriately model the phenomenon revealed in Figure 1, we use the data-fitting (or regression) scheme to find out a function taking the Mahalanobis distance as the input, *i.e.*, $\mathrm{E}(\Delta_{ij})$ , which hopefully can approximate the relationship between the empirical pairwise classification error rate and the corresponding Mahalanobis distance. Data fitting is a mathematical optimization method which, when given a series of data points $(u_i, v_i)$ with $i = 1,\ldots,n$ , attempts to find a function $G(u_i)$ whose output $\tilde{v}_i$ closely approximates $v_i$ . That is, it minimizes the sum of the squared error (or the squares of the ordinate differences) between the points $(u_i, \tilde{v}_i)$ and their corresponding points $(u_i, v_i)$ in the data set.

In our work, for example, if $\mathrm{E}(\Delta_{ij})$ is supposed to be a quadratic polynomial, *i.e.*, $\mathrm{E}(\Delta_{ij}) = a\Delta_{ij}^2 + b\Delta_{ij} + c$ , then given all of the data points $(\Delta_{ij}, ER_{ij})$ shown in Figure 1, we can estimate the coefficients of $\mathrm{E}(\Delta_{ij})$ , *a*, *b*, and *c*, by minimizing the sum of the squares of $(\mathrm{E}(\Delta_{ij}) - ER_{ij})$ shown below:

$$\{\hat{a},\hat{b},\hat{c}\} = \arg\min_{a,b,c} \frac{1}{2}\sum_{i,j=1}^{C}\left(\left(a\Delta_{ij}^2 + b\Delta_{ij} + c\right) - ER_{ij}\right)^2 \qquad (9)$$

The derived error function $\hat{\mathrm{E}}(\Delta_{ij})$ , as graphed in Figure 2 for polynomials of degrees 1 up to 5, can be used to predict the pairwise classification error rate, and to approximate the empirical accuracy for any class pair *i* and *j*, defined by $\hat{\mathrm{A}}(\Delta_{ij}) = 1 - \hat{\mathrm{E}}(\Delta_{ij})$ . This will lead to a new weighting function designed by

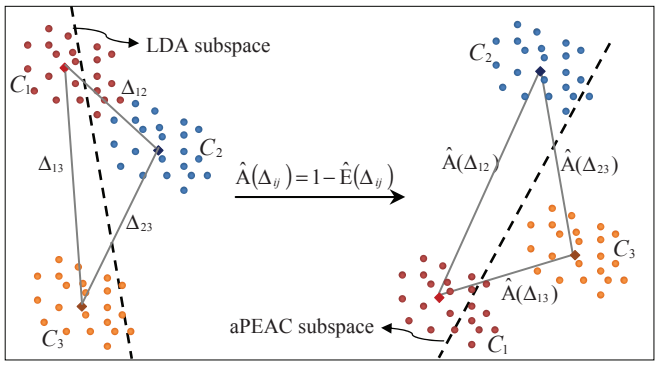


**Figure 3**: *Geometrical interpretation of* aPEAC.

$$w(i,j) = \frac{\hat{\mathrm{A}}(\Delta_{ij})}{\Delta_{ij}^2} = \frac{1 - \hat{\mathrm{E}}(\Delta_{ij})}{\Delta_{ij}^2} \qquad (10)$$

which is obviously invariant to the whitening transformation $\mathbf{S}_W^{-1/2}$ . To see how reasonable our proposed weighting function is, we can first simply substitute the right-hand part in (10) into (5) to form the *approximate pairwise empirical accuracy criterion* (aPEAC) in the whitened space. As can be seen from the following derivation, what aPEAC tries to do is maximizing the average pairwise empirical error rate in the transformed subspace with a specific dimensionality, as opposed to LDA that tries to maximize the weighted sum of the squared Euclidean distance in the whitened space.

$$\begin{aligned} J_{aPEAC}(\hat{\boldsymbol{\Theta}}) &= \mathrm{trace}\left(\frac{1}{2}\sum_{i,j=1}^{C} p_i p_j \frac{1 - \hat{\mathrm{E}}(\Delta_{ij})}{\Delta_{ij}^2} \hat{\boldsymbol{\Theta}}^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\boldsymbol{\Theta}}\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j \left(1 - \hat{\mathrm{E}}(\Delta_{ij})\right) \mathrm{trace}\left(\frac{\hat{\boldsymbol{\Theta}}^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)}{\Delta_{ij}} \frac{(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\boldsymbol{\Theta}}}{\Delta_{ij}}\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j \left(1 - \hat{\mathrm{E}}(\Delta_{ij})\right) \mathrm{trace}\left(\frac{(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \hat{\boldsymbol{\Theta}}}{\Delta_{ij}} \frac{\hat{\boldsymbol{\Theta}}^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)}{\Delta_{ij}}\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j \left(1 - \hat{\mathrm{E}}(\Delta_{ij})\right) \end{aligned}$$

Besides, the role that the weighting function plays is not merely to control the contribution of each class pair, but to endow such a criterion with more preferable notions for classification problems. As illustrated in Figure 3, which shows a geometrical comparison between LDA and aPEAC, the subspace (the projecting direction, the dashed line) generated by LDA is dominated by the class pairs (*e.g.* classes 1 and 3) with large Mahalanobis distance in the original space. However, in aPEAC, the subspace is instead dominated by the class pairs (*e.g.* classes 1 and 2) with large empirical classification accuracy in the tranformed space instead.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Experimental setup

The speech corpus for LVCSR consists of about 200 hours of MATBN Mandarin television news [8]. All the 200 hours of speech data are equipped with corresponding orthographic transcripts, in which about 25 hours of speech data were used to bootstrap the acoustic training. Another set of 1.5 hour speech data of were reserved for testing. On the other hand, the acoustic models chosen here for speech recognition are 112 right-context-dependent INITIAL's and 38 context-independent FINAL's. The

| Polynomial Regressions | without MLLT | with MLLT |
|---|---|---|
| Linear (1st degree) | 30.62 | 28.37 |
| Quadratic (2nd degree) | 30.40 | 28.15 |
| **Cubic (3rd degree)** | **30.59** | **27.80** |
| 4th degree | 30.69 | 28.57 |
| 5th degree | 30.47 | 28.03 |

**Table 1.** *The CER results (%) of* aPEAC*, with respect to various degrees of polynomials.*

| Methods | without MLLT | with MLLT |
|---|---|---|
| LDA | 31.44 | 28.95 |
| **aPEAC (3rd degree)** | **30.59** | **27.80** |
| Loog's (Eq. (6)) | 30.39 | 28.51 |
| Lee's (Eq. (8), $\alpha = 0.5$) | 30.76 | 28.17 |
| HLDA | 44.56 | 28.38 |

**Table 2.** *Comparison among the CER results (%) of* aPEAC *and various LDA-based approaches.*

acoustic models were trained using the Expectation-Maximization (EM) algorithm.

The recognition lexicon consists of 72K words. The language models used in this paper consist of unigram, bigram and trigram models, which were estimated using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA). The *N*-gram language models were trained using the SRI Language Modeling Toolkit (SRILM).

The baseline system with the Mel-frequency cepstral coefficient (MFCC) features resulted in a character error rate (CER) of 32.16 %.

### 4.2. Experimental results

The feature extraction was performed using LDA, aPEAC, and other methods on speech feature vectors consisting of 162 dimensions, which were first spliced by every 9 consecutive 18-dimensional Mel-filterbank feature vectors and then reduced to 39 dimensions. The states of each HMM were taken as the unit for class assignment, and a well-trained HMM-based recognition system was performed to obtain the class alignment of the training utterances. During the speech recognition process we kept track of full state alignment for obtaining the state-level transcriptions of the training data; by comparison with the correct ones, we thus derived the empirical pairwise classification error rates with (8).

Table 1 shows the results for aPEAC with polynomial regressions of various degrees, which are derived from the training data and its preliminary recognition results. The experiments on speech recognition were further performed in conjunction with a heteroscedastic decorrelation, namely the maximum likelihood linear transform (MLLT) [9], which is used for obtaining a projection maximizing the likelihood of the projected data under a diagonal-covariance assumption. As the 3th degree polynomial regression is being used, aPEAC yields the lowest CER, which has relative improvements of about 14 % over the system with MFCC baseline and about 4 % over LDA. The possible reasons why aPEAC did not significantly outperform LDA can be conjectured as follows:

1. In Figure 1, with the same limitation as polynomial regression, there are still many outliers, which lie far apart from the regression curves, especially in the vertical direction, and aPEAC cannot deal with them well. Practically speaking, other irregular or insignificant factors, which might affect the generations of the polynomial regressions, need to be expelled.
2. We can observe from Figure 2 that the variations of those polynomials are very slight when $\Delta_{ij} > 3$. But in Figure 1, there are still many class pairs with $\Delta_{ij} > 3$ but also with higher $ER_{ij}$ that might be deemphasized by the proposed weighting functions.

We then compare aPEAC with the two variants of LDA mentioned in Section 2.2, as those expressed in (6) and (7), respectively. If we look into the results summarized in Table 2 (the right-most column), we can observe that they also provide moderate improvements over LDA, which, however, seem slightly inferior to that achieved by aPEAC (3rd degree) for the LVCSR task studied here. Moreover, we also try to compare aPEAC with HLDA (heteroscedastic linear discriminant analysis) [10], one of the widely-used feature transformation approaches in ASR, which is usually derived on a maximum likelihood basis while relaxing the equal-covariance assumption of LDA. As can be seen from the last row of Table 2, the improvement of HLDA is less pronounced than aPEAC (3rd degree).

## 5. CONCLUSIONS

In this paper, we have proposed a new LDA-based criterion called aPEAC to generalize LDA by modulating the contribution of each class pair on between-class scatter through the use of a pairwise classification error function. aPEAC has successfully converted the original LDA criterion of distance maximization to that of empirical error rate minimization.

As part of our future work, we are going to make an in-depth comparison with other well-known but more complicated methods, such as fMPE [11]. Furthermore, not only will a combination of both theoretical and empirical parts of the classification accuracy be considered, but also the within-class scatter matrix will be reestimated on the basis of our proposed approach.

## 6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC96-2628-E-003-015-MY3, NSC95-2221-E-003-014-MY3, and NSC97-2631-S-003-003.

## 7. REFERENCES

[1] S. S. Wilks, *Mathematical Statistics*. New York: John Wiley & Sons, 1962, pp. 577-578.

[2] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988, pp. 298-300.

[3] O. C. Hamsici and A. M. Martinez, "Bayes Optimality in Linear Discriminant Analysis," *IEEE Trans. on PAMI*, vol. 30, no. 4, pp. 647-657, 2008.

[4] M. Loog et al., "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria," *IEEE Trans. on PAMI*, vol. 23, no. 2, 2001.

[5] H. S. Lee and B. Chen, "Linear Discriminant Feature Extraction Using Weighted Classification Confusion Information," in *Proc. Interspeech*, 2008.

[6] Y. Li et al., "Weighted Pairwise Scatter to Improve Linear Discriminant Analysis," in *Proc. ICSLP*, 2000.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990, pp. 446-447.

[8] B. Chen et al., "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. ICASSP*, 2004.

[9] R. A. Gopinath, "Maximum Likelihood Modeling with Gaussian Distributions for Classification," in *Proc. ICASSP*, 1998.

[10] N. Kumar and A. G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition," *Speech Communication*, vol. 26, 1998.

[11] D. Povey et al., "fMPE: Discriminatively Trained Features for Speech Recognition," in *Proc. ICASSP*, 2005.