Linear Discriminant Feature Extraction Using Weighted Classification Confusion Information

Hung-Shin Lee and Berlin Chen

Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
g95470153@csie.ntnu.edu.tw, berlin@csie.ntnu.edu.tw

Abstract
Linear discriminant analysis (LDA) can be viewed as a two-stage procedure geometrically. The first stage conducts an orthogonal and whitening transformation of the variables. The second stage involves a principal component analysis (PCA) on the transformed class means, which is intended to maximize the class separability along the principal axes. In this paper, we demonstrate that the second stage does not necessarily guarantee better classification accuracy. Furthermore, we propose a generalization of LDA, weighted LDA (WLDA), by integrating the empirical classification confusion information between each class pair, such that the separability and the classification error rate can be taken into consideration simultaneously. WLDA can be efficiently solved by a lightweight eigen-decomposition and easily combined with other modifications to the LDA criterion. The experiment results show that WLDA can yield a relative character error reduction of 4.6% over LDA on the Mandarin LVCSR task.

Index Terms: speech recognition, feature extraction, linear discriminant analysis, confusion information

1. Introduction
Linear discriminant analysis (LDA) has been extensively used to obtain discriminative acoustic features for a wide variety of speech sound classification problems. It aims to derive a transformation matrix, projecting feature vectors from an n-dimensional space to a p-dimensional space (p < n), that can maximize the ratio of between-class scatter to within-class scatter [1]. In practice, LDA is usually taken as a standard technique for projecting a spliced feature vector, which is formed by concatenating the features from a fixed number of frames immediately neighboring to the observation vector, into a lower dimensional space with a minimal loss in discrimination. Thus, the hidden Markov models (HMMs) can be more precisely modeled due to the additional incorporation of the temporal information, and the number of HMM parameters can be further reduced to avoid the problem of the curse of dimensionality, especially for problems having small sample sizes. It has been shown that LDA can lead to consistent performance improvements for small-vocabulary recognition tasks and mixed results on large-vocabulary applications [2].

In the recent past, several attempts, e.g. HLDA [3], HDA [4] and PLDA [5], have been made to relax the well-known assumption of LDA that all classes share the same within-class covariance [6]. Besides, Li argued that the canonical between-class scatter expressed in the original objective function of LDA in some sense assumes that each class is equally confusable with all other classes and LDA is likely to ignore the class discrimination information for class pairs which are close to each other [7]. They defined a new weighted pairwise scatter based on the Euclidean distance metric to moderate the contributions of large-distance class pairs on the derivation of the transformation matrix. Coincidentally, Loog also made a similar statement [8]. They demonstrated that LDA tends to preserve distances of class pairs that are already well separated, which might have the side effect of resulting in a large overlap of classes that are already well separated. More precisely, they introduced a weighting function to adjust the contributions of the class pairs based on the minimum Bayes error criterion, which can be conducted on both homoscedastic and heteroscedastic models. In other words, the problem of maximum class separation in LDA has been skillfully converted to the problem of minimum Bayes errors.

From the aforementioned two research works, we can see that the original LDA criterion does not necessarily guarantee better classification accuracy, whereas trying to maximize the class separation. Essentially, LDA is only optimal for Gaussian distributed classes with equal covariance [6]. If such a condition does not hold, the derivation of the LDA transformation might be overwhelmed with the “over-emphasis” problem caused by the large-distance class pairs. In this paper, we attempt to shed light on this problem with a geometrical illustration, and explore an appropriate modification of the LDA criterion such that contributions made by the large-distance class pairs can be properly deemphasized by taking their empirical classification errors into consideration.

The rest of this paper is organized as follows. Section 2 reviews the classical LDA and provides a geometric illustration of LDA [9]. Then, in Section 3, a novel weighting method, which attempts to relate the derivation of LDA transformation to the empirical classification error rates, is proposed. Then, the experiment settings and a series of experiments conducted on the Mandarin large vocabulary continuous speech recognition (LVCSR) task are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Review of LDA
Theoretically, LDA can be interpreted from two aspects [10]. One comes from the Bayesian decision theory, while the other is Fisher’s LDA (extended to multiclass cases by C. R. Rao). Here only the latter is discussed.

2.1. Fisher’s LDA
Let $S_w \in \mathbb{R}^{n \times n}$ and $S_b \in \mathbb{R}^{n \times n}$, respectively, denote the within-class and between-class scatter matrices for K classes and are defined as follows: (Note that $S_w$ can be expressed in terms of class-mean differences [7].)
The derivation of the LDA transformation matrix can be geometrically viewed as a two-stage procedure [9]. As shown in Fig. 1(a), suppose there are three classes with equal covariance matrix $S_0$ and prior probability in a two-dimensional space. In the first stage, $S_{w/2} = S_{w}^{1/2}S_0 S_{w}^{-1/2}$ is used as an orthogonal and whitening transformation of the original feature vectors (or variables). After the rotation and scaling from the original axes $X_1$ and $X_2$ to the new orthogonal axes $Y_1$ and $Y_2$, the vector components of different dimensions of the resulting vectors become uncorrelated, and the distribution contour for each class turns to be a unit circle (see Fig. 1(b)). The second stage involves a principal component analysis (PCA) on the transformed class means, which seeks new axes that coincide with the directions having the maximum variations of the class means. Thus, in the second stage, the criterion in Eq. (3) becomes

$$J_{s} (Z) = \text{trace} \left( Z' S_{w}^{1/2} S_0 S_{w}^{-1/2} Z \right)$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j \text{trace} \left( Z' S_{w}^{1/2} (m_i - \bar{m})(m_j - \bar{m})' S_{w}^{-1/2} Z \right).$$

(4)

Note that the principal components $z_i$'s (the column vectors of $Z$) are orthonormal. Since the traditional Euclidean distance is an appropriate distance measurement between any two classes with the identity covariance matrix, the between-class scatter, which has been transformed by $S_{w/2}$, can be precisely used to derive the principal component $z_i$ (the dashed line in Fig. 1(b)). Maximizing the criterion defined in Eq. (4) is equivalent to solving a typical eigenvector problem of $S_{w/2}^{-1} S_{0} S_{w}^{-1/2}$ with

$$\left( S_{w}^{-1/2} S_0 S_{w}^{-1/2} \right) z = \lambda z,$$

(5)

To get back the transformation matrix $\Theta$ for the original variables, $Z$ has to be premultiplied by $S_{w}^{1/2}$. Actually, it can be algebraically proven that $S_{w}^{1/2} Z$ also maximizes the LDA criterion in the original feature space (as expressed in Eq. (3)) by showing that $S_{w}^{1/2} Z$ is the eigenvector matrix of $S_{w}^{1/2} S_0 S_{w}^{-1/2}$. As follows:

$$S_{w}^{1/2} S_0 S_{w}^{-1/2} Z = \Lambda' Z,$$

(6)

where $\Lambda'$ is the diagonal matrix with eigenvalues $\lambda'$’s.

Moreover, it is worth mentioning that, according to the geometric analysis of LDA, at least two possible directions are offered to further generalize LDA. One is to obtain more effective estimates of the within-class scatter, while the other is to modify the between-class scatter for better class discrimination. Here we are inclined to the latter, which can also easily retain the analytical simplicity as LDA does.

2.3. Discussions

From the above illustrations, we can figure out some facts. First, an alternative procedure to find the LDA transformation matrix $\Theta$ can be written as follows:

**Algorithm I.** An alternative procedure of LDA

1. Find matrix $Z_{(m,p)}$, which is made up of the eigenvectors corresponding to the $p$ largest eigenvalues of $S_{w}^{1/2} S_0 S_{w}^{-1/2}$.
2. Derive $\Theta = S_{w}^{1/2} Z$.

Second, from Fig. 1(b), we can observe that the principal component $z_i$ is dominated by classes 2 and 3 because of their larger distance than classes 1 and 2, or 1 and 3. If the optimal conditions, i.e., all classes are Gaussian distributed with the same within-class covariance, do not hold for LDA (see Fig. 2(a)), it may result in a sub-optimal solution since it does not necessarily ensure the minimum of the sum of overlaps between the class distribution contours after the projection, notwithstanding the maximum of the sum of the distance between each class mean pairs. As graphically illustrated in Fig. 2(b), the overlap between classes 1 and 2 along the
principal component \( z_1 \), primarily determined by classes 2 and 3, is much larger than that in Fig. 1(b). That is, LDA will concentrate only on the geometric separability of class means by their distances, but instead ignore entirely the possibility that more classification errors (overlaps) will be produced. Furthermore, LDA cannot necessarily well serve the classifiers, such as the continuous HMMs, whose allocation rules for samples (or feature vectors) are not performed totally based on the comparison of the Euclidean distances between them to all the class means (cf. [11]). Based on the above discussions, it can be concluded that higher geometric separability in the front-end processing made by LDA does not imply a lower empirical classification error rate of the back-end classifier. This also gives a possible reason why LDA would produce mixed results on most LVCSR systems, which adopt continuous HMMs and Gaussian mixtures for acoustic modeling.

3. Generalization of LDA

As discussed above, in order to diminish the overlaps of the class distribution contours that LDA fails to achieve and the inherent inconsistency between LDA and the classifier, a compromise between the distance (for the geometric separability) and the empirical error rate (for the performance of the back-end classifier) for each class pair must be reached. Thus, we propose a novel linear dimensionality reduction method by combining two kinds of information sources: the empirical confusion information and the Euclidean distance for each class pair in a straightforward manner.

3.1. Some preliminaries

Suppose there are \( N \) labeled training samples (or feature vectors) with \( K \) classes. Let \( n_i \) denotes the sample size of class \( i \), and \( E_{ij} \) denotes the number of samples that originally belong to class \( i \) but are misallocated to class \( j \) by the classifier (or recognizer). The confusion matrix \( C_i \in \mathbb{R}^{K \times K} \) is constructed with

\[
C_{ij} = \begin{cases} 
E_{ij} / n_i & \text{if } i \neq j, \\
0 & \text{if } i = j,
\end{cases}
\]

(7)

where \( C_{ij} \) denotes the \((i, j)\)-th entry of \( C_i \), and when \( i \neq j \), it represents the classification error rate for a sample that belongs to class \( i \) but is misclassified to class \( j \). \( C_{ij} \), in a sense, can be used to measure the confusability between classes \( i \) and \( j \). That is, for class \( i \), the higher \( C_{ij} \) the more confusable it would be with class \( j \).

Therefore, we define a generalized between-class scatter

\[
\tilde{S}_g = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} p_i p_j (w(i,j)Z'\tilde{S}_g Z)
\]

(8)

where \( \tilde{S}_g = S_{g1}^{1/2}(m_i - m_j)(m_i - m_j)'S_{g1}^{1/2} \). Note that, \( \tilde{S}_g \) represents a distance matrix after the first-stage transformation, as mentioned in Section 2.2., of which the eigenvalue equals the squared Euclidean distance between the means of classes \( i \) and \( j \) (cf. [11]). The weighting function \( w(i,j) \) is employed to adjust the contribution of \( \tilde{S}_g \). For LDA, \( w(i,j) = 1 \), which means that the contributions are the same for all class pairs.

3.2. Weighted confusion information based LDA

Here, the weighted confusion information based LDA (WLDA) is introduced. In order to take both class separability and empirical classification error rate into consideration simultaneously, we define \( w(i,j) \) as

\[
w(i,j) = \alpha + (1-\alpha)\times C_{ij}
\]

(9)

where \( \alpha (0 \leq \alpha \leq 1) \) is a parameter used to control the ratio of the empirical confusability to the weighting function. The generalized criterion in Eq. (8) becomes

\[
J_{\alpha}(Z) = \text{trace}\left\{ \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} p_i p_j (\alpha + (1-\alpha)\times C_{ij})Z'\tilde{S}_g Z \right\}
\]

(10)

\( \alpha \) and \( C_{ij} \) play important roles in controlling the contribution of \( \tilde{S}_g \) to the optimization criterion expressed in Eq. (10). We can see that if \( \alpha = 1 \), WLDA apparently reduces to LDA; and if \( \alpha = 0 \), not only will the class pairs with \( C_{ij} \neq 0 \) dominate the whole weightings, but the class pairs with \( C_{ij} = 0 \) will get completely ignored. We can suitably set \( \alpha \) as a positive number so that each class pair can be taken into account.

Besides, to a certain extent, \( C_{ij} \) represents the degree of separation (or the undesirable overlap) between classes \( i \) and \( j \) for the classifier. If class \( i \) has a smaller \( C_{ij} \), by implication, then it has been well-separated from class \( j \) (or has less overlaps with class \( j \) for the classifier), and the weighting \( (\alpha + (1-\alpha)\times C_{ij}) \) also becomes relatively smaller, which makes their distance weighted less. On the contrary, the class pair with larger \( C_{ij} \) deserves a larger weighting on their distance.

Analogous to Algorithm I, the WLDA transformation matrix \( \Theta \) can be easily derived by the eigen-analysis, and we describe the algorithm as follows:

Algorithm II. WLDA

1. Classify the labeled training samples with \( K \) groups that have been transformed by LDA, and derive the confusion matrix \( C_i \in \mathbb{R}^{K \times K} \).
2. Find matrix \( Z_{(m \times p)} \), which is made up of eigenvectors corresponding to the \( p \) largest eigenvalues of

\[
\frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} p_i p_j (\alpha + (1-\alpha)\times C_{ij})\tilde{S}_g
\]

3. Derive \( \Theta = S_{g1}^{1/2}Z \).

3.3. Combination with the other method

In addition to the merit of simplicity for implementation, WLDA can be easily combined with the other modifications to LDA for better classification accuracy. For example, Loog [8] replaced the between-class scatter in Eq. (4) with a mean pairwise and Bayes-error based criterion, which we simplify as \( L \) here. We can apply WLDA on the criterion to form a new criterion:

\[
J_{\alpha}(Z) = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} p_i p_j (\alpha + (1-\alpha)\times C_{ij})\text{trace}(Z'\tilde{L}Z).
\]

(11)

where \( C_{ij} \) is derived based on the training samples transformed by Loog’s method alone.

4. Experiments and results

4.1. Experiment setup

The speech corpus consists of about 200 hours of MATBN Mandarin television news [12]. 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 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 acoustic models were trained with the maximum likelihood criterion.

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 speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree search as well as a lexical prefix tree organization of the lexicon. The recognition hypotheses were organized into a word graph for further language model rescoring [12]. The baseline system with the Mel frequency cepstral coefficient (MFCC) features resulted in a character error rate (CER) of 32.16%.

### 4.2. Experiment results

The feature extraction was performed using LDA, WLDA, Loog’s method [8], and Li’s method [7] for 162-dimensional feature vectors, which were first spliced by every 9 consecutive 18-dimensional filterbank feature vectors and then reduced to 39 dimensions. The states of each HMM were taken as the unit for class assignment [2], and a well-trained HMM-based recognition system was performed to obtain the class alignment of the training utterances. In LDA and Loog’s method, during the speech recognition process we kept track of full state alignment for obtaining the state-level transcriptions of the training data, and by comparison with the correct transcriptions, we derived the confusion matrices by Eq. (7).

Table 1 shows the results for WLDA with the confusion matrix derived after applying LDA. With $\alpha = 0.5$, WLDA yields the lowest CER, which has a relative improvement of 4.6% over the LDA baseline. It is noticeable that with $\alpha = 0$, WLDA did not make much improvement over LDA. One of the reasons may be that after applying WLDA with $\alpha = 0$, the class pair with $C_{ij} = 0$, which got thoroughly ignored, became closer in their reason. Another reason possibly lies in the mismatch between training and test data. Table 2 shows that with a proper setting of the parameter ($\alpha = 0.5$), WLDA can outperform the other modifications to LDA, such as Loog’s method (with $\gamma = 0$) for the homoscedastic model [8], and Li’s method (with $f(t) = 1/t^2$) [7]. As mentioned in Section 3.3, WLDA can also be integrated with the other modifications to LDA for achieving more improvements. Table 3 shows the results that WLDA (with $\alpha = 0.7$) can help Loog’s method (with $\gamma = 0$) for the decrease of CER by 1.14%.

### 5. Conclusions

In this paper, we have proposed a new weighting approach called WLDA to generalize LDA by additionally using the confusion information between each class pair. The experiment results demonstrated that WLDA could successfully combine confusion information with the geometric distance to get better improvements of CER for the classifier. Moreover, WLDA could also be easily and effectively integrated with other methods. Future work includes determining the parameter $\alpha$ analytically and performing WLDA in conjunction with a heteroscedastic decorrelation, such as the maximum likelihood linear transform (MLLT) [13].

### 6. References