特集 Acoustic Feature Transformation Based on Discriminant Analysis Preserving Local Structure for Speech Recognition*

坂井　誠　Makoto SAKAI
北岡教英　Norihide KITAOKA
武田一哉　Kazuya TAKEDA

To improve speech recognition performance, feature transformation based on discriminant analysis has been widely used to reduce the redundant dimensions of acoustic features. Linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA) are often used for this purpose, and a generalization method for LDA and HDA, called power LDA (PLDA), has been proposed. However, these methods may result in an unexpected dimensionality reduction for multimodal data. It is important to preserve the local structure of the data when reducing the dimensionality of multimodal data. In this paper we introduce two methods, locality-preserving HDA and locality-preserving PLDA, to reduce dimensionality of multimodal data appropriately. We also propose an approximate calculation scheme to calculate sub-optimal projections rapidly. Experimental results show that the locality-preserving methods yield better performance than the traditional ones in speech recognition.

Key words：Speech Recognition, Feature Extraction, Multidimensional Signal Processing

1.　INTRODUCTION

Cepstrum-based feature vectors, such as mel frequency cepstral coefficients (MFCCs), are generally used in speech recognition. These feature vectors are estimated over 20 to 30 milliseconds and accurately extract static information from speech signals. In addition to static feature vectors, dynamic information that describes temporal change among several successive features is usually included in the feature vector. Several methods for integrating dynamic information have been proposed [1)2)3)]. The simplest method including dynamic information is to concatenate several successive features into a single feature vector. The concatenated high-dimensional vectors often include nonessential information and incur heavy computational load. Therefore, to reduce dimensionality, a feature transformation method is often applied to concatenated vectors.

Linear discriminant analysis (LDA), also known as Fisher discriminant analysis (FDA), is widely used to reduce dimensionality, and is a powerful tool to preserve discriminative information [4)5)]. LDA assumes that each class shares a common class covariance [6)]. However, this assumption does not necessarily hold for a real data set. In order to overcome the limitation, heteroscedastic discriminant analysis (HDA) has been proposed [7)]. HDA employs individual weighted contributions of the classes for its objective function. In addition, a generalization method for LDA and HDA has been proposed, which is called power LDA (PLDA) [8)].

These methods may result in an unexpected dimensionality reduction if the data in a certain class consist of several clusters, i.e., multimodal, because they implicitly assume that data are generated from a single Gaussian distribution. In speech recognition, speech signals for acoustic model training tend to be multimodal because they are generally collected under various conditions, such as gender, age and noise environment. Therefore, each class such as a phone is generally represented as a Gaussian mixture model (GMM) or HMM whose states are represented by GMMs in a speech recognizer. Hence, dimensionality reduction methods without handling multimodality may give unsatisfactory performance, so a dimensionality reduction method for multimodal data is desired to improve speech recognition performance.

Recently, several methods have been proposed to reduce the dimensionality of multimodal data in the machine learning community [9)-12)]. It is important to preserve the local structure of data in reducing the dimensionality of multimodal data appropriately. Locality preserving projection (LPP) [10)] finds a projection such that the data pairs close to each other in the original space remain close in the projected space. Thus, LPP reduces dimensionality without losing information on local structure. Local Fisher discriminant analysis (LFDA) [11)] is also proposed as a supervised method for mul-

＊一般社団法人 電子情報通信学会の了解を得て，英文論文誌（Information and Systems: D）Vol.E93-D, no.5 MAY 2010より一部加筆して転載

timodal data, while LPP is an unsupervised method. To deal with multimodal data, LFDA combines the ideas of FDA and LPP, maximizes between-class separability and preserves within-class local structure. Thus, LFDA is an extension of LDA to reduce the dimensionality of multimodal data.

Since LFDA is based on LDA which assumes homoscedasticity, the effectiveness of LFDA may be limited. To reduce the dimensionality of multimodal data appropriately, we extend HDA which assumes heteroscedasticity. To deal with multimodal data using HDA, we combine the ideas of LPP and HDA, and propose locality-preserving HDA. In addition, we also propose locality-preserving PLDA. These extensions can be expected to yield better performance because they reduce the dimensionality of multimodal data appropriately.

Locality-preserving methods such as LFDA and the proposed methods require considerable computational time to obtain optimal projections when there are many features. In order to slash time, we propose an approximate calculation scheme. Experimental results show that the locality-preserving dimensionality reduction methods yield better performance than traditional ones.

The paper is organized as follows. Feature transformation methods are reviewed in Section 2. Existing locality-preserving dimensionality reduction methods are reviewed in Section 3. Proposed methods are introduced in Section 4. An approximate calculation to obtain a sub-optimal projection is given in Section 5. Experimental results are presented in Section 6. Finally, conclusions are given in Section 7.

2. LINEAR DIMENSIONALITY REDUCTION METHODS

We formulate the problem of linear dimensionality reduction. Given n-dimensional features $\mathrm{x}_j \in \mathbb{R}^n$ where $j=1,2,...,N,$ e.g., concatenated speech frames, and associated class labels $y_j \in \{1,2,...,\mathrm{K}\}$, e.g., phonemes, let us find a projection matrix $\mathbf{B} \in \mathbb{R}^{n \times p}$ that transforms these features to p-dimensional features $\mathbf{z}_j \in \mathbb{R}^p$, where , $p<n$, $\mathbf{z}_j = \mathbf{B}^\top \mathbf{x}_j$, K denotes the number of classes, and N denotes the number of features. $\mathbf{X}^\top$ denotes the transpose of the matrix **X**. Here, we briefly review existing dimensionality reduction methods. The aim of the techniques are to find a projection matrix B.

2.1 Linear Discriminant Analysis

In LDA, within-class, between-class and mixture covariance matrices are used to formulate its objective function. These covariance matrices are defined as follows 4)5):

$$\mathbf{C}^{(W)} = \frac{1}{N}\sum_{k=1}^{K}\sum_{j:y_j=k}(\mathbf{x}_j - \boldsymbol{\mu}_k)(\mathbf{x}_j - \boldsymbol{\mu}_k)^\top,$$

$$\mathbf{C}^{(B)} = \sum_{k=1}^{K} P_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top,$$

$$\begin{aligned}\mathbf{C}^{(M)} &= \frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^\top \\ &= \mathbf{C}^{(W)} + \mathbf{C}^{(B)},\end{aligned}$$

where $\boldsymbol{\mu}_k$ is the mean of features in class k, $\boldsymbol{\mu}$ is the mean of all features regardless of their class assignments, and P_k is the weight for class k. In general, P_k is empirically given by $P_k=N_k/N$, where N_k is the number of features in class k.

There are several definitions of LDA objective functions. Typical objective functions are the following 4)5):

$$J_{LDA_1}(\mathbf{B}) = \frac{\left|\mathbf{B}^\top\mathbf{C}^{(B)}\mathbf{B}\right|}{\left|\mathbf{B}^\top\mathbf{C}^{(W)}\mathbf{B}\right|}, \quad (1)$$

$$J_{LDA_2}(\mathbf{B}) = \frac{\left|\mathbf{B}^\top\mathbf{C}^{(M)}\mathbf{B}\right|}{\left|\mathbf{B}^\top\mathbf{C}^{(W)}\mathbf{B}\right|}, \quad (2)$$

$$J_{LDA_3}(\mathbf{B}) = tr\left((\mathbf{B}^\top\mathbf{C}^{(W)}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{C}^{(B)}\mathbf{B}\right), \quad (3)$$

where $|\mathbf{X}|$ is the determinant of the matrix **X**, and $tr(\mathbf{X})$ is the trace of the matrix **X**. A projection matrix is obtained to maximize the objective function with respect to **B**. The optimizations of Eqs. (1) to (3) result in the same projection 4).

In Eqs. (1) to (3), within-class scatter, $\mathbf{S}^{(W)}$, between-class scatter, $\mathbf{S}^{(B)}$, and mixture scatter, $\mathbf{S}^{(M)}$, may be employed in place of , $\mathbf{C}^{(W)}$, $\mathbf{C}^{(B)}$ and $\mathbf{C}^{(M)}$, respectively. These scatters are given by $\mathbf{S}^{(W)}=N\mathbf{C}^{(W)}$, $\mathbf{S}^{(B)}=N\mathbf{C}^{(B)}$, and $\mathbf{S}^{(M)}=N\mathbf{C}^{(M)}$. The same solution is obtained even if , $\mathbf{C}^{(W)}$, $\mathbf{C}^{(B)}$ and $\mathbf{C}^{(M)}$ in Eqs. (1) to (3) are replaced with , $\mathbf{S}^{(W)}$, $\mathbf{S}^{(B)}$ and $\mathbf{S}^{(M)}$, respectively.

2.2 Heteroscedastic Discriminant Analysis

HDA uses the following objective function which incorporates individual weighted contributions of the class variances [7]:

$$J_{HDA}(\mathbf{B}) = \prod_{k=1}^{K} \left(\frac{\left|\mathbf{B}^\top \mathbf{C}^{(B)} \mathbf{B}\right|}{\left|\mathbf{B}^\top \mathbf{C}_k \mathbf{B}\right|} \right)^{N_k} \propto \frac{\left|\mathbf{B}^\top \mathbf{C}^{(B)} \mathbf{B}\right|}{\prod_{k=1}^{K} \left|\mathbf{B}^\top \mathbf{C}_k \mathbf{B}\right|^{P_k}}, \tag{4}$$

where $\mathbf{C}_k$ is a class covariance matrix in class k and is given by

$$\mathbf{C}_k = \frac{1}{N_k} \sum_{j:y_j=k} (\mathbf{x}_j - \boldsymbol{\mu}_k)(\mathbf{x}_j - \boldsymbol{\mu}_k)^\top.$$

$\mathbf{C}_k$ and $\mathbf{C}^{(W)}$ satisfy $\mathbf{C}^{(W)} = \sum_{k=1}^{K} P_k \mathbf{C}_k$.

The solution to maximize Eq. (4) is not analytically obtained. Therefore, a numerical optimization technique, such as BFGS [13], is performed to maximize Eq. (4) with respect to **B**.

2.3 Power Linear Discriminant Analysis

We have proposed the following objective function, which integrates LDA and HDA [8)14)†]:

$$J_{PLDA_1}(\mathbf{B}, m) = \frac{\left|\mathbf{B}^\top \mathbf{C}^{(B)} \mathbf{B}\right|}{\left|\left(\sum_{k=1}^{K} P_k (\mathbf{B}^\top \mathbf{C}_k \mathbf{B})^m \right)^{1/m}\right|}, \tag{5}$$

where m denotes a control parameter. We have referred to it as power linear discriminant analysis (PLDA). Intuitively, as m becomes larger, the classes with larger variances become dominant in the denominator of Eq. (5). Conversely, as m becomes smaller, the classes with smaller variances become dominant. Thus, by varying the control parameter m, the objective function can represent various objective functions. If m is set to one/zero, the objective function corresponds to the LDA/HDA objective function [14].

The following objective function is given as another definition of PLDA:

$$J_{PLDA_2}(\mathbf{B}, m) = \frac{\left|\mathbf{B}^\top \mathbf{C}^{(M)} \mathbf{B}\right|}{\left|\left(\sum_{k=1}^{K} P_k (\mathbf{B}^\top \mathbf{C}_k \mathbf{B})^m \right)^{1/m}\right|}, \tag{6}$$

If m is set to zero, the objective function corresponds to heteroscedastic linear discriminant analysis [15], which is shown in [14]. One issue regarding PLDA in practice is how to select the optimal control parameter m. In [16], the method for selecting a sub-optimal control parameter is provided.

3. Existing Dimensionality Reduction Preserving Locality of Data Structure

Recently, several linear dimensionality reduction methods for multimodal data have been proposed in the machine learning community [9)-12)]. Here, we review two methods: locality preserving projection (LPP) [10] and local Fisher discriminant analysis (LFDA) [11].

3.1 Locality Preserving Projection

Let **A** be a symmetric $N \times N$ matrix, which represents an affinity between features [10]. The (i,j)-element $A_{i,j}$ of **A** is the affinity between $\mathbf{x}_i$ and $\mathbf{x}_j$. An affinity element $A_{i,j}$ becomes a large value if $\mathbf{x}_i$ and $\mathbf{x}_j$ are located close to each other. Contrarily, $A_{i,j}$ becomes a small value if $\mathbf{x}_i$ and $\mathbf{x}_j$ are located far from each other. There are several different definitions of **A**, e.g., the nearest neighbor [17], the heat kernel [18] or the local scaling [19]. The objective function of LPP is defined as follows [10]:

$$J_{LPP}(\mathbf{B}) = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} ||\mathbf{B}^\top \mathbf{x}_i - \mathbf{B}^\top \mathbf{x}_j||^2, \quad \text{s.t.} \quad \mathbf{B}^\top \mathbf{X} \mathbf{D} \mathbf{X}^\top \mathbf{B} = \mathbf{I}, \tag{7}$$

where $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_N]$, is the identity matrix, and **D** is a diagonal matrix whose (i,i)-element is given by $D_{i,i} = \sum_{j=1}^{N} A_{ij}$. Minimizing Eq. (7) with respect to **B**, LPP seeks for a projection matrix **B** such that nearby data pairs in the original space remain close in the projected space. To ignore a trivial solution, i.e., **B** = **0**, LPP imposes the constraint (7). Thus, LPP is an unsupervised dimensionality reduction method preserving locality of features in the original space.

† We let a function *f* of a symmetric positive definite matrix **A** equal , $\mathbf{U}diag(f(\lambda_1),\ldots,f(\lambda n))\mathbf{U}^T = \mathbf{U}(f(\mathbf{A}))\mathbf{U}^T$, where $\mathbf{A} = \mathbf{U}\mathbf{A}\mathbf{U}^T$, **U** denotes the matrix of *n* eigenvectors, and **A** denotes the diagonal matrix of eigenvalues, λi. We may define the function *f* as some power **A**.

3.2 Local Fisher Discriminant Analysis

A supervised dimensionality reduction method preserving locality of features has been proposed by Sugiyama [11)20)] and has been referred to as local Fisher discriminant analysis (LFDA). LFDA combines the ideas of LDA (FDA) and LPP.

Within-class scatter and between-class scatter explained in Section 2.1 can be rewritten in a pairwise manner:

$$\mathbf{S}^{(W)} = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(W)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{8}$$

$$\mathbf{S}^{(B)} = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(B)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{9}$$

where

$$W_{ij}^{(W)} = \begin{cases} 1/N_1 & \text{if } y_i = y_j = 1, \\ \vdots & \vdots \\ 1/N_K & \text{if } y_i = y_j = K, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{10}$$

$$W_{ij}^{(B)} = \begin{cases} 1/N - 1/N_1 & \text{if } y_i = y_j = 1, \\ \vdots & \vdots \\ 1/N - 1/N_K & \text{if } y_i = y_j = K, \\ 1/N & \text{if } y_i \neq y_j. \end{cases} \tag{11}$$

LDA searches for a projection matrix **B** such that data pairs in the same class are close to each other and data pairs in different classes are separate from each other. A more formal interpretation of this is given in [11)]. Based on an affinity matrix **A** and the pairwise expressions of the between / within-class scatter, a *local* within-class scatter and a *local* between-class scatter are defined as follows [11)]:

$$\mathbf{S}^{(LW)} = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(LW)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{12}$$

$$\mathbf{S}^{(LB)} = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(LB)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{13}$$

$$W_{ij}^{(LW)} = \begin{cases} A_{ij}/N_1 & \text{if } y_i = y_j = 1, \\ \vdots & \vdots \\ A_{ij}/N_K & \text{if } y_i = y_j = K, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{14}$$

$$W_{ij}^{(LB)} = \begin{cases} A_{ij}(1/N - 1/N_1) & \text{if } y_i = y_j = 1, \\ \vdots & \vdots \\ A_{ij}(1/N - 1/N_K) & \text{if } y_i = y_j = K, \\ 1/N & \text{if } y_i \neq y_j. \end{cases} \tag{15}$$

Both $\mathbf{S}^{(LW)}$ and $\mathbf{S}^{(LB)}$ put a weight on data pairs in the same class, which is proportional to their affinity. The objective function of LFDA corresponding to Eq. (3) is defined as follows [11)20)]:

$$J_{LFDA_3}(\mathbf{B}) = tr\left(\left(\mathbf{B}^\top\mathbf{S}^{(LW)}\mathbf{B}\right)^{-1}\mathbf{B}^\top\mathbf{S}^{(LB)}\mathbf{B}\right). \tag{16}$$

LFDA searches for a projection matrix **B** such that nearby data pairs in the same class remain close and the data pairs in different classes are separate from each other; far-apart data pairs in the same class are not forced to be close. Thus, LFDA is a supervised dimensionality reduction method preserving locality. If A_{ij} is taken to be one for all in-class pairs, LFDA corresponds exactly to LDA because $\mathbf{S}^{(LW)}$ and $\mathbf{S}^{(LB)}$agree with $\mathbf{S}^{(W)}$ and $\mathbf{S}^{(B)}$, respectively. Thus, LFDA is an extension of LDA to deal with multimodal data.

In the same fashion as the definition of LDA objective functions, the following function could be defined as other objective functions of LFDA:

$$J_{LFDA_1}(\mathbf{B}) = \frac{\left|\mathbf{B}^\top\mathbf{S}^{(LB)}\mathbf{B}\right|}{\left|\mathbf{B}^\top\mathbf{S}^{(LW)}\mathbf{B}\right|}, \tag{17}$$

$$J_{LFDA_2}(\mathbf{B}) = \frac{\left|\mathbf{B}^\top\mathbf{S}^{(LM)}\mathbf{B}\right|}{\left|\mathbf{B}^\top\mathbf{S}^{(LW)}\mathbf{B}\right|}, \tag{18}$$

where a *local* mixture scatter $\mathbf{S}^{(LM)}$ is given by

$$\mathbf{S}^{(LM)} = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}^{(LM)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{19}$$

and $W_{ij}^{(LM)}$ is given by

$$W_{ij}^{(LM)} = W_{ij}^{(LW)} + W_{ij}^{(LB)} = \begin{cases} A_{ij}/N & \text{if } y_i = y_j, \\ 1/N & \text{if } y_i \neq y_j. \end{cases} \tag{20}$$

The optimizations of Eqs. (16) to (18) result in the same projection.

Local within-class covariance, $\mathbf{C}^{(LW)}$, local between-class covariance, $\mathbf{C}^{(LB)}$, and local mixture covariance, $\mathbf{C}^{(LM)}$, can be defined as $\mathbf{C}^{(LW)}= \frac{1}{N}\mathbf{S}^{(LW)}$,$\mathbf{C}^{(LB)}= \frac{1}{N}\mathbf{S}^{(LB)}$ and

$\mathbf{C}^{(LB)} = \frac{1}{N}\mathbf{S}^{(LM)}$, respectively. The same solution is obtained when $\mathbf{S}^{(LW)}$, $\mathbf{S}^{(LB)}$ and $\mathbf{S}^{(LM)}$ in Eqs. (16) to (18) are replaced with $\mathbf{C}^{(LW)}$, $\mathbf{C}^{(LB)}$ and $\mathbf{C}^{(LM)}$, respectively.

4. Extensions of HDA and PLDA to Deal with Multimodality

We first describe limitations facing the existing methods: LDA, HDA, PLDA and LFDA. Next, in order to ease the limitations, we propose two methods that extend HDA and PLDA.

4.1 Limitations of Existing Methods

While LDA is widely used to reduce dimensionality because of its simplicity and effectiveness, it assumes that each class shares common class covariance (i.e., homoscedasticity) [6]. Therefore, if this assumption is far from the real data, LDA sometimes does not work well. In order to overcome the limitation, HDA has been proposed, which can deal with unequal class covariances (i.e., heteroscedasticity). These two methods, however, sometimes does not work well because the fixed weight of each class covariance in the two methods cannot be necessarily suitable for any kind of data [14]. So we previously proposed PLDA to generalize LDA and HDA to control the class weights. Unfortunately, all these methods implicitly assume that data are generated from a single Gaussian distribution. Therefore, they cannot deal with multimodal data appropriately. To deal with multimodal data, LFDA has been proposed as explained in Section 3.2. It extends the between-class covariance and the within-class covariance to preserve locality of data structure. Nevertheless, since LFDA is based on LDA that assumes homoscedasticity, the effectiveness of LFDA may be limited.

In the following sections, we extend HDA that assumes heteroscedasticity using locality-preserving class covariances that can deal with multimodal data. We also propose locality-preserving PLDA. These extensions can be expected to yield better performance because they do not assume homoscedasticity and can reduce dimensionality of multimodal data appropriately.

4.2 Local Heteroscedastic Discriminant Analysis

To deal with multimodality using LDA, LFDA extends the within-class and between-class covariances in the LDA objective function to the *local* within-class and between-class covariances, respectively. The HDA objective function uses class covariances instead of a within-class covariance. Therefore, we will extend class covariances, similar to the *local* within-class and *local* between-class covariances. We first rearrange a class covariance matrix in a pairwise manner:

$$\mathbf{C}_k = \frac{1}{2N_k}\sum_{i,j=1}^{N} W_{k,ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top,$$

where

$$W_{k,ij} = \begin{cases} 1/N_k & \text{if } y_i = y_j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

$W_{ij}^{(W)}$ and $W_{k,ij}$ satisfy $W_{ij}^{(W)} = \sum_{k=1}^{K} W_{k,ij}$.

Similar to LDA, HDA also searches for a projection matrix $\mathbf{B}$ so that data pairs in the same class are close to each other and data pairs in different classes are separate from each other. A more formal interpretation is given in [22].

A class covariance matrix can extend to preserve locality of the data structure, similar to the extensions of $\mathbf{S}^{(W)}$ and $\mathbf{S}^{(B)}$. Let us define a *local* class covariance matrix $\mathbf{C}_k^{(L)}$ as follows:

$$\mathbf{C}_k^{(L)} = \frac{1}{2N_k}\sum_{i,j=1}^{N} W_{k,ij}^{(L)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{22}$$

where

$$W_{k,ij}^{(L)} = \begin{cases} A_{ij}/N_k & \text{if } y_i = y_j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

From Eqs. (14) and (23), $W_{ij}^{(LW)} = \sum_{k=1}^{K} W_{k,ij}^{(L)}$. In addition, $\mathbf{C}_k^{(L)}$ and $\mathbf{C}^{(LW)}$ satisfy $\mathbf{C}^{(LW)} = \sum_{k=1}^{K} P_k \mathbf{C}_k^{(L)}$. Replacing class and the between-class covariance matrices with local class and the local between-class ones, the objective function of HDA preserving locality is defined as follows:

$$J_{LHDA}(\mathbf{B}) = \frac{\left|\mathbf{B}^\top \mathbf{C}^{(LB)}\mathbf{B}\right|}{\prod_{k=1}^{K}\left|\mathbf{B}^\top \mathbf{C}_k^{(L)}\mathbf{B}\right|^{P_k}}. \tag{24}$$

We call it local HDA. If A_{ij} is taken to be one for all in-class pairs, LHDA is proportionate to HDA because $\mathbf{C}_k^{(L)}$ corresponds to $\mathbf{C}_k$. Since the only difference between Eqs. (4) and (24) is the definitions of their covariance matrices, the solution to maximize Eq. (24) with respect to **B** is obtained through the same numerical optimization procedure of HDA.

4.3 Local Power Linear Discriminant Analysis

As in the case of LHDA, using *local* class covariances $\mathbf{C}_k^{(L)}$, we extend a PLDA objective function as follows:

$$J_{LPLDA_1}(\mathbf{B}, m) = \frac{\left|\mathbf{B}^\top \mathbf{C}^{(LB)}\mathbf{B}\right|}{\left|\left(\sum_{k=1}^{K} P_k (\mathbf{B}^\top \mathbf{C}_k^{(L)} \mathbf{B})^m\right)^{1/m}\right|}. \tag{25}$$

We call it local PLDA (LPLDA). From Eqs. (17) and (24), LPLDA corresponds exactly to LFDA when m and LPLDA corresponds exactly to LHDA when $m \rightarrow 0$. Since the only difference between Eqs. (5) and (25) is the definitions of their covariance matrices, the solution to maximize Eq. (25) with respect to **B** is obtained through the same numerical optimization procedure of PLDA [8)14)]. We can also extend the other definition of PLDA as follows:

$$J_{LPLDA_2}(\mathbf{B}, m) = \frac{\left|\mathbf{B}^\top \mathbf{C}^{(LM)}\mathbf{B}\right|}{\left|\left(\sum_{k=1}^{K} P_k (\mathbf{B}^\top \mathbf{C}_k^{(L)} \mathbf{B})^m\right)^{1/m}\right|}.$$

LPLDA corresponds exactly to PLDA when A_{ij} is taken to be one for all in-class pairs.

5. Approximate Computations of Local Covariances

To obtain the optimal projections by LFDA, LHDA and LPLDA, $\mathbf{C}_k^{(L)}$, $\mathbf{C}^{(LW)}$, $\mathbf{C}^{(LM)}$ and $\mathbf{C}^{(LB)}$ must be calculated in advance. Throughout the paper, these covariance matrices are called *local* covariance matrices. Each local covariance matrix requires N^2 times calculations from their definitions. Therefore, their computational complexities are proportional to N^2. Since acoustic models in a speech recognition system are generally trained using a large amount of speech data, the value of N tends to become large, e.g., 10^6 to 10^9. Hence, the computational costs of local covariance matrices tend to be high.

5.1 Approximation of Local Class Covariances

For rapid calculation of local covariances, we first consider an approximate computation of local class covariances. In general, each class is represented as GMMs or HMMs in a speech recognizer. Therefore, we assume that the distribution of each class is constructed from several separate clusters. In addition, we approximate a local class covariance by the average of covariances of the clusters. The relation between a local class covariance and covariances of clusters is similar to that between the within-class covariance and class covariances. Then, we have

$$\mathbf{C}_k^{(L)} \approx \sum_{m=1}^{M_k} P_{k,m}\mathbf{C}_{k,m} \equiv \tilde{\mathbf{C}}_k^{(L)}, \tag{26}$$

where M_k is the number of clusters in class k, $P_{k,m}$ is the weight of the m-th cluster in class k, and $\mathbf{C}_{k,m}$ is an m-th cluster covariance in class k. $\tilde{\mathbf{C}}_k^{(L)}$ denotes an approximated local class covariance matrix. $\tilde{\mathbf{C}}_k^{(L)}$ agrees with $\mathbf{C}_k^{(L)}$ when the affinity matrix is defined as follows: $A_{ij}=1/P_k$, if x_i and x_j are assigned to the same cluster m in a class k, otherwise A_{ij}=0. If the number of clusters equals one, $\tilde{\mathbf{C}}_k^{(L)}$ corresponds to $\mathbf{C}_k$. To obtain $P_{k,m}$ and $\mathbf{C}_{k,m}$, we employ the Expectation-Maximization (EM) algorithm. Since the computational complexities of the E-step and the M-step in the EM algorithm are proportional to the number of data, we can rapidly calculate $\mathbf{C}_k^{(L)}$ by using Eq. (26).

5.2 Approximation of Other Local Covariances

$\mathbf{C}^{(LW)}$, $\mathbf{C}^{(LM)}$, and $\mathbf{C}^{(LB)}$ can be rewritten using $\mathbf{C}_k^{(L)}$ as follows:

$$\mathbf{C}^{(LW)} = \sum_{k=1}^{K} P_k \mathbf{C}_k^{(L)}, \tag{27}$$

$$\mathbf{C}^{(LM)} = \mathbf{C}^{(M)} - \sum_{k=1}^{K} P_k^2 (\mathbf{C}_k - \mathbf{C}_k^{(L)}), \tag{28}$$

$$\mathbf{C}^{(LB)} = \mathbf{C}^{(LM)} - \mathbf{C}^{(LW)}. \tag{29}$$

The derivation of Eq. (28) is given in [22)], Since the computational cost of $\mathbf{C}_k^{(L)}$ is proportional to N^2, these covariances involve considerable computational costs.

To calculate these covariances rapidly, we replace all

$\mathbf{C}_k^{(L)}$ in Eqs. (27) - (29) by $\tilde{\mathbf{C}}_k^{(L)}$:

$$\mathbf{C}^{(LW)} \approx \sum_{k=1}^{K} P_k \tilde{\mathbf{C}}_k^{(L)} \equiv \tilde{\mathbf{C}}^{(LW)}, \tag{30}$$

$$\mathbf{C}^{(LM)} \approx \mathbf{C}^{(M)} - \sum_{k=1}^{K} P_k^2 (\mathbf{C}_k - \tilde{\mathbf{C}}_k^{(L)}) \equiv \tilde{\mathbf{C}}^{(LM)}, \tag{31}$$

$$\mathbf{C}^{(LB)} \approx \tilde{\mathbf{C}}^{(LM)} - \tilde{\mathbf{C}}^{(LW)} \equiv \tilde{\mathbf{C}}^{(LB)}. \tag{32}$$

$\tilde{\mathbf{C}}^{(LW)}$, $\tilde{\mathbf{C}}^{(LM)}$ and $\tilde{\mathbf{C}}^{(LB)}$ denote approximated $\mathbf{C}^{(LW)}$, $\mathbf{C}^{(LM)}$and $\mathbf{C}^{(LB)}$, respectively. Since the computational costs of $\mathbf{C}^{(M)}$ and $\mathbf{C}_k$ are proportional to the number of data, there are no N^2 times calculations in Eqs. (30) - (32). Once we calculate $\mathbf{C}^{(M)}$ and $\mathbf{C}_k$, and estimate $P_{k,m}$ and $\mathbf{C}_{k,m}$ for $\tilde{\mathbf{C}}_k^{(L)}$ using the EM algorithm, we can calculate $\tilde{\mathbf{C}}^{(LW)}$, $\tilde{\mathbf{C}}^{(LB)}$ and $\tilde{\mathbf{C}}^{(LM)}$ immediately. Thus, the computational costs are significantly reduced.

6. Experiments

We conducted experiments on CENSREC-3 database [21], which is designed as an evaluation framework for Japanese isolated word recognition in real in-car environments. Speech data were collected using two microphones: a close-talking (CT) microphone and a hands-free (HF) microphone. We only used the speech data collected using a CT microphone. For training of HMMs, a driver's speech of phonetically-balanced sentences was recorded under two conditions: while idling and driving on city streets under a normal in-car environment without air-conditioner noise. A total of 14,050 utterances by 293 drivers (202 males and 91 females) were recorded with a CT microphone. For evaluation, we used driver's speech of isolated words recorded with a CT microphone under three different conditions: an in-car environment without A / C noise (*normal*), with low fan-speed noise (*fan low*), and with high fan-speed noise (*fan high*). Originally, the aim of feature transformation is to reduce redundant information and not to treat mismatched conditions explicitly. However, the transformations should not compromise the system's robustness and so we also investigate robustness under different noise conditions. Although one can use various noise conditions, to make the problem simple, we selected fan noise for the investigation. There are 2,646, 2,637 and 2,695 speech utterances for *normal, fan low* and *fan high* conditions, respectively. The speech signals for training and evaluation were both sampled at 16 kHz.

6.1 Experimental setup

For an evaluation procedure, we followed the CENSREC-3 baseline scripts except that fifty similar-sounding words were added to the vocabulary (total 100 words) to make the recognition task difficult. The acoustic models consist of triphone HMMs. Each HMM has five states, and three of them have output distributions. Each distribution is represented with 32 mixture diagonal Gaussians. The total number of states with the distributions is 2,000. The baseline performance was calculated with 39 dimensional feature vectors that consist of 12 MFCCs and log-energy with their corresponding delta and acceleration coefficients. Eleven successive frames, whose center is the current frame, were used to obtain dynamic coefficients because delta and acceleration window sizes were three and two, respectively. At the beginning and end of the speech, the first or last vector is replicated five-fold. Frame length is 20 ms and frame shift is 10 ms. In the Mel-filter bank analysis, a cut-off is applied to frequency components lower than 250 Hz. Throughout the experiments, cepstral mean normalization is not applied to the features [‡].

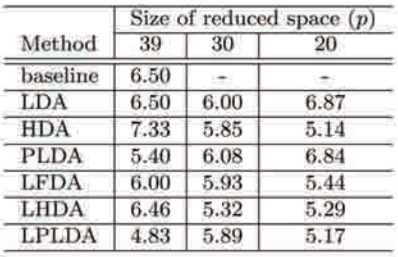
6.2 Feature Transformation Procedure

Feature transformation was performed using LDA, HDA [7], PLDA [8], LFDA [11], LHDA and LPLDA for spliced features. Eleven successive frames (143 dimensions), whose center is the current frame, were reduced to 20, 30 and 39 to investigate the effectiveness of the feature transformation methods. At the beginning and end of the speech, the first or last vector is replicated five-fold. In PLDA and LPLDA, we used the limited-memory BFGS algorithm as a numerical optimization technique, and their control parameters were experimentally selected. The LDA transformation matrix was used as the initial gradient. In LFDA, LHDA and LPLDA, the number of mixtures was four for each class, while the number of mixtures was one for the classes that

[‡] In CENSREC-3, there is no difference in the recording conditions between the training data and the evaluation data from the standpoint of convolutional noises such as reverberation. Therefore, the effectiveness of cepstral mean normalization is limited. In practice, preliminary experimental results have showed that cepstral mean normalization did not improve recognition performance in almost all the cases.

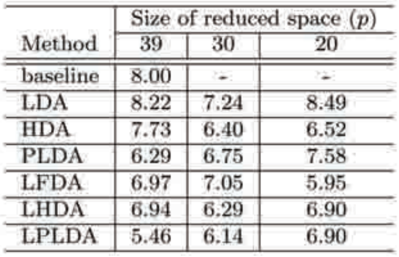
have training data of less than one percent of the total. In addition, to obtain projection matrices by LFDA, LHDA and LPLDA, we employed an approximate computation scheme for calculating covariances. To assign one of the classes to every feature vector, HMM state labels were generated for the training data by a state-level forced alignment algorithm using a well-trained HMM system. The number of classes was 40.

Table 1　Word error rates(%)under a *normal* condition.



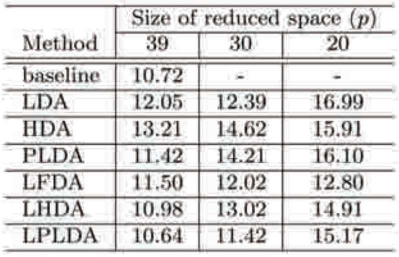
Method	Size of reduced space (p)		
	39	30	20
baseline	6.50	-	-
LDA	6.50	6.00	6.87
HDA	7.33	5.85	5.14
PLDA	5.40	6.08	6.84
LFDA	6.00	5.93	5.44
LHDA	6.46	5.32	5.29
LPLDA	4.83	5.89	5.17

Table 2　Word error rates(%)under a *fan low* condition.



Method	Size of reduced space (p)		
	39	30	20
baseline	8.00	-	-
LDA	8.22	7.24	8.49
HDA	7.73	6.40	6.52
PLDA	6.29	6.75	7.58
LFDA	6.97	7.05	5.95
LHDA	6.94	6.29	6.90
LPLDA	5.46	6.14	6.90

Table 3　Word error rates(%)under a *fan hign* condition.



Method	Size of reduced space (p)		
	39	30	20
baseline	10.72	-	-
LDA	12.05	12.39	16.99
HDA	13.21	14.62	15.91
PLDA	11.42	14.21	16.10
LFDA	11.50	12.02	12.80
LHDA	10.98	13.02	14.91
LPLDA	10.64	11.42	15.17

6.3　Results

Experimental results are presented in **Table 1** to **Table 3**. The noise condition for the evaluation data used in **Table 1** matches that for training data. The evaluation data used in **Table 2** and the data used in **Table 3** contain low air-conditioner noise and high air-conditioner noise, respectively. These noises are not contained in training data.

We first discuss the results of the feature transformation methods when the size of a reduced space is 39 (i.e., $p=39$). The size is equal to that of baseline. **Table 1** showed that the locality-preserving dimensionality reduction methods consistently yielded better performance than the traditional methods. This result suggests that projected features using the locality-preserving methods have higher separability among acoustic classes than those using the traditional methods because the locality-preserving methods can consider multimodality of data. Especially, LPLDA yielded the lowest word error rate (WER) among all dimensionality reduction methods. **Table 2** showed a similar tendency to **Table 1**. The locality-preserving dimensionality reduction methods also yielded better performance. These results were obtained from the fact that the difference between a *normal* condition and a *fan low* condition is slight because A/C noise with a low fan-speed is small. In addition, the combinations of heteroscedasticity and locality-preservation worked well. On the other hand, **Table 3** showed a different tendency from the others. The feature transformation methods excluding LPLDA gave worse performance than at baseline (MFCC+Δ+Δ Δ). In general, the degree of confusability of acoustic features among different classes would change when the noise in training differs considerably from that in evaluation. Therefore, a feature transformation estimated under a *normal* noise environment in training did not necessarily work well under a *fan high* noise environment in evaluation. Nevertheless, LPLDA kept comparable performance with the baseline whether or not the noise condition in evaluation matches when training because it would transform features that have sufficiently high separability among different classes even in a mismatch noise condition.

Next, we discuss the results of the feature transformation methods when $p=20$ and $p=30^{\S}$. As shown in **Table 1** and **Table 2**, under matched and almost matched noise conditions between training and evaluation, the optimal dimensions of most feature transformation methods are lower than 39. On the other hand, **Table 3** showed that all methods degraded recognition performance under a mismatched noise condition when the dimensions were relatively small. These results imply that feature transformation methods might obtain lower dimensions in matched conditions, whereas in mismatched conditions, redundant information can contribute to the improvement of recognition performance. **Table 1** to **Table 3** also showed that while the proposed methods did not necessarily yield comparable performance of the other methods when $p = 20$, they consistently yielded the lowest word error rate when $p \geq 30$.

7. Conclusions

In this paper, we proposed two-dimensionality reduction methods; HDA preserving the local structure of the data (LHDA) and PLDA preserving the local structure (LPLDA), to reduce dimensionality of multimodal data appropriately. In general, to obtain the optimal projections by the locality-preserving methods, considerable computational time is required. In order to overcome this problem, we proposed an approximate calculation scheme. Experimental results showed that the locality-preserving dimensionality reduction methods yielded better performance than traditional ones, especially under matched noise conditions. In particular, LPLDA outperformed the others whether or not the noise condition in evaluation matched that in training.

8. Acknowledgment

The study presented was conducted using the CENSREC-3 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

References

1) S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech and Signal Processing, vol.34, no.1, 1986.

2) S. Nakagawa et al., "Evaluation of segmental unit input HMM, "Proc. ICASSP, 1996.

3) R. Haeb-Umbach et al., "Linear discriminant analysis for improved large vocabulary continuous speech recognition, "Proc. ICASSP, 1992.

4) K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, 1990.

5) R. O. Duda et al., Pattern Classification, John Wiley & Sons, New York, 2001.

6) N. A. Campbell, "Canonical variate analysis - a general model formulation, "Australian Journal of Statistics, vol.4, 1984.

7) G. Saon et al., "Maximum likelihood discriminant feature spaces," Proc. ICASSP, 2000.

8) M. Sakai et al., "Generalization of linear discriminant analysis used in segmental unit input HMM for speech recognition," Proc. ICASSP, 2007.

9) T. Hastie et al., "Discriminant analysis by Gaussian mixtures," Journal of the Royal Statistical Society, vol.58, no.1, 1996.

10) X. He et al., "Locality preserving projections, "Advances in Neural Information Processing Systems, 2004.

11) M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, "Journal of Machine Learning Research, vol.8, 2007.

12) F. de la Torre et al., "Multimodal oriented discriminant analysis," International Conference on Machine Learning, 2005.

13) J. Nocedal et al., Numerical Optimization, Springer-Verlag, 1999.

14) M. Sakai et al., "Linear discriminant analysis using a generalized mean of class covariances and its application to speech recognition, "IEICE Transactions on Information and Systems, vol.E91-D, no.3, 2008.

15) N. Kumar et al., "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," Speech Communication, 1998.

16) M. Sakai et al., "Selection of optimal dimensionality reduction method using Chernoff bound for segmental unit input HMM, "Proc. Interspeech, 2007.

17) S. T. Roweis et al., "Nonlinear dimensionality reduction by locally linear embedding, "Science, vol.290, no.5500, 2000.

18) M. Belkin et al., "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Computation, vol.15, no.6, 2003.

19) L. Zelnik-Manor et al., "Self-tuning spectral clustering," Advances in Neural Information Processing Systems, 2005.

20) M. Sugiyama, "Local Fisher discriminant analysis for supervised dimensionality reduction, "International Conference on Machine Learning, 2006.

21) M. Fujimoto et al., "CENSREC-3: An evaluation framework for Japanese speech recognition in real driving-car environments, "IEICE Transactions on Information and Systems, vol.E89-D, no.11, 2006.

§ In some preliminary experiments, the degradation of recognition performance was found when p>39 and p<20 with a few exceptions. While the best performances of PLDA and LPLDA were found at p=50 under a fan high condition, the differences of the performances between p=50 and p=39 were not so large. Although the best results using different methods under the different environments are obtained with a few variety of p as explained here, these facts does not affect the overall conclusion of this paper.

22) M. Sakai et al., "Acoustic Feature Transformation Based on Discriminant Analysis Preserving Local Structure for Speech Recognition, "IEICE Transactions on Information and Systems, vol.E93-D, no.5, 2010.

＜著　者＞

坂井　誠
（さかい　まこと）
エレクトロニクス研究部
博士（情報科学）
ヒューマンマシンインター
フェースの研究に従事

北岡　教英
（きたおか　のりひで）
名古屋大学大学院情報科学
研究科 准教授 博士（工学）
音声認識，音声対話システム，
マルチモーダルインタフェース
の研究に従事

武田　一哉
（たけだ　かずや）
名古屋大学大学院情報科学
研究科 教授 博士（工学）
音声符号化，空間音響処理，
音声情報処理など音声音響
言語処理の研究に従事