IEICE TRANS. INF. & SYST., VOL.E91–D, NO.3 MARCH 2008

PAPER Special Section on Robust Speech Processing in Realistic Environments

Linear Discriminant Analysis Using a Generalized Mean of Class Covariances and Its Application to Speech Recognition

Makoto SAKAI[†,††a)], **Norihide KITAOKA**[††b)], ***Members***, ***and*** **Seiichi NAKAGAWA**[†††c)], ***Fellow***

SUMMARY To precisely model the time dependency of features is one of the important issues for speech recognition. Segmental unit input HMM with a dimensionality reduction method has been widely used to address this issue. Linear discriminant analysis (LDA) and heteroscedastic extensions, e.g., heteroscedastic linear discriminant analysis (HLDA) or heteroscedastic discriminant analysis (HDA), are popular approaches to reduce dimensionality. However, it is difficult to find one particular criterion suitable for any kind of data set in carrying out dimensionality reduction while preserving discriminative information. In this paper, we propose a new framework which we call power linear discriminant analysis (PLDA). PLDA can be used to describe various criteria including LDA, HLDA, and HDA with one control parameter. In addition, we provide an efficient selection method using a control parameter without training HMMs nor testing recognition performance on a development data set. Experimental results show that the PLDA is more effective than conventional methods for various data sets.

key words: *speech recognition, feature extraction, multidimensional signal processing*

1. Introduction

Although Hidden Markov Models (HMMs) have been widely used to model speech signals for speech recognition, they cannot precisely model the time dependency of feature parameters. In order to overcome this limitation, many extensions have been proposed [1]–[5]. Segmental unit input HMM [1] has been widely used for its effectiveness and tractability. In segmental unit input HMM, the immediate use of several successive frames as an input vector inevitably increases the number of parameters. Therefore, a dimensionality reduction method is applied to feature vectors.

Linear discriminant analysis (LDA) [6], [7] is widely used to reduce dimensionality and a powerful tool to preserve discriminative information. LDA assumes each class has the same class covariance [8]. However, this assumption does not necessarily hold for a real data set. In order to overcome this limitation, several methods have been proposed. Heteroscedastic linear discriminant analysis (HLDA) could deal with unequal covariances because the maximum likelihood estimation was used to estimate parameters for different Gaussians with unequal covariances [9]. Heteroscedastic discriminant analysis (HDA) was proposed as another objective function which employed individual weighted contributions of the classes [10]. The effectiveness of these methods for some data sets has been experimentally demonstrated. However, it is difficult to find one particular criterion suitable for any kind of data set.

In this paper we show that these three methods have a strong mutual relationship, and provide a new interpretation for them. Then, we propose a new framework that we call *power linear discriminant analysis* (PLDA), which can describe various criteria including LDA, HLDA, and HDA with one control parameter. Because PLDA can describe various criteria for dimensionality reduction, it can flexibly adapt to various environments such as a noisy environment. Thus, PLDA can provide robustness to a speech recognizer in realistic environments. Unfortunately, we cannot know which control parameter is the most effective before training HMMs and testing the performances of each control parameter on a development set. In general, this training and testing process requires more than several dozen hours. Moreover, the computational time is proportional to the number of variations of the control parameters under test. PLDA requires much time to find an optimal control parameter because its control parameter can be set to a real number. We will provide an efficient selection method of an optimal control parameter without training of HMMs nor testing recognition performance on a development set.

The paper is organized as follows: LDA, HLDA, and HDA are reviewed in Sect. 2. Then, a new framework of PLDA is proposed in Sect. 3. A selection method of an optimal control parameter is derived in Sect. 4. Experimental results are presented in Sect. 5. Finally, conclusions are given in Sect. 6.

2. Segmental Unit Input HMM

For an input symbol sequence $\mathbf{o} = (\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T)$ and a state sequence $\mathbf{q} = (q_1, q_2, \cdots, q_T)$, the output probability of segmental unit input HMM is given by the following equations [1]:

$$P(\mathbf{o}_1, \cdots, \mathbf{o}_T) = \sum_{\mathbf{q}} \prod_i P(\mathbf{o}_i \mid \mathbf{o}_1, \cdots, \mathbf{o}_{i-1}, q_1, \cdots, q_i)$$

Manuscript received July 2, 2007.
Manuscript revised September 14, 2007.

[†]The author is with DENSO CORPORATION, Nisshin-shi, 470–0111 Japan.
[††]The authors are with Nagoya University, Nagoya-shi, 464–8603 Japan.
[†††]The author is with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
a) E-mail: msakai@rlab.denso.co.jp
b) E-mail: kitaoka@nagoya-u.jp
c) E-mail: nakagawa@slp.ics.tut.ac.jp
DOI: 10.1093/ietisy/e91–d.3.478

$$\times P(q_i \mid q_1, \cdots, q_{i-1}) \tag{1}$$

$$\approx \sum_{\mathbf{q}} \prod_i P(\mathbf{o}_i \mid \mathbf{o}_{i-(d-1)}, \cdots, \mathbf{o}_{i-1}, q_i) P(q_i \mid q_{i-1}) \tag{2}$$

$$\approx \sum_{\mathbf{q}} \prod_i P(\mathbf{o}_{i-(d-1)}, \cdots, \mathbf{o}_i \mid q_i) P(q_i \mid q_{i-1}), \tag{3}$$

where T denotes the length of input sequence and d denotes the number of successive frames used in probability calculation at a time frame. The immediate use of several successive frames as an input vector inevitably increases the number of parameters. When the number of dimensions increases, several problems generally occur: heavier computational load and larger memory are required, and the accuracy of a parameter estimation degrades. Then, dimensionality reduction methods, e.g., principal component analysis (PCA), LDA, HLDA or HDA, are used to reduce dimensionality [1], [3], [9], [10].

Here, we will briefly review LDA, HLDA and HDA, then investigate the effectiveness of these methods for some artificial data sets.

2.1 Linear Discriminant Analysis

Given n-dimensional feature vectors $\mathbf{x}_j \in \mathbb{R}^n$ $(j = 1, 2, \ldots, N)$ [†], e.g., $\mathbf{x}_j = \left[\mathbf{o}_{j-(d-1)}^T, \cdots, \mathbf{o}_j^T\right]^T$, let us find a projection matrix $\mathbf{B}_{[p]} \in \mathbb{R}^{n \times p}$ that projects these feature vectors to p-dimensional feature vectors $\mathbf{z}_j \in \mathbb{R}^p$ $(j = 1, 2, \ldots, N)$ $(p < n)$, where $\mathbf{z}_j = \mathbf{B}_{[p]}^T \mathbf{x}_j$, and N denotes the number of all features.

Within-class and between-class covariance matrices are defined as follows [6], [7]:

$$\begin{aligned}\boldsymbol{\Sigma}_w &= \frac{1}{N} \sum_{k=1}^{c} \sum_{\mathbf{x}_j \in D_k} \left(\mathbf{x}_j - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_j - \boldsymbol{\mu}_k\right)^T \\ &= \sum_{k=1}^{c} P_k \boldsymbol{\Sigma}_k, \end{aligned} \tag{4}$$

$$\boldsymbol{\Sigma}_b = \sum_{k=1}^{c} P_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T, \tag{5}$$

where c denotes the number of classes, D_k denotes the subset of feature vectors labeled as class k, $\boldsymbol{\mu}$ is the mean vector for all the classes, $\boldsymbol{\mu}_k$ is the mean vector of the class k, $\boldsymbol{\Sigma}_k$ is the covariance matrix of the class k, and P_k is the class weight, respectively.

There are several ways to formulate objective functions for multi-class data [6]. Typical objective functions are the following:

$$J_{LDA}\left(\mathbf{B}_{[p]}\right) = \frac{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_b \mathbf{B}_{[p]}\right|}{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_w \mathbf{B}_{[p]}\right|}, \tag{6}$$

$$J_{LDA}\left(\mathbf{B}_{[p]}\right) = \frac{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_t \mathbf{B}_{[p]}\right|}{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_w \mathbf{B}_{[p]}\right|}, \tag{7}$$

where $\boldsymbol{\Sigma}_t$ denotes the covariance matrix of all features, namely a total covariance, which equals $\boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_w$.

LDA finds a projection matrix $\mathbf{B}_{[p]}$ that maximizes Eqs. (6) or (7). The optimization of (6) and (7) results in the same projection.

2.2 Heteroscedastic Extensions

LDA is not the optimal projection when the class distributions are heteroscedastic. Campbell [8] has shown that LDA is related to the maximum likelihood estimation of parameters for a Gaussian model with an identical class covariance. However, this condition is not necessarily satisfied for a real data set.

In order to overcome this limitation, several extensions have been proposed [9]–[12]. This paper focuses on two heteroscedastic extensions called heteroscedastic linear discriminant analysis (HLDA) and heteroscedastic discriminant analysis (HDA) [9], [10].

2.2.1 Heteroscedastic Linear Discriminant Analysis

In HLDA, the full-rank linear projection matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ is constrained as follows: the first p columns of $\mathbf{B}$ span the p-dimensional subspace in which the class means and variances are different and the remaining $n - p$ columns of $\mathbf{B}$ span the $(n - p)$-dimensional subspace in which the class means and variances are identical. Let the parameters that describe the class means and covariances of $\mathbf{B}^T\mathbf{x}$ be $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$, respectively:

$$\hat{\boldsymbol{\mu}}_k = \begin{bmatrix} \mathbf{B}_{[p]}^T \boldsymbol{\mu}_k \\ \mathbf{B}_{[n-p]}^T \boldsymbol{\mu} \end{bmatrix}, \tag{8}$$

$$\hat{\boldsymbol{\Sigma}}_k = \begin{bmatrix} \mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_k \mathbf{B}_{[p]} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{[n-p]}^T \boldsymbol{\Sigma}_t \mathbf{B}_{[n-p]} \end{bmatrix}, \tag{9}$$

where $\mathbf{B} = \left[\mathbf{B}_{[p]} | \mathbf{B}_{[n-p]}\right]$ and $\mathbf{B}_{[n-p]} \in \mathbb{R}^{n \times (n-p)}$.

Kumar et al. [9] incorporated the maximum likelihood estimation of parameters for differently distributed Gaussians. An HLDA objective function is derived as follows:

$$J_{HLDA}(\mathbf{B}) = \frac{|\mathbf{B}|^{2N}}{\left|\mathbf{B}_{[n-p]}^T \boldsymbol{\Sigma}_t \mathbf{B}_{[n-p]}\right|^N} \frac{1}{\prod_{k=1}^{c} \left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_k \mathbf{B}_{[p]}\right|^{N_k}}. \tag{10}$$

N_k denotes the number of features of class k. The solution to maximize Eq. (10) is not analytically obtained. Therefore, its maximization is performed using a numerical optimization technique. Alternatively, a computationally efficient scheme is given in [13].

[†]This paper uses the following notation: capital bold letters refer to matrices, e.g., $\mathbf{A}$, bold letters refer to vectors, e.g., $\mathbf{b}$, and scalars are not bold, e.g., c. Where submatrices are used they are indicated, for example, by $\mathbf{A}_{[p]}$, this is an $n \times p$ matrix. $\mathbf{A}^T$ is the transpose of the matrix, $|\mathbf{A}|$ is the determinant of the matrix, and $tr(\mathbf{A})$ is the trace of the matrix.

2.2.2 Heteroscedastic Discriminant Analysis

HDA uses the following objective function which incorporates individual weighted contributions of the class variances [10]:

$$J_{HDA}\left(\mathbf{B}_{[p]}\right) = \prod_{k=1}^{c}\left(\frac{\left|\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_b\mathbf{B}_{[p]}\right|}{\left|\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_k\mathbf{B}_{[p]}\right|}\right)^{N_k} \tag{11}$$

$$= \frac{\left|\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_b\mathbf{B}_{[p]}\right|^N}{\prod_{k=1}^{c}\left|\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_k\mathbf{B}_{[p]}\right|^{N_k}}. \tag{12}$$

In contrast to HLDA, this function is not considered $(n - p)$ dimensions. Only a projection matrix $\mathbf{B}_{[p]}$ is estimated. There is no closed-form solution to obtain projection matrix $\mathbf{B}_{[p]}$ similar to HLDA.

2.3 Dependency on Data Set

In Fig. 1, two-dimensional, two- or three-class data features are projected onto one-dimensional subspaces by LDA and HDA. Here, HLDA projections were omitted because they were close to HDA projections. Figure 1 (a) shows that HDA has higher separability than LDA for the data set used in [10]. On the other hand, as shown in Fig. 1 (b), LDA has higher separability than HDA for another data set. Figure 1 (c) shows the case with another data set where both LDA and HDA have low separabilities. Thus, LDA and HDA do not always classify the given data set appropriately. All results show that the separabilities of LDA and HDA depend significantly on data sets.

3. Generalization of Discriminant Analysis

As shown above, it is difficult to separate appropriately every data set with one particular criterion such as LDA, HLDA, or HDA. Here, we concentrate on providing a framework which integrates various criteria.

3.1 Relationship between HLDA and HDA

By using Eqs. (8) and (9), let us rearrange $\mathbf{B}^T\boldsymbol{\Sigma}_t\mathbf{B}$ as follows:

$$\mathbf{B}^T\boldsymbol{\Sigma}_t\mathbf{B} = \mathbf{B}^T\boldsymbol{\Sigma}_b\mathbf{B} + \mathbf{B}^T\boldsymbol{\Sigma}_w\mathbf{B} \tag{13}$$

$$= \sum_k P_k(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^T + \sum_k P_k\hat{\boldsymbol{\Sigma}}_k \tag{14}$$

$$= \begin{bmatrix} \mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_t\mathbf{B}_{[p]} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{[n-p]}^T\boldsymbol{\Sigma}_t\mathbf{B}_{[n-p]} \end{bmatrix}, \tag{15}$$

where $\hat{\boldsymbol{\mu}} = \mathbf{B}^T\boldsymbol{\mu}$.

The determinant of this is

$$|\mathbf{B}^T\boldsymbol{\Sigma}_t\mathbf{B}| = |\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_t\mathbf{B}_{[p]}||\mathbf{B}_{[n-p]}^T\boldsymbol{\Sigma}_t\mathbf{B}_{[n-p]}|. \tag{16}$$

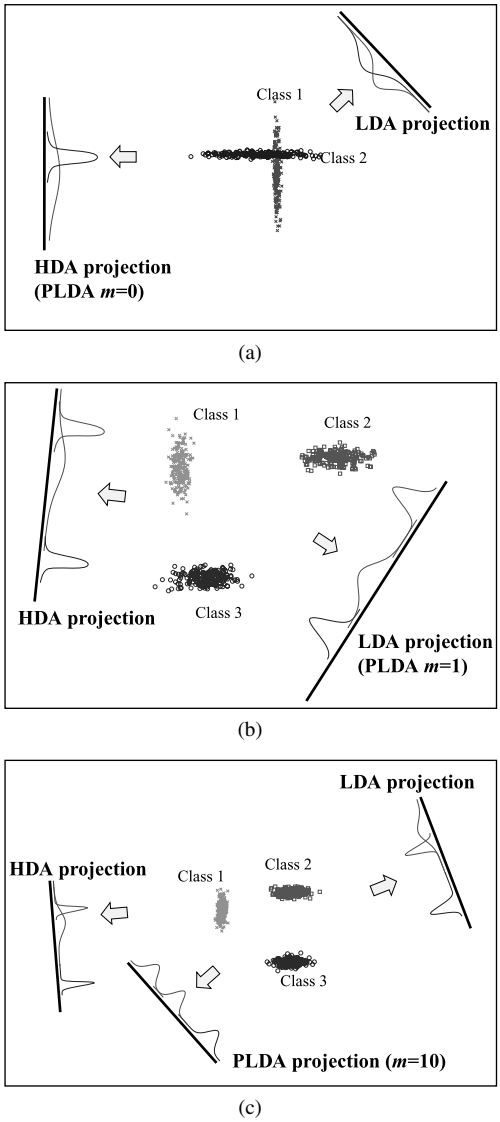


Fig. 1 Examples of dimensionality reduction by LDA, HDA and PLDA.

Inserting this in (10) and removing a constant term yields

$$J_{HLDA}\left(\mathbf{B}_{[p]}\right) \propto \frac{\left|\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_t\mathbf{B}_{[p]}\right|^N}{\prod_{k=1}^{c}\left|\mathbf{B}_{[p]}^T\boldsymbol{\Sigma}_k\mathbf{B}_{[p]}\right|^{N_k}}. \tag{17}$$

From (12) and (17), the difference between HLDA and HDA lies in their numerators, i.e., the total covariance matrix versus the between-class covariance matrix. This difference is the same as the difference between the two LDAs shown in (6) and (7). Thus, (12) and (17) can be viewed as the same formulation.

3.2 Relationship between LDA and HDA

The LDA and HDA objective functions can be rewritten as

$$J_{LDA}\left(\mathbf{B}_{[p]}\right) = \frac{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_b \mathbf{B}_{[p]}\right|}{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_w \mathbf{B}_{[p]}\right|} = \frac{\left|\tilde{\boldsymbol{\Sigma}}_b\right|}{\left|\sum_{k=1}^{c} P_k \tilde{\boldsymbol{\Sigma}}_k\right|}, \tag{18}$$

$$J_{HDA}\left(\mathbf{B}_{[p]}\right) = \frac{\left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_b \mathbf{B}_{[p]}\right|^N}{\prod_{k=1}^{c} \left|\mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_k \mathbf{B}_{[p]}\right|^{N_k}} \propto \frac{\left|\tilde{\boldsymbol{\Sigma}}_b\right|}{\prod_{k=1}^{c} \left|\tilde{\boldsymbol{\Sigma}}_k\right|^{P_k}}, \tag{19}$$

where $\tilde{\boldsymbol{\Sigma}}_b = \mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_b \mathbf{B}_{[p]}$ and $\tilde{\boldsymbol{\Sigma}}_k = \mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_k \mathbf{B}_{[p]}$ are between-class and class k covariance matrices in the projected p-dimensional space, respectively.

Both numerators denote determinants of the between-class covariance matrix. In Eq. (18), the denominator can be viewed as a determinant of *the weighted arithmetic mean* of the class covariance matrices. Similarly, in Eq. (19), the denominator can be viewed as a determinant of *the weighted geometric mean* of the class covariance matrices. Thus, the difference between LDA and HDA is the definitions of the mean of the class covariance matrices. Moreover, to replace their numerators with the determinants of the total covariance matrices, the difference between LDA and HLDA is the same as the difference between LDA and HDA.

3.3 Power Linear Discriminant Analysis

As described above, Eqs. (18) and (19) give us a new integrated interpretation of LDA and HDA. As an extension of this interpretation, their denominators can be replaced by a determinant of *the weighted harmonic mean*, or a determinant of *the root mean square*.

In the econometric literature, a more general definition of a mean is often used, called *the weighted mean of order* m [14]. We extend this notion to a determinant of a matrix mean and propose a new objective function as follows[†]:

$$J_{PLDA}\left(\mathbf{B}_{[p]}, m\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\left|\left(\sum_{k=1}^{c} P_k \tilde{\boldsymbol{\Sigma}}_k^m\right)^{1/m}\right|}, \tag{20}$$

where $\tilde{\boldsymbol{\Sigma}}_n \in \{\tilde{\boldsymbol{\Sigma}}_b, \tilde{\boldsymbol{\Sigma}}_t\}$, $\tilde{\boldsymbol{\Sigma}}_t = \mathbf{B}_{[p]}^T \boldsymbol{\Sigma}_t \mathbf{B}_{[p]}$, and m is a control parameter. By varying the control parameter m, the proposed objective function can represent various criteria. Some typical objective functions are enumerated below.

- $m = 2$ (root mean square)

$$J_{PLDA}\left(\mathbf{B}_{[p]}, 2\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\left|\left(\sum_{k=1}^{c} P_k \tilde{\boldsymbol{\Sigma}}_k^2\right)^{1/2}\right|}. \tag{21}$$

- $m = 1$ (arithmetic mean)

$$J_{PLDA}\left(\mathbf{B}_{[p]}, 1\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\left|\sum_{k=1}^{c} P_k \tilde{\boldsymbol{\Sigma}}_k\right|} = J_{LDA}\left(\mathbf{B}_{[p]}\right). \tag{22}$$

- $m \to 0$ (geometric mean)

$$J_{PLDA}\left(\mathbf{B}_{[p]}, 0\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\prod_{k=1}^{c} \left|\tilde{\boldsymbol{\Sigma}}_k\right|^{P_k}} \propto J_{HDA}\left(\mathbf{B}_{[p]}\right). \tag{23}$$

- $m = -1$ (harmonic mean)

$$J_{PLDA}\left(\mathbf{B}_{[p]}, -1\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\left|\left(\sum_{k=1}^{c} P_k \tilde{\boldsymbol{\Sigma}}_k^{-1}\right)^{-1}\right|}. \tag{24}$$

See Appendix A for the proof that the PLDA objective function tends to HDA objective function when m tends to zero. The following equations are also obtained under a particular condition (see Appendix B).

- $m \to \infty$

$$J_{PLDA}\left(\mathbf{B}_{[p]}, \infty\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\max_k \left|\tilde{\boldsymbol{\Sigma}}_k\right|}. \tag{25}$$

- $m \to -\infty$

$$J_{PLDA}\left(\mathbf{B}_{[p]}, -\infty\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\min_k \left|\tilde{\boldsymbol{\Sigma}}_k\right|}. \tag{26}$$

Intuitively, as m becomes larger, the classes with larger variances become dominant in the denominator of Eq. (20). Conversely, as m becomes smaller, the classes with smaller variances become dominant.

We call this new discriminant analysis formulation *Power Linear Discriminant Analysis* (PLDA). Figure 1 (c) shows that PLDA can have a higher separability for a data set with which LDA and HDA have lower separability. To maximize the PLDA objective function with respect to $\mathbf{B}$, we can use numerical optimization techniques such as the Nelder-Mead method [15] or the SANN method [16]. These methods need no derivatives of the objective function. However, it is known that these methods converge slowly. In some special cases below, using a matrix differential calculus [17], the derivatives of the objective function are obtained. Hence, we can use some fast convergence methods, such as the quasi-Newton method and conjugate gradient method [18].

3.3.1 Order m Constrained to be an Integer

Assuming that a control parameter m is constrained to be an integer, the derivatives of the PLDA objective function are formulated as follows:

[†]We let the function f of a symmetric positive definite matrix $\mathbf{A}$ equal $\mathbf{U}diag(f(\lambda_1), \ldots, f(\lambda_n))\mathbf{U}^T = \mathbf{U}(f(\boldsymbol{\Lambda}))\mathbf{U}^T$, where $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, $\mathbf{U}$ denotes the matrix of n eigenvectors, and $\boldsymbol{\Lambda}$ denotes the diagonal matrix of eigenvalues, λ_i's. We may define the function f as some power or the logarithm of $\mathbf{A}$.

$$\frac{\partial}{\partial \mathbf{B}_{[p]}} \log J_{PLDA}\left(\mathbf{B}_{[p]}, m\right) = 2\boldsymbol{\Sigma}_n \mathbf{B}_{[p]} \tilde{\boldsymbol{\Sigma}}_n^{-1} - 2\mathbf{D}_m, \quad (27)$$

where

$$\mathbf{D}_m = \begin{cases} \dfrac{1}{m} \displaystyle\sum_{k=1}^{c} P_k \boldsymbol{\Sigma}_k \mathbf{B}_{[p]} \sum_{j=1}^{m} \mathbf{X}_{m,j,k}, & \text{if } m > 0 \\ \displaystyle\sum_{k=1}^{c} P_k \boldsymbol{\Sigma}_k \mathbf{B}_{[p]} \tilde{\boldsymbol{\Sigma}}_k^{-1}, & \text{if } m = 0 \\ -\dfrac{1}{m} \displaystyle\sum_{k=1}^{c} P_k \boldsymbol{\Sigma}_k \mathbf{B}_{[p]} \sum_{j=1}^{|m|} \mathbf{Y}_{m,j,k}, & \text{otherwise} \end{cases}$$

$$\mathbf{X}_{m,j,k} = \tilde{\boldsymbol{\Sigma}}_k^{m-j} \left(\sum_{l=1}^{c} P_l \tilde{\boldsymbol{\Sigma}}_l^{m} \right)^{-1} \tilde{\boldsymbol{\Sigma}}_k^{j-1},$$

and

$$\mathbf{Y}_{m,j,k} = \tilde{\boldsymbol{\Sigma}}_k^{m+j-1} \left(\sum_{l=1}^{c} P_l \tilde{\boldsymbol{\Sigma}}_l^{m} \right)^{-1} \tilde{\boldsymbol{\Sigma}}_k^{-j}.$$

This equation is used for acoustic models with full covariance.

3.3.2 $\tilde{\boldsymbol{\Sigma}}_k$ Constrained to be Diagonal

Because of computational simplicity, the covariance matrix in class k is often assumed to be diagonal [9], [10]. Since a diagonal matrix multiplication is commutative, the derivatives of the PLDA objective function are simplified as follows:

$$J_{PLDA}\left(\mathbf{B}_{[p]}, m\right) = \frac{\left|\tilde{\boldsymbol{\Sigma}}_n\right|}{\left|\left(\sum_{k=1}^{c} P_k diag(\tilde{\boldsymbol{\Sigma}}_k)^m \right)^{1/m}\right|}, \quad (28)$$

$$\frac{\partial}{\partial \mathbf{B}_{[p]}} \log J_{PLDA}\left(\mathbf{B}_{[p]}, m\right) = 2\boldsymbol{\Sigma}_n \mathbf{B}_{[p]} \tilde{\boldsymbol{\Sigma}}_n^{-1} - 2\mathbf{F}_m \mathbf{G}_m, \quad (29)$$

where

$$\mathbf{F}_m = \sum_{k=1}^{c} P_k \boldsymbol{\Sigma}_k \mathbf{B}_{[p]} diag\left(\tilde{\boldsymbol{\Sigma}}_k\right)^{m-1}, \quad (30)$$

$$\mathbf{G}_m = \left(\sum_{k=1}^{c} P_k diag\left(\tilde{\boldsymbol{\Sigma}}_k\right)^{m} \right)^{-1}, \quad (31)$$

and *diag* is an operator which sets zero for off-diagonal elements. In Eq. (28), the control parameter m can be any real number, unlike in Eq. (27).

When m is equal to zero, the PLDA objective function corresponds to the diagonal HDA (DHDA) objective function introduced in [10].

4. Selection of an Optimal Control Parameter

As shown in the previous section, PLDA can describe various criteria by varying its control parameter m. One way of obtaining an optimal control parameter m is to train HMMs and test recognition performance on a development set changing m and to choose the m with the smallest error. Unfortunately, this raises a considerable problem in a speech recognition task. In general, to train HMMs and to test recognition performance on a development set for finding an optimal control parameter requires several dozen hours. PLDA requires considerable time to select an optimal control parameter because it is able to choose a control parameter within a real number.

In this section we focus on a class separability error of the features in the projected space instead of using a recognition error on a development set. Better recognition performance can be obtained under the lower class separability error of projected features. Consequently, we measure the class separability error and use it as a criterion for the recognition performance comparison. We will define a class separability error of projected features.

4.1 Two-Class Problem

This subsection focuses on the two-class case. We first consider the Bayes error of the projected features on training data as a class separability error:

$$\varepsilon = \int \min[P_1 p_1(\mathbf{x}), P_2 p_2(\mathbf{x})] d\mathbf{x}, \quad (32)$$

where P_i denotes a prior probability of class i and $p_i(\mathbf{x})$ is a conditional density function of class i. The Bayes error ε can represent a classification error, assuming that training data and evaluation data come from the same distributions. However, it is difficult to directly measure the Bayes error. Instead, we use the Chernoff bound between class 1 and class 2 as a class separability error [6]:

$$\varepsilon_u^{1,2} = P_1^s P_2^{1-s} \int p_1^s(\mathbf{x}) p_2^{1-s}(\mathbf{x}) d\mathbf{x} \quad \text{for} \quad 0 \le s \le 1 \quad (33)$$

where ε_u indicates an upper bound of ε. In addition, when the $p_i(\mathbf{x})$'s are normal with mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$, the Chernoff bound between class 1 and class 2 becomes

$$\varepsilon_u^{1,2} = P_1^s P_2^{1-s} \exp(-\eta^{1,2}(s)), \quad (34)$$

where

$$\eta^{1,2}(s) = \frac{s(1-s)}{2} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{12}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_{12}|}{|\boldsymbol{\Sigma}_1|^s |\boldsymbol{\Sigma}_2|^{1-s}}, \quad (35)$$

where $\boldsymbol{\Sigma}_{12} \equiv s\boldsymbol{\Sigma}_1 + (1-s)\boldsymbol{\Sigma}_2$. In this case, ε_u can be obtained analytically and calculated rapidly.

In Fig. 2, two-dimensional two-class data are projected onto one-dimensional subspaces by two methods. To compare with their Chernoff bounds, the lower class separability error is obtained from the projected features by Method 1 as

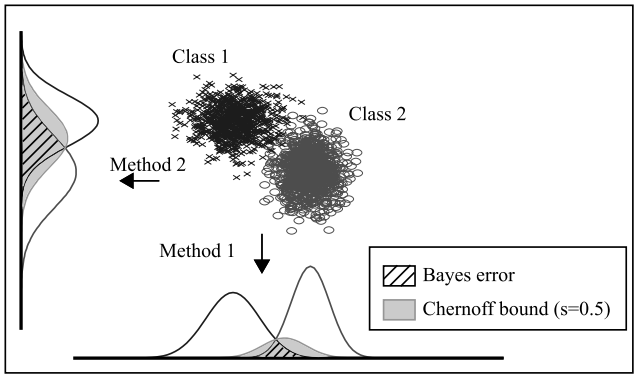


Fig. 2 Comparison of Bayes error and Chernoff bound.

compared with those by Method 2. In this case, Method 1 preserving the lower class separability error should be selected.

4.2 Extension to Multi-Class Problem

In the previous subsection, we defined a class separability error for two-class data. Here, we extend a two-class case to a multi-class case. Unlike the two-class case, it is possible to define several error functions for multi-class data. We define an error function as follows:

$$\tilde{\varepsilon}_u = \sum_{i=1}^{c} \sum_{j=1}^{c} I(i,j) \varepsilon_u^{i,j} \tag{36}$$

where $I(\cdot)$ denotes an indicator function. We consider the following three formulations as an indicator function.

4.2.1 Sum of Pairwise Approximated Errors

The sum of all the pairwise Chernoff bounds is defined using the following indicator function:

$$I(i,j) = \begin{cases} 1, & \text{if } j > i, \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

4.2.2 Maximum Pairwise Approximated Error

The maximum pairwise Chernoff bound is defined using the following indicator function:

$$I(i,j) = \begin{cases} 1, & \text{if } j > i \text{ and } (i,j) = (\hat{i},\hat{j}), \\ 0, & \text{otherwise,} \end{cases} \tag{38}$$

where $(\hat{i},\hat{j}) \equiv \arg\max_{i,j} \varepsilon_u^{i,j}$.

4.2.3 Sum of Maximum Approximated Errors in Each Class

The sum of the maximum pairwise Chernoff bounds in each class is defined using the following indicator function:

$$I(i,j) = \begin{cases} 1, & \text{if } j = \hat{j}_i, \\ 0, & \text{otherwise,} \end{cases} \tag{39}$$

where $\hat{j}_i \equiv \arg\max_{j} \varepsilon_u^{i,j}$.

5. Experiments

We conducted the experiments using the CENSREC-3 database [19]. The CENSREC-3 is designed as an evaluation framework of Japanese isolated word recognition in real driving car environments. Speech data were collected using 2 microphones, a close-talking (CT) microphone and a hands-free (HF) microphone. For training, driver's speech of phonetically-balanced sentences was recorded under two conditions: while idling and driving on a city street with normal in-car environment. A total of 14,050 utterances spoken by 293 drivers (202 males and 91 females) were recorded with both microphones. We used all utterances recorded with CT and HF microphones for training. For evaluation, driver's speech of isolated words was recorded under 16 environmental conditions using combinations of three kinds of vehicle speeds (idling, low-speed driving on a city street, and high-speed driving on an expressway) and six kinds of in-car environments (normal, with hazard lights on, with the air-conditioner on (fan low/high), with the audio CD player on, and with windows open). In these conditions, the "hazard lights on" condition was used only when idling. We only used three kinds of vehicle speeds in normal in-car environment and evaluated 2,646 utterances spoken by 18 speakers (8 males and 10 females) for each microphone. The speech signals for training and evaluation were both sampled at 16 kHz.

5.1 Baseline System

In the CENSREC-3, the baseline scripts are designed to facilitate HMM training and evaluation by HTK [20]. The acoustic models consisted of triphone HMMs. Each HMM had five states and three of them had output distributions. Each distribution was represented with 32 mixture diagonal Gaussians. The total number of states with the distributions were 2,000. The feature vector consisted of 12 MFCCs and log-energy with their corresponding delta and acceleration coefficients (39 dimensions). Frame length was 20 msec and frame shift was 10 msec. In the Mel-filter bank analysis, a cut-off was applied to frequency components lower than 250 Hz. The decoding process was performed without any language model. The vocabulary size was 100 words which included the original fifty words and another fifty similar-sounding words.

5.2 Dimensionality Reduction Procedure

The dimensionality reduction was performed using PCA, LDA, HDA, DHDA [10], and PLDA for the spliced features. Eleven successive frames (143 dimensions) were reduced to 39 dimensions. In (D)HDA and PLDA, to optimize Eq. (28),

we assumed that projected covariance matrices were diagonal and used the limited-memory BFGS algorithm as a numerical optimization technique [18]. The LDA projection matrix was used as the initial gradient matrix. To assign one of the classes to every feature after dimensionality reduction, HMM state labels were generated for the training data by a state-level forced alignment algorithm using a well-trained HMM system. The class number was 43 corresponding to the number of the monophones.

5.3 Experimental Results

Tables 1 and 2 show the word error rates and class separability errors according to Eqs. (37)–(39) for each dimensionality reduction criterion. The evaluation sets used in Tables 1 and 2 were recorded with CT and HF microphones, respectively. For the evaluation data recorded with a CT microphone, Table 1 shows that PLDA with $m = -0.5$ yields the lowest WER. For the evaluation data recorded with a HF microphone, the lowest WER is obtained by PLDA with a different control parameter ($m = -1.5$) in Table 2. In both cases with CT and HF microphones, PLDA with the optimal control parameters consistently outperformed the other criteria. Two data sets recorded with different microphones had different optimal control parameters. The analysis on the training data revealed that the voiced sounds had larger variances while the unvoiced sounds had smaller ones. As described in Sect. 3.3, PLDA with a smaller control parameter gives greater importance to the discrimination of classes with smaller variances. Thus, PLDA with a smaller control parameter has better ability to discriminate unvoiced sounds. In general, under noisy environment as with an HF microphone, discrimination of unvoiced sounds becomes difficult. Therefore, the optimal control parameter m for an HF microphone is smaller than with a CT microphone.

Table 1 Word error rates (%) and class separability errors according to Eqs. (37)–(39) for the evaluation set with a CT microphone. The best results are highlighted in bold.

	m	WER	Eq. (37)	Eq. (38)	Eq. (39)
MFCC +Δ +ΔΔ	-	7.45	2.31	0.0322	0.575
PCA	-	10.58	3.36	0.0354	0.669
LDA	-	8.78	3.10	0.0354	0.641
HDA	-	7.94	2.99	0.0361	0.635
PLDA	−3.0	6.73	2.02	0.0319	0.531
PLDA	−2.0	7.29	2.07	0.0316	0.532
PLDA	−1.5	6.27	**1.97**	0.0307	0.523
PLDA	−1.0	6.92	1.99	0.0301	**0.521**
PLDA	−0.5	**6.12**	2.01	**0.0292**	0.525
DHDA (PLDA)	- (0.0)	7.41	2.15	0.0296	0.541
PLDA	0.5	7.29	2.41	0.0306	0.560
PLDA	1.0	9.33	3.09	0.0354	0.641
PLDA	1.5	8.96	4.61	0.0394	0.742
PLDA	2.0	8.58	4.65	0.0404	0.745
PLDA	3.0	9.41	4.73	0.0413	0.756

Table 2 Word error rates (%) and class separability errors according to Eqs. (37)–(39) for the evaluation set with an HF microphone.

	m	WER	Eq. (37)	Eq. (38)	Eq. (39)
MFCC +Δ+ΔΔ	-	15.04	2.56	0.0356	0.648
PCA	-	19.39	3.65	0.0377	0.738
LDA	-	15.80	3.38	0.0370	0.711
HDA	-	17.16	3.21	0.0371	0.697
PLDA	−3.0	15.04	2.19	0.0338	0.600
PLDA	−2.0	12.32	2.26	0.0339	0.602
PLDA	−1.5	**10.70**	**2.18**	0.0332	**0.5921**
PLDA	−1.0	11.49	2.23	**0.0327**	0.5922
PLDA	−0.5	12.51	2.31	0.0329	0.598
DHDA (PLDA)	- (0.0)	14.17	2.50	0.0331	0.619
PLDA	0.5	13.53	2.81	0.0341	0.644
PLDA	1.0	16.97	3.38	0.0370	0.711
PLDA	1.5	17.31	5.13	0.0403	0.828
PLDA	2.0	15.91	5.22	0.0412	0.835
PLDA	3.0	16.36	5.36	0.0424	0.850

In comparing dimensionality reduction criteria without training HMMs nor testing recognition performance on a development set, we used $s = 1/2$ for the Chernoff bound computation because there was no *a priori* information about weights of two class distributions. In the case of $s = 1/2$, Eq. (33) is called the Bhattacharyya bound. Two covariance matrices in Eq. (35) were treated as diagonal because diagonal Gaussians were used to model HMMs. The parameter selection was performed as follows: To select the optimal control parameter for the data set recorded with a CT microphone, all the training data with a CT microphone were labeled with monophones using a forced alignment recognizer. Then, each monophone was modeled as a unimodal normal distribution, and the mean vector and covariance matrix of each class were calculated. Chernoff bounds were obtained using these mean vectors and covariance matrices. The optimal control parameter for the data set with an HF microphone was obtained using all of the training data with an HF microphone through the same process as a CT microphone. Both Tables 1 and 2 show that the results of the proposed method and relative recognition performance agree well. There was little difference in the parameter selection performances among Eqs. (37)–(39) in parameter selection accuracy. The proposed selection method yielded sub-optimal performance without training HMMs nor testing recognition performance on a development set, although it neglected time information of speech feature sequences to measure a class separability error and modeled a class distribution as a unimodal normal distribution. In addition, the optimal control parameter value can vary with different speech features, a different language, or a different noise environment. The proposed selection method can adapt to such variations.

5.4 Discriminative Training Results

PLDA can combine a discriminative training technique of HMMs, such as maximum mutual information (MMI) and minimum phone error (MPE) [21]–[23]. We also conducted the same experiments using MMI and MPE by HTK and

Table 3 Word error rates (%) using a maximum likelihood training and three discriminative trainings for the evaluation set with a CT microphone.

	ML	MMI	MPE (approx.)	MPE (exact)
MFCC +Δ+ΔΔ	7.45	7.14	6.92	6.95
PLDA (m=-0.5)	6.12	5.71	5.06	4.99

Table 4 Word error rates (%) using a maximum likelihood training and three discriminative trainings for the evaluation set with an HF microphone.

	ML	MMI	MPE (approx.)	MPE (exact)
MFCC +Δ+ΔΔ	15.04	14.44	18.67	15.99
PLDA (m=-1.5)	10.70	10.39	9.44	10.28

Table 5 Computational costs with the conventional and proposed method.

conventional	220 h = (15 h (training) + 5 h (test)) × 11 conditions
proposed	0.87 h = 30 min (mean and variance calculations) + 2 min (Chernoff bound calculation) × 11 conditions

compared a maximum likelihood (ML) training, MMI, approximate MPE and exact MPE. The approximate MPE assigns approximate correctness to phones while the exact MPE assigns exact correctness to phones. The former is faster in computation for assigning correctness, and the latter is more precise in correctness. The results are shown in Tables 3 and 4. By combining PLDA and the discriminative training techniques, we obtained better performance than the PLDA with a maximum likelihood criterion training. There appears to be no consistent difference between approximate and exact MPE as reported in a discriminative training study [23].

5.5 Computational Costs

The computational costs for the evaluation of recognition performance versus the proposed selection method are shown in Table 5. Here, the computational cost involves the optimization procedure of the control parameter. In this experiment, we evaluate the computational costs on the evaluation data set with a Pentium IV 2.8 GHz computer. For every dimensionality reduction criterion, the evaluation of recognition performance required 15 hours for training of HMMs and 5 hours for test on a development set. In total, 220 hours were required for comparing 11 dimensionality reduction criteria (PLDAs using 11 different control parameters). On the other hand, the proposed selection method only required approximately 30 minutes for calculating statistical values such as mean vectors and covariance matrices of each class in the original space. After this, 2 minutes were required to calculate Eqs. (37)–(39) for each dimensionality reduction criterion. In total, only 0.87 hour was required for predicting the optimal criterion among the 11 dimensionality reduction criteria described above. Thus, the proposed method could perform the prediction process much faster than a conventional procedure that included training of HMMs and test of recognition performance on a development set.

6. Conclusions

In this paper we proposed a new framework for integrating various criteria to reduce dimensionality. The new framework which we call power linear discriminant analysis (PLDA) includes LDA, HLDA, and HDA criteria as special cases. Next, an efficient selection method of an optimal PLDA control parameter was introduced. The method used the Chernoff bound as a measure of a class separability error which was the upper bound of the Bayes error. The experimental results on the CENSREC-3 database demonstrated that segmental unit input HMM with PLDA gave better performance than the others and that PLDA with a control parameter selected by the proposed efficient selection method yielded sub-optimal performance with a drastic reduction of computational costs.

Acknowledgments

The authors would like to thank Dr. Manabu Otsuka for his useful comments.

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

References

[1] S. Nakagawa and K. Yamamoto, "Evaluation of segmental unit input HMM," Proc. ICASSP, pp.439–442, 1996.
[2] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," IEEE Trans. Acoust. Speech Signal Process., vol.37, no.12, pp.1857–1869, 1989.
[3] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," Proc. ICASSP, pp.13–16, 1992.
[4] H. Gish and M. Russell, "Parametric trajectory models for speech recognition," Proc. ICSLP, pp.466–469, 1996.
[5] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features," Proc. Eurospeech 2003, pp.865–868, 2003.
[6] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1990.
[7] R.O. Duda, P.B. Hart, and D.G. Stork, Pattern Classification, John Wiley & Sons, New York, 2001.
[8] N.A. Campbell, "Canonical variate analysis — A general model formulation," Australian Journal of Statistics, vol.4, pp.86–96, 1984.
[9] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," Speech Commun., pp.283–297, 1998.
[10] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," Proc. ICASSP, pp.129–132, 2000.
[11] F. de la Torre and T. Kanade, "Oriented discriminant analysis," British Machine Vision Conference, pp.132–141, 2004.

[12] M. Loog and R. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The chernoff criterion," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.6, pp.732–739, 2004.
[13] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE Trans. Speech Audio Process., vol.7, no.3, pp.272–281, 1999.
[14] J.R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, John Wiley & Sons, 1999.
[15] J.A. Nelder and R. Mead, "A simplex method for function minimization," Comput. J., vol.7, pp.308–313, 1965.
[16] C.J.P. Belisle, "Convergence theorems for a class of simulated annealing algorithms," J. Applied Probability, vol.29, pp.885–892, 1992.
[17] S.R. Searle, Matrix Algebra Useful for Statistics, Wiley Series in Probability and Mathematical Statistics, New York, 1982.
[18] J. Nocedal and S.J. Wright, Numerical Optimization, Springer-Verlag, 1999.
[19] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-3: An evaluation framework for Japanese speech recognition in real driving-car environments," IEICE Trans. Inf. & Syst., vol.E89-D, no.11, pp.2783–2793, Nov. 2006.
[20] HTK Web site. http://htk.eng.cam.ac.uk/
[21] L. Bahl, P. Brown, P. de Sousa, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," Proc. ICASSP, pp.49–52, 1986.
[22] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," Proc. ICASSP, pp.105–108, 2002.
[23] D. Povey, Discriminative Training for Large Vocabulary Speech Recognition, Ph.D. Thesis, Cambridge University, 2003.

Appendix A: Proof of (23)

Theorem 1. Let $\tilde{\mathbf{\Sigma}}_k$ $(1 \le k \le c)$ be symmetric positive definite matrices. Then,

$$\lim_{m\to 0} J_{PLDA}(\mathbf{B}_{[p]}, m) = \frac{\left|\tilde{\mathbf{\Sigma}}_n\right|}{\prod_{k=1}^{c}\left|\tilde{\mathbf{\Sigma}}_k\right|^{P_k}}. \quad (\text{A}\cdot 1)$$

Proof. Here, we focus on the denominator of a PLDA objective function. We let

$$f(m) = \log\left|\sum_{i=1}^{c} P_i \tilde{\mathbf{\Sigma}}_i^m\right| \quad (\text{A}\cdot 2)$$

and $g(m) = m$, so that

$$\log\left|\left(\sum_{i=1}^{c} P_i \tilde{\mathbf{\Sigma}}_i^m\right)^{1/m}\right| = \frac{1}{m}\log\left|\sum_{i=1}^{c} P_i \tilde{\mathbf{\Sigma}}_i^m\right| = \frac{f(m)}{g(m)}. \quad (\text{A}\cdot 3)$$

Then $f(0) = g(0) = 0$, and

$$\frac{\partial f(m)}{\partial m} = tr\left(\mathbf{Z}_m \sum_i P_i \frac{\partial}{\partial m}\tilde{\mathbf{\Sigma}}_i^m\right) \quad (\text{A}\cdot 4)$$

$$= tr\left(\mathbf{Z}_m \sum_i P_i \mathbf{U}_i \left(\frac{\partial}{\partial m}\mathbf{\Lambda}_i^m\right)\mathbf{U}_i^T\right) \quad (\text{A}\cdot 5)$$

$$= tr\left(\mathbf{Z}_m \sum_i P_i \mathbf{U}_i \mathbf{\Lambda}_i^m (\log \mathbf{\Lambda}_i)\,\mathbf{U}_i^T\right), \quad (\text{A}\cdot 6)$$

$$\frac{\partial g(m)}{\partial m} = 1, \quad (\text{A}\cdot 7)$$

where $\mathbf{Z}_m = \left(\sum_j P_j \tilde{\mathbf{\Sigma}}_j^m\right)^{-1}$, $\mathbf{U}_i$ denotes the matrix of eigenvectors of $\tilde{\mathbf{\Sigma}}_i^m$, and $\mathbf{\Lambda}_i$ denotes the diagonal matrix of eigenvalues of $\tilde{\mathbf{\Sigma}}_i^m$.

By l'Hôpital's rule,

$$\lim_{m\to 0}\frac{f(m)}{g(m)} = \lim_{m\to 0}\frac{f'(m)}{g'(m)} = \frac{f'(0)}{g'(0)} \quad (\text{A}\cdot 8)$$

$$= \sum_i P_i tr(\log \mathbf{\Lambda}_i) \quad (\text{A}\cdot 9)$$

$$= \log\prod_i \left|\tilde{\mathbf{\Sigma}}_i\right|^{P_i}, \quad (\text{A}\cdot 10)$$

and (A· 1) follows. □

Appendix B: Proofs of (25) and (26)

Theorem 2. Let $|\tilde{\mathbf{\Sigma}}_k| = \max_i |\tilde{\mathbf{\Sigma}}_i|$ (k is not necessarily unique). If $\tilde{\mathbf{\Sigma}}_k$ satisfies $\tilde{\mathbf{\Sigma}}_k^m \ge \sum_i P_i \tilde{\mathbf{\Sigma}}_i^m$, then

$$J_{PLDA}\left(\mathbf{B}_{[p]}, \infty\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_n\right|}{\max_i \left|\tilde{\mathbf{\Sigma}}_i\right|}, \quad (\text{A}\cdot 11)$$

$$J_{PLDA}\left(\mathbf{B}_{[p]}, -\infty\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_n\right|}{\min_i \left|\tilde{\mathbf{\Sigma}}_i\right|}, \quad (\text{A}\cdot 12)$$

where $\mathbf{X} \ge \mathbf{Y}$ denotes that $(\mathbf{X} - \mathbf{Y})$ is a positive semidefinite matrix.

Proof. To prove (A· 11), let

$$\phi(m) = \left|\left(\sum_i P_i \tilde{\mathbf{\Sigma}}_i^m\right)^{1/m}\right|. \quad (\text{A}\cdot 13)$$

We have the following inequality[†]:

$$\left|\sum_i P_i \tilde{\mathbf{\Sigma}}_i^m\right| \ge \left|P_k \tilde{\mathbf{\Sigma}}_k^m\right|. \quad (\text{A}\cdot 14)$$

Using this, for $m > 0$, we obtain

$$\phi(m) \ge \left|P_k \tilde{\mathbf{\Sigma}}_k^m\right|^{1/m} \quad (\text{A}\cdot 15)$$

$$= \left|P_k^{1/m} \tilde{\mathbf{\Sigma}}_k\right| \quad (\text{A}\cdot 16)$$

$$= \left|\tilde{\mathbf{\Sigma}}_k\right| \ (m \to \infty). \quad (\text{A}\cdot 17)$$

Added to this, since we assume that $\tilde{\mathbf{\Sigma}}_k$ satisfies $\tilde{\mathbf{\Sigma}}_k^m \ge \sum_i P_i \tilde{\mathbf{\Sigma}}_i^m$, we also obtain

$$\left|\tilde{\mathbf{\Sigma}}_k^m\right| = \left|\sum_i P_i \tilde{\mathbf{\Sigma}}_i^m + \tilde{\mathbf{\Sigma}}_k^m - \sum_i P_i \tilde{\mathbf{\Sigma}}_i^m\right| \quad (\text{A}\cdot 18)$$

[†]The inequality follows from observing that $|\mathbf{X} + \mathbf{Y}| \ge |\mathbf{X}|$ for symmetric positive definite matrices $\mathbf{X}$ and $\mathbf{Y}$.

$$\geq \left| \sum_i P_i \tilde{\boldsymbol{\Sigma}}_i^m \right|. \tag{A·19}$$

Hence,

$$\left|\tilde{\boldsymbol{\Sigma}}_k\right| \geq \left| \sum_i P_i \tilde{\boldsymbol{\Sigma}}_i^m \right|^{1/m} = \phi(m). \tag{A·20}$$

(A·17) and (A·20) imply

$$\lim_{m\to\infty} \phi(m) = \left|\tilde{\boldsymbol{\Sigma}}_k\right| = \max_i \left|\tilde{\boldsymbol{\Sigma}}_i\right|. \tag{A·21}$$

(A·11) then follows.

To prove (A·12), let $n = -m$ and $\check{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{\Sigma}}_i^{-1}$. Then

$$\phi(m) = \left(\left| \sum_i P_i \check{\boldsymbol{\Sigma}}_i^n \right|^{1/n} \right)^{-1} \tag{A·22}$$

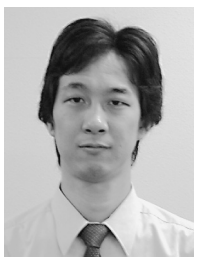
and hence

$$\lim_{m\to-\infty} \phi(m) = \lim_{n\to\infty} \left(\left| \sum_i P_i \check{\boldsymbol{\Sigma}}_i^n \right|^{1/n} \right)^{-1} \tag{A·23}$$

$$= \left(\max_i \left|\check{\boldsymbol{\Sigma}}_i\right| \right)^{-1} \tag{A·24}$$

$$= \min_i |\tilde{\boldsymbol{\Sigma}}_i|, \tag{A·25}$$

and (A·12) follows. □



Makoto Sakai received his B.E. and M.E. degrees from Nagoya Institute of Technology in 1997 and 1999, respectively. He is currently with the Research Laboratories, DENSO CORPORATION, Japan. He is also pursuing a Ph.D. degree in the Department of Media Science, Nagoya University. His research interests include automatic speech recognition, signal processing, and machine learning.

Norihide Kitaoka received his B.E. and M.E. degrees from Kyoto University in 1992 and 1994, respectively, and a Dr. Eng. degree from Toyohashi University of Technology in 2000. He joined DENSO CORPORATION, Japan in 1994. He then joined the Department of Information and Computer Sciences at Toyohashi University of Technology as a research associate in 2001 and was a lecturer from 2003 to 2006. Currently he is an Associate Professor of the Graduate School of Information Science, Nagoya University. His research interests include speech processing, speech recognition and spoken dialog. He is a member of the Information Processing Society of Japan (IPSJ), the Acoustical Society of Japan (ASJ) and the Japan Society for Artificial Intelligence (JSAI).

Seiichi Nakagawa received Dr. of Eng. degree from Kyoto University in 1977. He joined the Faculty of Kyoto University, in 1976, as a Research Associate in the Department of Information Sciences. He moved to Toyohashi University of Technology in 1980. From 1980 to 1983, he was an Assistant Professor, and from 1983 to 1990 he was an Associate Professor. Since 1990 he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, USA. He received the 1997/2001 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electro. Telecomm. Engrs. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence. He is a fellow of IPSJ.