We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

5,400
Open access books available

133,000
International authors and editors

165M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Feature Transformation Based on Generalization of Linear Discriminant Analysis

Makoto Sakai[1,2], Norihide Kitaoka[2] and Seiichi Nakagawa[3]
[1]DENSO CORP.,
[2]Nagoya University,
[3]Toyohashi University of Technology
Japan

1. Introduction

Hidden Markov models (HMMs) have been widely used to model speech signals for speech recognition. However, they cannot precisely model the time dependency of feature parameters. In order to overcome this limitation, several researchers have proposed extensions, such as segmental unit input HMM (Nakagawa & Yamamoto, 1996). Segmental unit input HMM 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 dimensions. The concatenated vectors may have strong correlations among dimensions, and may include nonessential information. In addition, high-dimensional data require a heavy computational load. Therefore, to reduce dimensionality, a feature transformation method is often applied. Linear discriminant analysis (LDA) is widely used to reduce dimensionality and a powerful tool to preserve discriminative information. LDA assumes each class has the same class covariance. However, this assumption does not necessarily hold for a real data set. In order to remove 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. Heteroscedastic discriminant analysis (HDA) was proposed as another objective function, which employed individual weighted contributions of the classes. 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 chapter we show that these three methods have a strong mutual relationship, and provide a new interpretation for them. Then, we present a new framework that we call power linear discriminant analysis (PLDA) (Sakai et al., 2007), which can describe various criteria including the discriminant analyses 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. Moreover, the presented technique can combine a discriminative training, such as maximum mutual information (MMI) and minimum phone error (MPE). Experimental results show the effectiveness of the presented technique.

2. Notations

This chapter 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 $\text{tr}(\mathbf{A})$ is the trace of the matrix.

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

3. 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 (Nakagawa & Yamamoto, 1996):

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

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

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

where T denotes the length of input sequence and d denotes the number of successive frames used in probability calculation at a current 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 parameter estimation degrades. Therefore, to reduce dimensionality, feature transformation methods, e.g., principal component analysis (PCA), LDA, HLDA or HDA, are often used (Nakagawa & Yamamoto, 1996; Haeb-Umbach & Ney, 1992; Kumar & Andreou, 1998; Saon et al., 2000). Here, we briefly review LDA, HLDA and HDA, and then investigate the effectiveness of these methods for some artificial data sets.

3.1 Linear discriminant analysis

Given n-dimensional feature vectors $\mathbf{x}_j \in \Re^n$ $(j = 1, 2, \cdots, N)$, e.g., $\mathbf{x}_j = \left[\mathbf{o}_{j-(d-1)}^T, \cdots, \mathbf{o}_j^T\right]^T$, let us find a transformation matrix $\mathbf{B}_{[p]} \in \Re^{n \times p}$ that projects these feature vectors to p-dimensional feature vectors $\mathbf{z}_j \in \Re^p$ $(j = 1, 2, \cdots, N)$ $(p < n)$, 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 (Fukunaga, 1990):

$$\Sigma_w = \frac{1}{N} \sum_{k=1}^{c} \sum_{x_j \in D_k} \left(x_j - \mu_k \right)\left(x_j - \mu_k \right)^T$$

$$= \sum_{k=1}^{c} P_k \Sigma_k, \tag{4}$$

$$\Sigma_b = \sum_{k=1}^{c} P_k \left(\mu_k - \mu \right)\left(\mu_k - \mu \right)^T, \tag{5}$$

where c denotes the number of classes, D_k denotes the subset of feature vectors labeled as class k, μ is the mean vector of all features, μ_k is the mean vector of the class k, Σ_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 (Fukunaga, 1990). Typical objective functions are the following:

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

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

where Σ_t denotes the covariance matrix of all features, namely a total covariance, which equals $\Sigma_b + \Sigma_w$.

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

3.2 Heteroscedastic extensions

LDA is not the optimal transformation when the class distributions are heteroscedastic. Campbell has shown that LDA is related to the maximum likelihood estimation of parameters for a Gaussian model with an identical class covariance (Campbell, 1984). However, this condition is not necessarily satisfied for a real data set.

In order to overcome this limitation, several extensions have been proposed. This chapter focuses on two heteroscedastic extensions called heteroscedastic linear discriminant analysis (HLDA) (Kumar & Andreou, 1998) and heteroscedastic discriminant analysis (HDA) (Saon et al., 2000).

3.2.1 Heteroscedastic linear discriminant analysis

In HLDA, the full-rank linear transformation matrix $\mathbf{B} \in \Re^{n \times n}$ 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 $n\text{-}p$ columns of $\mathbf{B}$ span the ($n\text{-}p$)-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 $\hat{\mu}_k$ and $\hat{\Sigma}_k$, respectively:

$$\hat{\mu}_k = \begin{bmatrix} B_{[p]}^T \mu_k \\ B_{[n-p]}^T \mu \end{bmatrix}, \tag{8}$$

$$\hat{\Sigma}_k = \begin{bmatrix} B_{[p]}^T \Sigma_k B_{[p]} & 0 \\ 0 & B_{[n-p]}^T \Sigma_t B_{[n-p]} \end{bmatrix}, \tag{9}$$

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

Kumar et al. incorporated the maximum likelihood estimation of parameters for differently distributed Gaussians. An HLDA objective function is derived as follows (Kumar & Andreou, 1998):

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

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.

3.2.2 Heteroscedastic discriminant analysis

HDA uses the following objective function, which incorporates individual weighted contributions of the class variances (Saon et al., 2000):

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

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

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

3.3 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. Fig. 1 (a) shows that HDA has higher separability than LDA for the data set used in (Saon et al., 2000). On the other hand, as shown in Fig. 1(b), LDA has higher separability than HDA for another data set. Fig. 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.
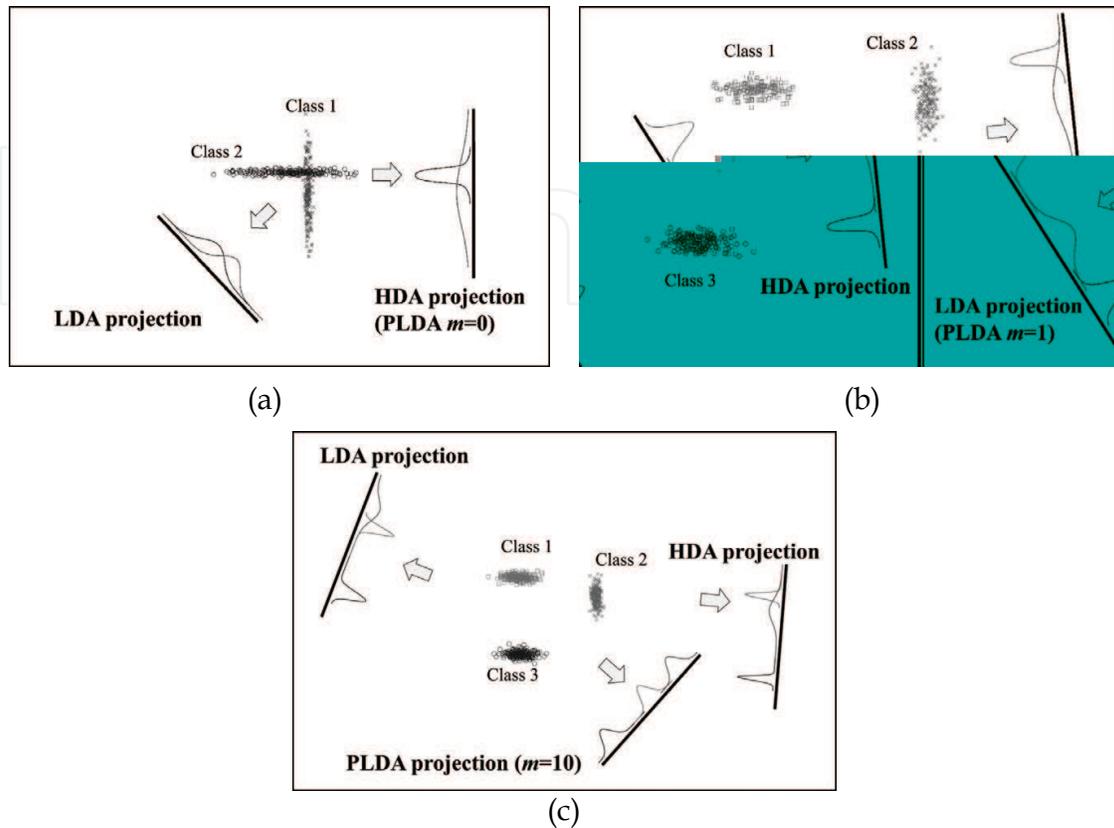

(a)　　　　　　　　　　　　　　　　　　(b)

(c)

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

4. 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.

4.1 Relationship between HLDA and HDA

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

$$\mathbf{B}^T \Sigma_t \mathbf{B} = \mathbf{B}^T \Sigma_b \mathbf{B} + \mathbf{B}^T \Sigma_w \mathbf{B} \tag{13}$$

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

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

where $\hat{\mu} = \mathbf{B}^T \mu$.
The determinant of this is

$$\left| \mathbf{B}^T \Sigma_t \mathbf{B} \right| = \left| \mathbf{B}_{[p]}^T \Sigma_t \mathbf{B}_{[p]} \right| \left| \mathbf{B}_{[n-p]}^T \Sigma_t \mathbf{B}_{[n-p]} \right|, \tag{16}$$

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

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

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.

4.2 Relationship between LDA and HDA

The LDA and HDA objective functions can be rewritten as

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

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

where $\tilde{\Sigma}_b = \mathbf{B}_{[p]}^T \Sigma_b \mathbf{B}_{[p]}$ and $\tilde{\Sigma}_k = \mathbf{B}_{[p]}^T \Sigma_k \mathbf{B}_{[p]}$ 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.

4.3 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* (Magnus & Neudecker, 1999). We have extended this notion to a determinant of a matrix mean and have proposed a new objective function as follows (Sakai et al., 2007):

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

where $\tilde{\Sigma}_n \in \left\{ \tilde{\Sigma}_b, \tilde{\Sigma}_t \right\}, \tilde{\Sigma}_t = \mathbf{B}_{[p]}^T \Sigma_t \mathbf{B}_{[p]}$, 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)

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

- $m=1$ (arithmetic mean)

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

- $m \to 0$ (geometric mean)

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

- $m=-1$ (harmonic mean)

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

The following equations are also obtained under a particular condition.

- $m \to \infty$

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

- $m \to -\infty$

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

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). Fig. 1 (c) shows that PLDA with $m=10$ 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 **B**, we can use numerical optimization techniques such as the Nelder-Mead method or the SANN method. 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, 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.

4.3.1 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:

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

where

$$\mathbf{D}_m = \begin{cases} \dfrac{1}{m} \sum_{k=1}^{c} P_k \Sigma_k \mathbf{B}_{[p]} \sum_{j=1}^{m} \mathbf{X}_{m,j,k}, & \text{if } m > 0 \\[2ex] \sum_{k=1}^{c} P_k \Sigma_k \mathbf{B}_{[p]} \tilde{\Sigma}_k^{-1}, & \text{if } m = 0 \\[2ex] -\dfrac{1}{m} \sum_{k=1}^{c} P_k \Sigma_k \mathbf{B}_{[p]} \sum_{j=1}^{|m|} \mathbf{Y}_{m,j,k}, & \text{otherwise} \end{cases}$$

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

and

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

This equation can be used for acoustic models with full covariance.

4.3.2 $\tilde{\Sigma}_k$ constrained to be diagonal

Because of computational simplicity, the covariance matrix of class *k* is often assumed to be diagonal (Kumar & Andreou, 1998; Saon et al., 2000). Since a diagonal matrix multiplication is commutative, the derivatives of the PLDA objective function are simplified as follows:

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

$$\frac{\partial}{\partial B_{[p]}} \log J_{PLDA}\left(B_{[p]}, m\right) = 2\Sigma_n B_{[p]} \tilde{\Sigma}_n^{-1} - 2F_m G_m, \tag{29}$$

where

$$F_m = \sum_{k=1}^{c} P_k \Sigma_k B_{[p]} diag(\tilde{\Sigma}_k)^{m-1}, \tag{30}$$

$$G_m = \left(\sum_{k=1}^{c} P_k diag(\tilde{\Sigma}_k)^m \right)^{-1}, \tag{31}$$

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 (Saon et al., 1990).

5. 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 define a class separability error of projected features.
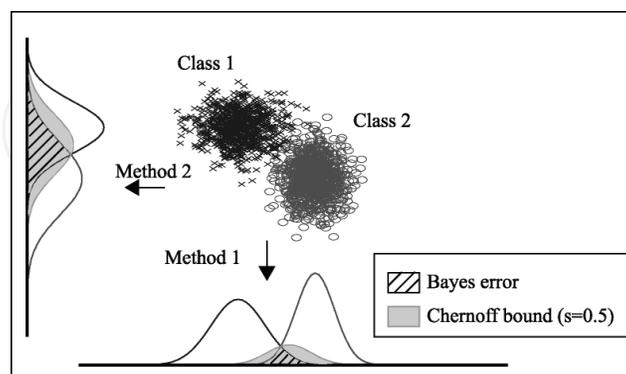

Fig. 2. Comparison of Bayes error and Chernoff bound.

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

$$\varepsilon = \int \min\left[P_1 p_1(\mathrm{x}), P_2 p_2(\mathrm{x}) \right] d\mathrm{x}, \tag{32}$$

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 (Fukunaga, 1990):

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

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

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

where

$$\eta^{1,2}(s) = \frac{s(1-s)}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma_{12}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \frac{\left| \Sigma_{12} \right|}{\left| \Sigma_1 \right|^s \left| \Sigma_2 \right|^{1-s}}, \tag{35}$$

where $\Sigma_{12} \equiv s\boldsymbol{\Sigma}_1 + (1-s)\boldsymbol{\Sigma}_2$. 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 compared with those by Method 2. In this case, Method 1 preserving the lower class separability error should be selected.

5.2 Extension to multi-class problem

In Section 5.1, 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.

5.2.1 Sum of pairwise approximated errors

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

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

5.2.2 Maximum pairwise approximated error

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

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

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

5.2.3 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:

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

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

6. Combination of feature transformation and discriminative training

Feature transformation aims to transform high dimensional features to low dimensional features in a feature space while separating different classes such as monophones. Discriminative trainings, such as maximum mutual information (MMI) (Bahl et al., 1986) and minimum phone error (MPE) (Povey & Woodland, 2002), estimate the acoustic models discriminatively in a model space (Fig. 3). Because feature transformation and discriminative training are adopted at different levels, a combination of them can have a complementary effect on speech recognition.
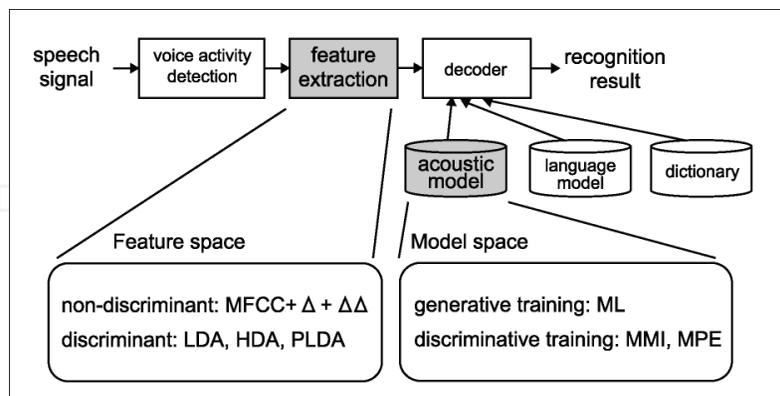

Fig. 3. Feature transformation and discriminative training.

6.1 Maximum mutual information (MMI)

The MMI criterion is defined as follows (Bahl et al., 1986):

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(O_r \mid s_r)^\kappa P(s_r)}{\sum_s p_\lambda(O_r \mid s)^\kappa P(s)}, \tag{40}$$

where λ is the set of HMM parameters, $\mathbf{O}_r$ is the r'th training sentence, R denotes the number of training sentences, κ is an acoustic de-weighting factor which can be adjusted to improve the test set performance, $p_\lambda(\mathbf{O}_r|s)$ is the likelihood given sentence s, and $P(s)$ is the language model probability for sentence s. The MMI criterion equals the multiplication of the posterior probabilities of the correct sentences s_r.

6.2 Minimum phone error (MPE)

MPE training aims to minimize the phone classification error (or maximize the phone accuracy) (Povey & Woodland, 2002). The objective function to be maximized by the MPE training is expressed as

$$F_{MPE}(\lambda) = \sum_{r=1}^{R} \frac{\sum_s p_\lambda(\mathrm{O}_r \mid s)^\kappa P(s) A(s, s_r)}{\sum_{s'} p_\lambda(\mathrm{O}_r \mid s')^\kappa P(s')}, \tag{41}$$

where $A(s, s_r)$ represents the raw phone transcription accuracy of the sentence s given the correct sentence s_r, which equals the number of correct phones minus the number of errors.

7. Experiments

We conducted experiments on CENSREC-3 database (Fujimoto et al., 2006), 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. The data recorded with an HF microphone tend to have higher noise than those recorded with a CT microphone because the HF microphone is attached to the driver's sun visor. 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. A total of 28,100 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, we used driver's speech of isolated words recorded with CT and HF microphones under a 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.

7.1 Baseline system

In the CENSREC-3, the baseline scripts are designed to facilitate HMM training and evaluation by HTK (available at http://htk.eng.cam.ac.uk/). 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 was 2,000. The feature vector consisted of 12 MFCCs and log-energy with their corresponding delta and acceleration coefficients (total 39 dimensions). Frame length and frame shift were 20 msec and 10 msec, respectively. 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.

7.2 Dimensionality reduction procedure

The dimensionality reduction was performed using PCA, LDA, HDA, DHDA (Saon et al., 2000), and PLDA for concatenated 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. The LDA transformation 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.

7.3 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 Section 4.3, 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. 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.

	m	WER	Eq. (37)	Eq. (38)	Eq. (39)
MFCC+ Δ + $\Delta\Delta$	-	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 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 + $\Delta\Delta$	-	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

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

7.4 Discriminative training results

We also conducted the same experiments using MMI and MPE by HTK and 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 (Povey, 2003).

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

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	15.04	14.44	18.67	15.99
PLDA	10.70	10.39	9.44	10.28

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

7.5 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 five hours for test on a development set. In total, 220 hours were required for comparing 11 feature transformations (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 feature transformation. In total, only 0.87 hour was required for predicting the sub-optimal feature transformation among the 11 feature transformation 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.

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

Table 5. Computational costs with the conventional and proposed methods.

8. Conclusions

In this chapter we presented a new framework for integrating various criteria to reduce dimensionality. The framework, termed power linear discriminant analysis, 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 presented efficient selection method yielded sub-optimal performance with a drastic reduction of computational costs.

9. References

Bahl, L., Brown, P., de Sousa, P. & Mercer, R. (1986). Maximul mutual information estimation of hidden Markov model parameters for speech recognition, *Proceedings of IEEE Int. Conf. on Acoustic Speech and Signal Processing*, pp. 49-52.

Campbell, N. A. (1984). Canonical variate analysis – A general model formulation, *Australian Journal of Statistics*, Vol.4, pp. 86-96.

Fujimoto, M., Takeda, K. & Nakamura, S. (2006). CENSREC-3 : An evaluation framework for Japanese speech recognition in real driving-car environments, *IEICE Trans. Inf. & Syst.*, Vol. E89-D, pp. 2783-2793.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, Academic Press, New York.

Haeb-Umbach, R. & Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. *Proceedings of IEEE Int. Conf. on Acoustic Speech and Signal Processing*, pp. 13-16.

Kumar, N. & Andreou, A. G. (1998). Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, *Speech Communication*, pp. 283-297.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons.

Nakagawa, S. & Yamamoto, K. (1996). Evaluation of segmental unit input HMM. *Proceedings of IEEE Int. Conf. on Acoustic Speech and Signal Processing*, pp. 439-442.

Povey, D. & Woodland, P. (2002). Minimum phone error and I-smoothing for improved discriminative training, *Proceedings of IEEE Int. Conf. on Acoustic Speech and Signal Processing*, pp. 105-108.

Povey, D. (2003). *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. Thesis, Cambridge University.

Sakai, M., Kitaoka, N. & Nakagawa, S. (2007). Generalization of linear discriminant analysis used in segmental unit input HMM for speech recognition, *Proceedings of IEEE Int. Conf. on Acoustic Speech and Signal Processing*, pp. 333-336.

Saon, G., Padmanabhan, M., Gopinath, R. & Chen, S. (2000). Maximum likelihood discriminant feature spaces, *Proceedings of IEEE Int. Conf. on Acoustic Speech and Signal Processing*, pp. 129-132.

Speech Recognition

Edited by France Mihelic and Janez Zibert

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Makoto Sakai, Norihide Kitaoka and Seiichi Nakagawa (2008). Feature Transformation Based on Generalization of Linear Discriminant Analysis, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from:
http://www.intechopen.com/books/speech_recognition/feature_transformation_based_on_generalization_of_linear_discriminant_analysis

INTECH
open science | open minds