GENERALIZATION OF LINEAR DISCRIMINANT ANALYSIS USED IN SEGMENTAL UNIT INPUT HMM FOR SPEECH RECOGNITION

Makoto Sakai[1,2]*, Norihide Kitaoka*[2,3]*, Seiichi Nakagawa*[2]

[1] DENSO CORPORATION, Nisshin 470-0111, Japan
[2] Toyohashi University of Technology, Toyohashi 441-8580, Japan
[3] Nagoya University, Nagoya 464-8601, Japan
msakai@rlab.denso.co.jp, {kitaoka, nakagawa}@slp.ics.tut.ac.jp

ABSTRACT

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 is widely used to address this issue. Linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA) are classical and 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 describe various criteria including LDA and HDA with one parameter. Experimental results show that the PLDA is more effective than PCA, LDA, and HDA for various data sets.

Index Terms— Speech recognition, Feature extraction, Multidimensional signal processing

1. INTRODUCTION

Hidden Markov Models (HMMs) have been widely used to model speech signals for speech recognition. However, HMMs cannot precisely model the time dependency of feature parameters. In order to overcome this limitation, many extensions have been proposed [1–3]. Segmental unit input HMM [1] is widely used for its effectiveness and tractability. In segmental unit input HMM, a feature vector is derived from several successive frames. The immediate use of several successive frames inevitably increases the dimensionality of parameters. Therefore, a dimensionality reduction method is performed to spliced frames.

Linear discriminant analysis (LDA) [4,5] is widely used for this purpose and a powerful tool to preserve discriminative information. LDA assumes each class has the same class covariance [6]. However, this assumption does not necessarily hold for a real data set. In order to overcome this limitation, several methods have been proposed. Kumar et al. incorporated the maximum likelihood estimation as an objective function to estimate parameters for different Gaussians with unequal covariances [7]. Saon et al. proposed another objective function similar to Kumar's and showed its relationship with a constrained maximum likelihood estimation [8]. Both Kumar's and Saon's heteroscedastic extensions are called heteroscedastic discriminant analysis (HDA). The effectiveness of these methods for some data sets has been experimentally shown. However, it is difficult to find one particular criterion suitable for any kind of data set.

In this paper, we focus on LDA and Saon's HDA, and give a new interpretation of them. Then, we propose a new framework which we call *power linear discriminant analysis* (PLDA). PLDA can describe various criteria including LDA and HDA with one parameter. Experimental results show the effectiveness for two data sets which collected using a close-talking microphone and a hands-free microphone.

The paper is organized as follows: Classical LDA and HDA are reviewed in Section 2. Then, a new framework of PLDA is proposed in Section 3. Experimental results are presented in Section 4. Finally, conclusions and future work are given in Section 5.

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].

$$
\begin{aligned}
&P(\mathbf{o}_1, \cdots, \mathbf{o}_T) \\
&= \sum_{\mathbf{q}} \prod_i P(\mathbf{o}_i \mid \mathbf{o}_1, \cdots, \mathbf{o}_{i-1}, q_1, \cdots, q_i) \\
&\qquad \times P(q_i \mid q_1, \cdots, q_{i-1}) \qquad (1) \\
&\approx \sum_{\mathbf{q}} \prod_i P(\mathbf{o}_i \mid \mathbf{o}_{i-(d-1)}, \cdots, \mathbf{o}_{i-1}, q_i) P(q_i \mid q_{i-1}) \qquad (2) \\
&\approx \sum_{\mathbf{q}} \prod_i P(\mathbf{o}_{i-(d-1)}, \cdots, \mathbf{o}_i \mid q_i) P(q_i \mid q_{i-1}), \qquad (3)
\end{aligned}
$$

where T denotes the length of input sequence and d denotes the number of successive frames. The immediate use of several successive frames as an input vector inevitably increases the dimension of parameters. Then, PCA, LDA, or HDA was used to reduce dimensionality [1, 3, 8].

Here, we briefly review LDA and HDA. In addition, we investigate the effectiveness of LDA and HDA for some artificial data sets.

2.1. Linear Discriminant Analysis

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

Within-class and between-class covariance matrices are defined

as follows [4, 5]:

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

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

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

In LDA, the objective function is defined as follows:

$$J_{LDA}\left(\mathbf{B}\right) = \frac{\left|\mathbf{B}^T \boldsymbol{\Sigma}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \boldsymbol{\Sigma}_w \mathbf{B}\right|}. \tag{6}$$

LDA finds a transformation matrix $\mathbf{B}$ that maximizes Eq. (6).

2.2. Heteroscedastic Discriminant Analysis

LDA is not the optimal transform when the class distributions are heteroscedastic. Kumar et al. incorporated the maximum likelihood estimation of parameters for differently distributed Gaussians [7]. Saon et al. proposed another objective function similar to Kumar's and showed its relationship with a constrained maximum likelihood estimation [8]. Both Kumar's and Saon's heteroscedastic extensions are called *heteroscedastic discriminant analysis* (HDA).

In this paper, we focus on Saon's HDA objective function:

$$J_{HDA}\left(\mathbf{B}\right) = \prod_{k=1}^{c}\left(\frac{\left|\mathbf{B}^T \boldsymbol{\Sigma}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \boldsymbol{\Sigma}_k \mathbf{B}\right|}\right)^{N_k}. \tag{7}$$

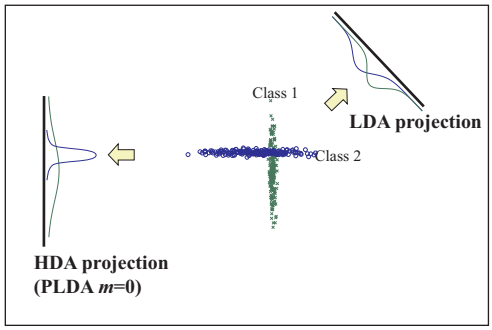
The solution to maximize Eq.(7) is not analytically obtained. Therefore, its maximization is performed using a numerical optimization technique.

2.3. Dependency on data set

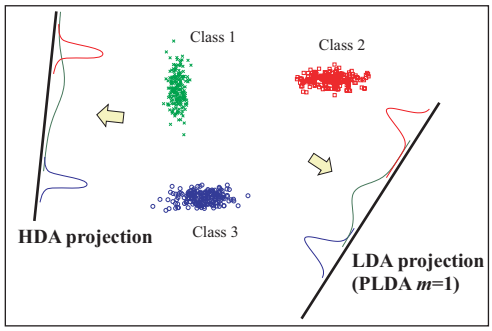
In Figure 1, two-dimensional two- or three-class data features are projected onto a one-dimensional subspace by LDA and HDA. Figure 1(a) shows that HDA has higher separability than LDA for the data set used in [8]. On the other hand, as shown in Figure 1(b) , LDA has higher separability than HDA for 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

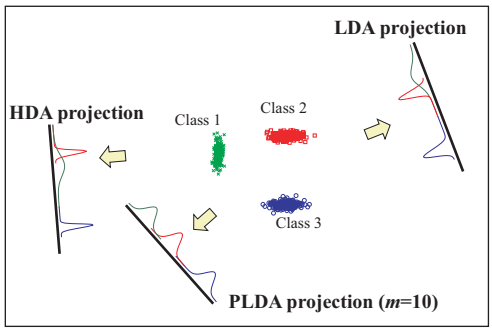
As shown above, it is difficult to separate appropriately every data set with one particular criterion such as LDA and HDA. Here, we concentrate on providing a framework which integrates various criteria.



(a)



(b)



(c)

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

3.1. Relationship between LDA and HDA

From a different viewpoint, LDA and HDA objective functions can be rewritten as

$$J_{LDA}\left(\mathbf{B}\right) = \frac{\left|\mathbf{B}^T \boldsymbol{\Sigma}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \boldsymbol{\Sigma}_w \mathbf{B}\right|} = \frac{\left|\tilde{\boldsymbol{\Sigma}}_b\right|}{\left|\sum_{k=1}^{c} P_k \tilde{\boldsymbol{\Sigma}}_k\right|}, \tag{8}$$

$$J_{HDA}\left(\mathbf{B}\right) = \prod_{k=1}^{c}\left(\frac{\left|\mathbf{B}^T \boldsymbol{\Sigma}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \boldsymbol{\Sigma}_k \mathbf{B}\right|}\right)^{N_k} \propto \frac{\left|\tilde{\boldsymbol{\Sigma}}_b\right|}{\left|\prod_{k=1}^{c} \tilde{\boldsymbol{\Sigma}}_k^{P_k}\right|}, \tag{9}$$

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

Both numerators denote determinants of the between-class covariance matrix. In Eq. (8), the denominator can be viewed as a

determinant of *the weighted arithmetic mean* of the class covariance matrices. Similarly, in Eq. (9), 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.

3.2. Power Linear Discriminant Analysis

As described above, Eqs. (8) and (9) give us a new integrated interpretation of LDA and HDA. As 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* [9]. We extend this notion to a determinant of a matrix mean and propose a new objective function as follows:

$$J_{PLDA}\left(\mathbf{B}, m\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_b\right|}{\left|\left(\sum_{k=1}^{c} P_k \tilde{\mathbf{\Sigma}}_k^m\right)^{1/m}\right|}, \tag{10}$$

where *m* denotes a control parameter. Intuitively, as *m* becomes larger, the classes with larger variances become dominant in the denominator of Eq. (10). Contrarily, as *m* becomes smaller, the classes with smaller variances become dominant. Thus, varying a parameter *m*, the proposed objective function can represent various objective ones. Some typical objective functions are enumerated below.

- $m = 2$ (root mean square)

$$J_{PLDA}\left(\mathbf{B}, 2\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_b\right|}{\left|\left(\sum_{k=1}^{c} P_k \tilde{\mathbf{\Sigma}}_k^2\right)^{1/2}\right|}.$$

- $m = 1$ (arithmetic mean)

$$J_{PLDA}\left(\mathbf{B}, 1\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_b\right|}{\left|\sum_{k=1}^{c} P_k \tilde{\mathbf{\Sigma}}_k\right|} = J_{LDA}\left(\mathbf{B}\right).$$

- $m = 0$ (geometric mean)

$$J_{PLDA}\left(\mathbf{B}, 0\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_b\right|}{\left|\prod_{k=1}^{c} \tilde{\mathbf{\Sigma}}_k^{P_k}\right|} \propto J_{HDA}\left(\mathbf{B}\right).$$

- $m = -1$ (harmonic mean)

$$J_{PLDA}\left(\mathbf{B}, -1\right) = \frac{\left|\tilde{\mathbf{\Sigma}}_b\right|}{\left|\left(\sum_{k=1}^{c} P_k \tilde{\mathbf{\Sigma}}_k^{-1}\right)^{-1}\right|}.$$

We call this new discriminant analysis formulation *Power Linear Discriminant Analysis* (PLDA). Figure 1(c) shows that PLDA can have a higher separability for the 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 [10] and SANN method [11]. These methods need no derivatives of the objective function. However, it is known that these methods converge slowly. In some special cases, the derivatives of the objective function are derived. Hence, we can use some fast convergence methods, such as the quasi-Newton method and conjugate gradient method [12].

3.2.1. 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:

$$\frac{\partial}{\partial \mathbf{B}} \log J_{PLDA}\left(\mathbf{B}, m\right) = 2\mathbf{\Sigma}_b \mathbf{B} \tilde{\mathbf{\Sigma}}_b^{-1} - 2\mathbf{D}_m, \tag{11}$$

where

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

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

and

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

3.2.2. $\tilde{\mathbf{\Sigma}}_k$ constrained to be diagonal

Because of computational simplicity, the covariance matrix in the class *k* is often assumed to be diagonal [7,8]. Since a diagonal matrix multiplication is commutative, the derivatives of the PLDA objective function are simplified as follows:

$$\frac{\partial}{\partial \mathbf{B}} \log J_{PLDA}\left(\mathbf{B}, m\right) = 2\mathbf{\Sigma}_b \mathbf{B} \tilde{\mathbf{\Sigma}}_b^{-1} - 2\left(\sum_{k=1}^{c} P_k \mathbf{\Sigma}_k \mathbf{B} diag\left(\tilde{\mathbf{\Sigma}}_k\right)^{m-1}\right)\left(\sum_{k=1}^{c} P_k diag\left(\tilde{\mathbf{\Sigma}}_k\right)^m\right)^{-1}, \tag{12}$$

where *diag* is an operator which sets zero to off-diagonal elements.

When *m* is equal to zero, the PLDA objective function corresponds to the diagonal HDA (DHDA) objective function introduced in [8]. When *m* is equal to one, the PLDA objective function can be viewed as a diagonal LDA (DLDA) similar to DHDA. Note that DLDA no longer has the global optimum unlike LDA.

4. EXPERIMENTS

We conducted the experiments using the CENSREC-3 database [13]. The CENSREC-3 is designed as an evaluation framework of Japanese isolated word recognition in real driving car environments. Speech data was 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 and six kinds of in-car environments. We only used three kinds of vehicle speeds in normal in-car environment for evaluation. A total of 2,646 utterances spoken

by 18 speakers (8 males and 10 females) were evaluated for each microphone. The speech signals for training and evaluation were both sampled at 16 kHz.

4.1. Baseline System

In the CENSREC-3, the baseline scripts are designed to facilitate HMM training and evaluation by HTK [14]. 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 250Hz. The decoding process was performed without any language model. The vocabulary size of the CENSREC-3 was 50 words. Fifty similar-sounding out-of-vocabulary words were added for the experiments.

4.2. Dimensionality Reduction Procedure

The dimensionality reduction was performed using PCA, (D)LDA, (D)HDA, and PLDA for the spliced features. Eleven successive frames (143 dimensions) were reduced to 39 dimensions. In HDA and PLDA, to optimize Eq. (10), we used the limited-memory BFGS algorithm as a numerical optimization technique [12]. Assuming that projected covariance matrices were diagonal, Eq. (12) was used to compute a gradient. The LDA transformation matrix was used for the initial gradient. To assign one of the classes to every feature after dimensionality reduction, HMM state labels were generated for the training data by state-level forced alignment algorithm using a well-trained HMM system. The class number was 43 corresponding to the number of the monophones.

Table 1. Word error rates (%) by PLDA and conventional methods.

Method	m	CT	HF	Overall
MFCC + Δ + $\Delta\Delta$	–	7.45	15.04	11.24
PCA	–	10.58	19.39	14.98
LDA	–	8.78	15.80	12.28
HDA	–	7.94	17.16	12.55
PLDA	-3.0	6.73	15.04	10.88
PLDA	-2.0	7.29	12.32	9.81
PLDA	-1.5	6.27	**10.70**	**8.48**
PLDA	-1.0	6.92	11.49	9.20
PLDA	-0.5	**6.12**	12.51	9.32
DHDA	(0.0)	7.41	14.17	10.79
PLDA	0.5	7.29	13.53	10.41
DLDA	(1.0)	9.33	16.97	13.15
PLDA	1.5	8.96	17.31	13.13
PLDA	2.0	8.58	15.91	12.24
PLDA	3.0	9.41	16.36	12.89

4.3. Experimental Results

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 ($m = -1.5$). Thus, these two data sets recorded with different microphones have different optimal control parameters. PLDA with the optimal control parameters consistently outperform the other methods.

5. CONCLUSIONS

In this paper we propose a new framework for integrating various criteria to reduce dimensionality. The new framework which we call power linear discriminant analysis (PLDA) includes LDA and Saon's HDA criteria as special cases. The experimental results on the CENSREC-3 database show that segmental unit input HMM with PLDA gives better performance than the others. Future work includes choosing the parameter m automatically to get optimal performance, combining PLDA with MLLT [15], and comparing PLDA+ MLLT with both LDA+MLLT and HDA+MLLT.

6. ACKNOWLEDGMENT

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

7. 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 and Signal Processing*, 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] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, New York, second edition, 1990.

[5] R. O. Duda, P. B. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, New York, 2001.

[6] N. A. Campbell, "Canonical variate analysis – a general model formulation," *Australian Journal of Statistics*, pp. 86–96, 1984.

[7] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, pp. 283–297, 1998.

[8] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," *Proc. ICASSP*, pp. 129–132, 2000.

[9] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley and Sons, 1999.

[10] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[11] C. J. P. Belisle, "Convergence theorems for a class of simulated annealing algorithms," *Journal of Applied Probability*, vol. 29, pp. 885–892, 1992.

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, 1999.

[13] M. Fujimoto, K. Takeda, and S. Nakamura, "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, pp. 2783–2793, 2006.

[14] *HTK Web site*, `http://htk.eng.cam.ac.uk/`.

[15] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," *Proc. ICASSP*, 1998.