Selection of optimal dimensionality reduction method using Chernoff bound for segmental unit input HMM

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

[1] DENSO CORPORATION, Nisshin 470-0111, Japan
[2] Nagoya University, Nagoya 464-8603, Japan
[3] Toyohashi University of Technology, Toyohashi 441-8580, Japan
msakai@rlab.denso.co.jp, kitaoka@nagoya-u.jp, nakagawa@slp.ics.tut.ac.jp

10.21437/Interspeech.2007-364

Abstract

To precisely model the time dependency of features, segmental unit input HMM with a dimensionality reduction method has been widely used for speech recognition. Linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA) are popular approaches to reduce the dimensionality. We have proposed another dimensionality reduction method called power linear discriminant analysis (PLDA) to select the best dimensionality reduction method that yields the highest recognition performance. This selection process on the basis of trial and error requires much time to train HMMs and to test the recognition performance for each dimensionality reduction method.

In this paper we propose a performance comparison method without training or testing. We show that the proposed method using the Chernoff bound can rapidly and accurately evaluate the relative recognition performance.

Index Terms: speech recognitoin, 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 sequences. In order to overcome this limitation, many extensions have been proposed [1–3]. Segmental unit input HMM [1] has been 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 which inevitably increases the dimension of the parameters. Therefore, a dimensionality reduction method is performed to spliced frames.

Linear discriminant analysis (LDA) [4,5] and heteroscedastic discriminant analysis (HDA) [6, 7] are used for this purpose. In addition, we have proposed a new framework which we call power linear discriminant analysis (PLDA) [8], which can describe various criteria including LDA and HDA with one control parameter. The effectiveness of these methods has been experimentally shown. Unfortunately, we cannot know which method is the most effective before training HMMs and testing the performances of all dimensionality reduction methods on the evaluation set. In general, this training and testing process requires more than several dozen hours. Moreover, the computational time is proportional to the number of dimensionality reduction methods. PLDA, especially, requires considerable time to compare several conditions because its control parameter can be set to a real number.

In this paper, we propose a performance comparison method without training of HMMs and test on an evaluation set. To evaluate the relative performance among a number of methods, we focus on a class separability error of projected features and measure it on evaluation data. We show that the proposed method can rapidly and accurately compare with the relative recognition performance.

The paper is organized as follows: Segmental unit input HMM with dimensionality reduction method is reviewed in Section 2. Then, a comparison method of the relative recognition performance is proposed in Section 3. Experimental results are presented in Section 4. Finally, conclusions 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]:

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

where T denotes the length of input sequence and d 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, HDA, or PLDA were used to reduce dimensionality [1, 3, 7, 8].

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.

In LDA, to obtain an optimal projection matrix $\mathbf{B}$, the objective function is defined as follows:

$$
J_{LDA}(\mathbf{B}) = \frac{\left|\mathbf{B}^T \boldsymbol{\Sigma}_b \mathbf{B}\right|}{\left|\sum_{k=1}^{c} P_k \mathbf{B}^T \boldsymbol{\Sigma}_k \mathbf{B}\right|}, \qquad (4)
$$

where $\mathbf{\Sigma}_b$ denotes a between-class covariance matrix, $\mathbf{\Sigma}_k$ the covariance matrix in the class k, P_k the class weight, and c the number of classes, respectively. LDA finds a projection matrix $\mathbf{B}$ that maximizes Eq. (4).

2.2. Heteroscedastic discriminant analysis

In HDA [7], the objective function is defined as follows:

$$J_{HDA}\left(\mathbf{B}\right) = \prod_{k=1}^{c} \left(\frac{\left|\mathbf{B}^T \mathbf{\Sigma}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \mathbf{\Sigma}_k \mathbf{B}\right|} \right)^{N_k}, \tag{5}$$

where N_k denotes the number of features labeled as class k. The solution to maximize Eq. (5) is not analytically obtained. Therefore, its maximization is performed using a numerical optimization technique.

2.3. Power linear discriminant analysis

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

$$J_{PLDA}\left(\mathbf{B}, m\right) = \frac{\left|\mathbf{B}^T \mathbf{\Sigma}_b \mathbf{B}\right|}{\left| \left(\sum_{k=1}^{c} P_k (\mathbf{B}^T \mathbf{\Sigma}_k \mathbf{B})^m \right)^{1/m} \right|}, \tag{6}$$

where *m* denotes a control parameter. We have referred to it as *Power Linear Discriminant Analysis* (PLDA). Intuitively, as *m* becomes larger, the classes with larger variances become dominant in the denominator of Eq. (6). Conversely, as *m* becomes smaller, the classes with smaller variances become dominant. Thus, by varying the control parameter *m*, the proposed objective function can represent various objective functions. Some typical objective functions are enumerated below.

- $m = 1$

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

- $m = 0$

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

where $\tilde{\mathbf{\Sigma}}_b = \mathbf{B}^T \mathbf{\Sigma}_b \mathbf{B}$ and $\tilde{\mathbf{\Sigma}}_k = \mathbf{B}^T \mathbf{\Sigma}_k \mathbf{B}$. To maximize the PLDA objective function with respect to $\mathbf{B}$, we can use some numerical optimization methods, such as the quasi-Newton method and conjugate gradient method [9].

2.4. Problem in selection of optimal dimensionality reduction method

As shown above, several methods to reduce dimensionality have been proposed, and their effectiveness has been experimentally demonstrated. Unfortunately, we cannot know which method is the most effective before training HMMs and testing the performances of all dimensionality reduction methods on the evaluation set. In general, this training and testing process requires more than several dozen hours. Moreover, the computational time is proportional to the number of dimensionality reduction methods. PLDA, especially, requires much time to compare a number of conditions because it is able to choose a control parameter within a real number.

3. Performance comparison method

In this section we focus on a class separability error of the features in the projected space instead of using a recognition error. 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.

3.1. Two-class problem

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

$$\varepsilon = \int \min[P_1 p_1(\mathbf{x}), P_2 p_2(\mathbf{x})] d\mathbf{x}, \tag{7}$$

where P_i denotes a prior probability of the class i and $p_i(\mathbf{x})$ is a conditional density function of the class i. The Bayes error ε can represent a classification error, assuming that the training data and the 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 [4]:

$$\varepsilon_u^{1,2} = P_1^s P_2^{1-s} \int p_1^s(\mathbf{x}) p_2^{1-s}(\mathbf{x}) d\mathbf{x} \quad \text{for} \quad 0 \le s \le 1 \tag{8}$$

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

$$\varepsilon_u^{1,2} = P_1^s P_2^{1-s} \exp(-\eta^{1,2}(s)), \tag{9}$$

where

$$\eta^{1,2}(s) = \frac{s(1-s)}{2} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \left(s\mathbf{\Sigma}_1 + (1-s)\mathbf{\Sigma}_2\right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} \ln \frac{\left|s\mathbf{\Sigma}_1 + (1-s)\mathbf{\Sigma}_2\right|}{\left|\mathbf{\Sigma}_1\right|^s \left|\mathbf{\Sigma}_2\right|^{1-s}}. \tag{10}$$

In this case, ε_u can be obtained analytically and calculated rapidly.

In Figure 1, two-dimensional two-class data are projected onto a one-dimensional subspace 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.

3.2. 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:

$$\tilde{\varepsilon}_u = \sum_{i=1}^{c} \sum_{j=1}^{c} I(i,j) \varepsilon_u^{i,j} \tag{11}$$

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

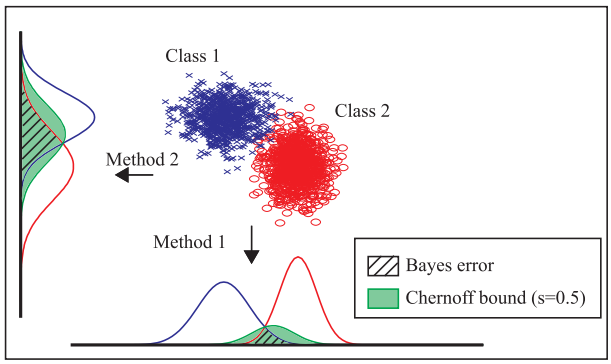

Figure 1: Examples of dimensionality reduction.

3.2.1. *Sum of pairwise approximated errors*

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

$$I(i,j) = \begin{cases} 1, & \text{if } j > i, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

3.2.2. *Maximum pairwise approximated error*

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

$$I(i,j) = \begin{cases} 1, & \text{if } j > i \text{ and } (i,j) = (\hat{i},\hat{j}), \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

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

3.2.3. *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:

$$I(i,j) = \begin{cases} 1, & \text{if } j = \hat{j}_i, \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

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

4. Experiments

To evaluate the effectiveness of the comparison method of the relative recognition performance, we first conducted the experiments using the CENSREC-3 database [10]. 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 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 [11]. 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 ms and frame shift was 10 ms. 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 of the CENSREC-3 was 50 words. Another fifty similar-sounding words were added to the vocabulary for the experiments.

4.2. Dimensionality reduction procedure

The dimensionality reduction was performed using PCA, LDA, HDA, DHDA [7], and PLDA for the spliced features. Eleven successive frames (143 dimensions) were reduced to 39 dimensions. In (D)HDA and PLDA, to optimize Eq. (6), we assumed that projected covariance matrices were diagonal and used the limited-memory BFGS algorithm as a numerical optimization technique [9]. 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 a state-level forced alignment algorithm using a well-trained HMM system. The class number was 43 corresponding to the number of the monophones.

4.3. Experimental results

Tables 1 and 2 show the word error rates and class separability errors according to Eqs. (12)-(14) for each dimensionality reduction method. 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 comparing dimensionality reduction methods, we used $s = 1/2$ for the Chernoff bound computation. In the case of $s = 1/2$, Eq. (8) is called the Bhattacharyya bound. Two covariance matrices in Eq. (10) were treated as diagonal because diagonal Gaussians were used to model HMMs. Both Tables 1 and 2 show that the results of the proposed method and relative recognition performance agree well. Eqs. (13) and (14) yield slightly better agreement of the recognition performance than Eq. (12). However, no comparison method among Eqs. (12)-(14) could predict the best dimensionality reduction methods simultaneously for both of the two evaluation sets. It is supposed that this results from neglecting time information of speech feature sequences to measure a class separability error and modeling a class distribution as a unimodal normal distribution.

Table 1: Word error rates (%) and class separability errors according to Eqs. (12)-(14) for the evaluation set with CT microphone. The best results are highlighted in bold.

Method	WER	Eq. (12)	Eq. (13)	Eq. (14)
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 ($m = -3$)	6.73	2.02	0.0319	0.531
PLDA ($m = -2$)	7.29	2.07	0.0316	0.532
PLDA ($m = -1.5$)	6.27	**1.97**	0.0307	0.523
PLDA ($m = -1$)	6.92	1.99	0.0301	**0.521**
PLDA ($m = -0.5$)	**6.12**	2.01	**0.0292**	0.525
DHDA (PLDA m=0)	7.41	2.15	0.0296	0.541
PLDA ($m = 0.5$)	7.29	2.41	0.0306	0.560
PLDA ($m = 1$)	9.33	3.09	0.0354	0.641
PLDA ($m = 1.5$)	8.96	4.61	0.0394	0.742
PLDA ($m = 2$)	8.58	4.65	0.0404	0.745
PLDA ($m = 3$)	9.41	4.73	0.0413	0.756

Table 2: Word error rates (%) and class separability errors according to Eqs. (12)-(14) for the evaluation set with HF microphone. The best results are highlighted in bold.

Method	WER	Eq. (12)	Eq. (13)	Eq. (14)
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 ($m = -3$)	15.04	2.19	0.0338	0.600
PLDA ($m = -2$)	12.32	2.26	0.0339	0.602
PLDA ($m = -1.5$)	**10.70**	**2.18**	0.0332	**0.5921**
PLDA ($m = -1$)	11.49	2.23	**0.0327**	0.5922
PLDA ($m = -0.5$)	12.51	2.31	0.0329	0.598
DHDA (PLDA m=0)	14.17	2.50	0.0331	0.619
PLDA ($m = 0.5$)	13.53	2.81	0.0341	0.644
PLDA ($m = 1$)	16.97	3.38	0.0370	0.711
PLDA ($m = 1.5$)	17.31	5.13	0.0403	0.828
PLDA ($m = 2$)	15.91	5.22	0.0412	0.835
PLDA ($m = 3$)	16.36	5.36	0.0424	0.850

4.4. Computational costs

The computational costs for the evaluation of recognition performance versus the proposed comparison method is shown in Table 3, for which the experiment is done with a Pentium IV 2.8 GHz computer. For every dimensionality reduction method, the evaluation of recognition performance required 15 hours for training of HMMs and 5 hours for testing on an evaluation set. In total, 300 hours were needed for comparing 15 dimensionality reduction methods (MFCC+Δ+$\Delta\Delta$, PCA, LDA, HDA, and PLDAs using 11 different control parameters). On the other hand, the proposed comparison method 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 needed to calculate Eqs. (12)-(14) for each dimensionality reduction method. In total, only one hour was needed for predicting the optimal method among the 15 dimensionality reduction methods described above. Thus, the proposed method could perform the prediction process up to 2 orders faster than a conventional method that included training of HMMs and test on an evaluation set.

Table 3: Computational costs with the conventional and proposed method.

conventional	proposed
300 hours	1 hour

5. Conclusions

In this paper we proposed a new method to compare dimensionality reduction methods and to select the best one. The proposed method used the Chernoff bound as a measure of a class separability error which was an upper bound of the Bayes error. Experimental results showed that the proposed method could evaluate the relative recognition performance without training of HMMs and test on an evaluation set. In addition, the proposed method yielded accurate performance comparison with a drastic reduction of computational costs.

6. Acknowledgment

The presented study was conducted using the CENSREC-3 database developed by the 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. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, pp. 283–297, 1998.

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

[8] M. Sakai, N. Kitaoka, and S. Nakagawa, "Generalization of linear discriminant analysis used in segmental unit input HMM for speech recognition," *Proc. ICASSP*, pp. 333–336, 2007.

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

[10] 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.

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