LETTER

Acoustic Feature Transformation Combining Average and Maximum Classification Error Minimization Criteria

Makoto SAKAI[†,††a)], **Norihide KITAOKA**[††], *and* **Kazuya TAKEDA**[††], *Members*

SUMMARY Acoustic feature transformation is widely used to reduce dimensionality and improve speech recognition performance. In this letter we focus on dimensionality reduction methods that minimize the average classification error. Unfortunately, minimization of the average classification error may cause considerable overlaps between distributions of some classes. To mitigate risks of considerable overlaps, we propose a dimensionality reduction method that minimizes the maximum classification error. We also propose two interpolated methods that can describe the average and maximum classification errors. Experimental results show that these proposed methods improve speech recognition performance.

key words: *speech recognition, dimensionality reduction, Bayes error*

1. Introduction

Using acoustic dynamic information that expresses temporal change in speech signals would improve speech recognition performance because temporal change is not adequately described by a hidden Markov model-based speech recognition system. Several methods for integrating dynamic information have been proposed [1], [2]. One popular approach is to compute first- and second-order differences of successive features [1]. It is well known that this approach can improve speech recognition performance. Another approach for integrating dynamic information is to concatenate several successive features into a single high-dimensional feature vector. Then, a feature transformation method is applied to the vector to reduce dimensionality without losing discriminative information. The latter approach includes the former one as a special case. In this letter the latter approach is investigated. Especially, we focus on dimensionality reduction methods that minimize misclassification in the sense of the Bayes classification error [3]–[5], while the former approach does not take the minimization of misclassification into account directly. We show that the purpose of the existing methods can be regarded as minimization of the average classification error (AveCE) among classes. While minimizing the AveCE suppresses total classification error, it cannot avoid the occurrence of considerable overlaps between distributions of some classes. Therefore, there may be class pairs that have little or no discriminative information on each other. Hence, the AveCE does not necessarily find a suitable projection for speech recognition. To avoid this, we propose an alternative dimensionality reduction method that minimizes the maximum classification error (MaxCE) among all class pairs. The proposed method can avoid considerable error between classes. Moreover, we propose interpolated methods including AveCE and MaxCE.

Manuscript received November 11, 2009.
Manuscript revised February 19, 2010.

[†]The author is with DENSO CORPORATION, Nisshin-shi, 470–0111 Japan.
[††]The authors are with Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: msakai@rlab.denso.co.jp

DOI: 10.1587/transinf.E93.D.2005

2. Minimization of Approximated Bayes Error

In this letter we focus on a minimization criterion of an approximation of the Bayes error [3], [5].

2.1 Bayes Error

Let us consider the discrimination problem of classifying an observation as coming from one of K possible classes $k \in \{1, 2, \cdots, K\}$. And, let $\mathbf{x}$ be an n-dimensional feature vector such as a concatenated speech frame. The error probability P_e of the optimal Bayes rule for the classification into K classes becomes [6], [7]

$$P_e = 1 - \int \max_k [\lambda_k p_k(\mathbf{x})]\, d\mathbf{x},$$

where λ_k and p_k denote a prior probability and a probability density function (pdf) for class k, respectively. We assume that the λ_k and p_k for $k = 1, \cdots, K$ are entirely known.

The number of the dimension of a feature vector $\mathbf{x}$ can be reduced to $p < n$ by a transformation $\mathbf{z} = \mathbf{B}^\top \mathbf{x}$ with a transformation matrix $\mathbf{B} \in \mathbb{R}^{n \times p}$ of rank p, where $\mathbf{B}^\top$ is the transpose of the matrix $\mathbf{B}$. Then, the error probability in the range space of $\mathbf{B}^\top$, $P_e^{\mathbf{B}}$, becomes:

$$P_e^{\mathbf{B}} = 1 - \int \max_k \left[\lambda_k p_k^{\mathbf{B}}(\mathbf{z})\right] d\mathbf{z},$$

where $p_k^{\mathbf{B}}$ denotes the pdf for class k in the projected space spanned by the column vectors of $\mathbf{B}$. Since the transformation $\mathbf{z} = \mathbf{B}^\top \mathbf{x}$ produces a linear combination of the components of the feature vector $\mathbf{x}$, discriminative information is generally lost and $P_e^{\mathbf{B}} \geq P_e$ [4].

The feature transformation problem could be stated as a selection of an n by p matrix $\hat{\mathbf{B}}$ from all n by p matrices of rank p such that

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} P_e^{\mathbf{B}}. \tag{1}$$

Unfortunately, it is generally difficult to calculate $P_e^{\mathbf{B}}$ directly.

2.2 Other Criteria for Estimating Error Probability

Instead of minimizing $P_e^{\mathbf{B}}$ directly, the following affinity between two pdfs have been often used:

$$\rho_{i,j} = \int \sqrt{p_i(\mathbf{x})p_j(\mathbf{x})}d\mathbf{x}. \tag{2}$$

$\rho_{i,j}$ is called the Bhattacharyya coefficient and is an upper bound on the Bayes error [3]. This coefficient can be regarded as a classification error between two pdfs. Clearly, $\rho_{i,j}$ lies between zero and one.

The Bhattacharyya coefficient in the range space of $\mathbf{B}^\top$ becomes:

$$\rho_{i,j}^{\mathbf{B}} \equiv \int \sqrt{p_i^{\mathbf{B}}(\mathbf{z})p_j^{\mathbf{B}}(\mathbf{z})}d\mathbf{z}. \tag{3}$$

If we assume that the p_k is a Gaussian distribution with a mean vector $\boldsymbol{\mu}_k$ and a covariance matrix $\mathbf{C}_k$, Eq. (3) has the closed form expression:

$$\rho_{i,j}^{\mathbf{B}} = \exp(-\eta_{i,j}^{\mathbf{B}}) \tag{4}$$

where

$$\eta_{i,j}^{\mathbf{B}} = \frac{1}{8}tr\left(\left(\mathbf{B}^\top\mathbf{C}_{ij}\mathbf{B}\right)^{-1}\mathbf{B}^\top\mathbf{M}_{ij}\mathbf{B}\right) + \frac{1}{2}\log\frac{\left|\mathbf{B}^\top\mathbf{C}_{ij}\mathbf{B}\right|}{\sqrt{\left|\mathbf{B}^\top\mathbf{C}_i\mathbf{B}\right|\left|\mathbf{B}^\top\mathbf{C}_j\mathbf{B}\right|}}, \tag{5}$$

$\mathbf{C}_{ij} = \frac{\mathbf{C}_i+\mathbf{C}_j}{2}$, and $\mathbf{M}_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top$. $\eta_{ij}^{\mathbf{B}}$ is called the Bhattacharyya distance.

Several extensions of Eq. (2) to handle multi-class problems have been proposed. Here, we briefly review two techniques.

2.2.1 Upper Bound on Bayes Error

The Bayes error is bounded from above by the following expression [5], [8]:

$$\sum_{i,j>i}\sqrt{\lambda_i\lambda_j}\rho_{i,j}. \tag{6}$$

Saon et al. [5] proposed the following objective function based on Eq. (6):

$$J_{bound}(\mathbf{B}) = \sum_{i,j>i}\sqrt{\lambda_i\lambda_j}\rho_{i,j}^{\mathbf{B}}. \tag{7}$$

2.2.2 Average Bhattacharyya Coefficient

Another natural extension to treat multi-class problems is the average Bhattacharyya coefficient as follows [3]:

$$\sum_{i,j}\lambda_i\lambda_j\rho_{i,j} \tag{8}$$

Based on the average Bhattacharyya coefficient, we can define the following objective function to reduce dimensionality:

$$J_{ave}(\mathbf{B}) = \sum_{i,j}\lambda_i\lambda_j\rho_{i,j}^{\mathbf{B}}. \tag{9}$$

3. Issue about Existing Methods

From $\rho_{i,i}^{\mathbf{B}} = 1$, $\rho_{i,j}^{\mathbf{B}} = \rho_{j,i}^{\mathbf{B}}$, and $\sum_i \lambda_i = 1$, we have

$$\sum_{i,j>i}\sqrt{\lambda_i\lambda_j}\rho_{i,j}^{\mathbf{B}} = \frac{1}{2}\left(\sum_{i,j}\sqrt{\lambda_i\lambda_j}\rho_{i,j}^{\mathbf{B}} - 1\right). \tag{10}$$

Using this, Eq. (7) can be rewritten as follows:

$$\begin{aligned} J_{bound}(\mathbf{B}) &\propto \sum_{i,j}\sqrt{\lambda_i\lambda_j}\rho_{i,j}^{\mathbf{B}} \\ &\propto \sum_{i,j}\frac{\sqrt{\lambda_i}}{Z}\frac{\sqrt{\lambda_j}}{Z}\rho_{i,j}^{\mathbf{B}} \\ &= \sum_{i,j}\lambda_i'\lambda_j'\rho_{i,j}^{\mathbf{B}}, \end{aligned} \tag{11}$$

where $Z \equiv \sum_k \sqrt{\lambda_k}$ is a normalizing constant, and $\lambda_k' \equiv \sqrt{\lambda_k}/Z$. Equations (9) and (11) are essentially the same objective function, and the only difference between them is their priors. Hence, both functions can be regarded as the average of Bhattacharyya coefficient $\rho_{i,j}^{\mathbf{B}}$. That is, both objective functions search for a projection matrix $\mathbf{B}$ so that the average classification error (AveCE) is minimized. Although minimizing the AveCE suppresses total classification error among classes, it cannot avoid the occurrence of considerable overlaps between distributions of some classes, which is critical for speech recognition because there may be class pairs that have little or no discriminative information on each other.

Figure 1 shows that two-dimensional three-class samples are projected onto one-dimensional subspace. Each class sample is synthetic data drawn from different Gaussians. The priors of classes 1 to 3 were 0.75, 0.125 and

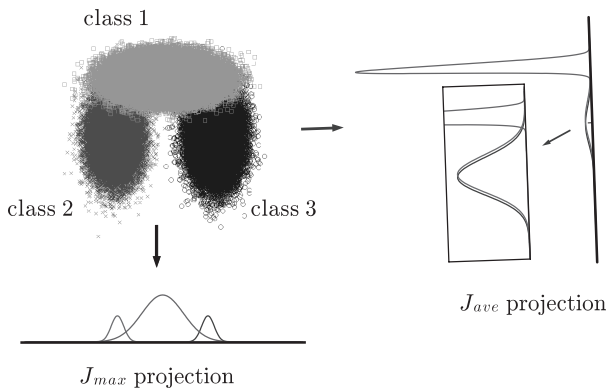

Fig. 1 Example of a synthetic data set comprising three classes. Two lines are the one-dimensional subspaces. The vertical line and the horizontal line are obtained using Eqs. (11) and (12), respectively.

0.125, respectively. The projection by J_{ave} gave high separabilities between classes 1 and 2, and between classes 1 and 3. On the other hand, there was a considerable overlap between classes 2 and 3. Here, let us regard the situation in Fig. 1 as a phone classification task. Suppose that classes 1 to 3 represent some phones (ex. /sil/, /a/, /o/, etc.). When we transform features by J_{ave}, classification becomes difficult between two phones associated with classes 2 and 3.

4. Minimization of Maximum Bhattacharyya Coefficient

To overcome the drawback of the AveCE described in the previous section, we propose an alternative objective function that minimizes the maximum classification error (MaxCE) among all class pairs. The proposed objective function can avoid considerable error between classes. Moreover, we propose generalized objective functions including two criteria.

4.1 Approximated Maximum Classification Error

To prevent less discrimination power of some class pairs, we define the alternative objective function that minimizes the maximum overlap among classes regardless of their priors, instead of AveCE, as follows:

$$J_{max}(\mathbf{B}) \equiv \max_{i,j} \rho_{i,j}^{\mathbf{B}}. \tag{12}$$

Unfortunately, minimization of Eq. (12) with respect to $\mathbf{B}$ is generally difficult. Instead, we approximate Eq. (12). Let $\mathbf{y}$ be an $n \times 1$ vector with positive components $\{y_i\}_{i=1}^n$, and let $\boldsymbol{\alpha}$ be an $n \times 1$ vector of positive weights $\{\alpha_i\}_{i=1}^n$, so that $0 < \alpha_i < 1$ and $\sum_{i=1}^n \alpha_i = 1$. To approximate Eq. (12), we focus on the generalized mean, also known as the weighted mean of order m. The generalized mean is given by [9]:

$$M(\mathbf{y}, \boldsymbol{\alpha}, m) = \left(\sum_{i=1}^{n} \alpha_i y_i^m \right)^{1/m}, \tag{13}$$

for any real m. Equation (13) can describe several means by changing m. For example, Eq. (13) with $m = 1$ corresponds to the arithmetic mean of $\{y_i\}_{i=1}^n$, and Eq. (13) with $m \to 0$ converges to the geometric mean of $\{y_i\}_{i=1}^n$. We especially focus on the following special case of the generalized mean:

$$\lim_{m \to \infty} M(\mathbf{y}, \boldsymbol{\alpha}, m) = \max_i y_i. \tag{14}$$

We approximate Eq. (12) using the generalized mean and sufficiently large value $\hat{m}$ as follows:

$$J_{max}(\mathbf{B}) = \lim_{m \to \infty} \left(\sum_{i,j} \lambda_i \lambda_j \left(\rho_{i,j}^{\mathbf{B}} \right)^m \right)^{1/m} \tag{15}$$

$$\approx \left(\sum_{i,j} \lambda_i \lambda_j \left(\rho_{i,j}^{\mathbf{B}} \right)^{\hat{m}} \right)^{1/\hat{m}}. \tag{16}$$

Equation (16) with $\hat{m} = 100$ was applied in Fig. 1. The result showed that the projection by J_{max} gave higher separability between class 2 and class 3 than that by J_{ave}. That is, J_{max} can offer greatly improved classification power between class 2 and class 3.

4.2 Interpolation between Two Criteria

In Fig. 1, the projection by J_{max} gave a more desirable result than by J_{ave}. However, similar to J_{ave}, J_{max} also does not necessarily find a suitable projection. If a number of class pairs have an overlap comparable to the maximum one, the total error increases significantly. In such a situation, speech recognition performance will deteriorate because most class pairs have only small discrimination power. Therefore, an interpolated criterion that minimizes MaxCE while minimizing AveCE would be effective. Here, we propose two interpolated functions between MaxCE and AveCE.

$$J_{interp1}(\mathbf{B}, \alpha) = (1 - \alpha) J_{ave}(\mathbf{B}) + \alpha J_{max}(\mathbf{B}),$$

$$J_{interp2}(\mathbf{B}, m) = \left(\sum_{i,j} \lambda_i \lambda_j \left(\rho_{i,j}^{\mathbf{B}} \right)^m \right)^{1/m},$$

where α and m denote control parameters so that $\alpha \in [0, 1]$ and $m \geq 1$, respectively. $J_{interp1}$ corresponds to J_{ave} when $\alpha = 0$ and to J_{max} when $\alpha = 1$. From Eq. (9), $J_{interp2}$ corresponds to J_{ave} when $m = 1$. Similarly, from Eq. (15), $J_{interp2}$ converges to J_{max} when $m \to \infty$. As α becomes larger, only one class pair with the maximum overlap between class distributions becomes dominant in $J_{interp1}$. On the other hand, as m becomes larger, several class pairs with large overlaps become dominant in $J_{interp2}$.

5. Experiments

We conducted experiments on a CENSREC-3 database [10], which is designed as an evaluation framework for Japanese isolated word recognition in real in-car environments. For training of HMMs, we used drivers speech of phonetically-balanced sentences recorded under two conditions: while idling and driving on city streets under a normal in-car environment. A total of 14,050 utterances by 293 drivers (202 males and 91 females) were recorded with a close-talking (CT) microphone. For evaluation, we used 2,646 utterances by 18 drivers (8 males and 10 females) recorded under an in-car environment. The speech signals were sampled at 16 kHz. We followed the CENSREC-3 baseline scripts as the evaluation procedure except that fifty similar-sounding words (ex. *aim* for *game* and *tops* for *pops*) were added to the vocabulary. The total vocabulary size became 100. In CENSREC-3, the baseline scripts are designed to facilitate HMM training and evaluation by HTK. The acoustic models consist of triphone HMMs. Each HMM has five states three of which have output distributions. Each distribution is represented with a 32 mixture of diagonal Gaussians. The total number of states with the distributions is

Table 1 Word error rates (WER) (%).

	WER		WER
MFCC + Δ + ΔΔ	6.50	LDA	6.12
J_{ave}	5.85	J_{max}	5.36
$J_{interp1}$ ($\alpha = 0.6$)	4.72	$J_{interp2}$ ($m = 16$)	3.32

Table 2 WER of $J_{interp1}$ vs. α.

α	0	0.2	0.4	0.6	0.8	1.0
WER	5.85	5.78	5.74	4.72	5.10	5.36

Table 3 WER of $J_{interp2}$ vs. m.

m	1	2.5	6	16	30	100
WER	5.85	4.57	4.00	3.32	4.19	5.36

2,000. The baseline performance was evaluated with 39 dimensional feature vectors that consist of 12 MFCCs and log-energy, and their delta and delta-delta coefficients. A delta coefficient was calculated from seven successive frames of MFCCs, and a delta-delta from five successive frames of delta. Consequently, a feature vector was calculated using eleven successive MFCC vectors. The frame length and the frame shift are 20 ms and 10 ms, respectively.

5.1 Feature Transformation Procedure

Eleven successive frames were concatenated into one feature vector (143 dimensions), which is the same number of frames used for calculating delta and delta-delta coefficients. Feature transformation was performed by LDA [7], J_{ave}, J_{max}, $J_{interp1}$ and $J_{interp2}$ for the concatenated features. The concatenated vectors were reduced to 39, which are the same number of dimensions of the baseline feature vectors, and then MLLT [11] was applied. The number of classes was 40.

5.2 Experimental Results

The experimental result is presented in Table 1. Optimal control parameters of $J_{interp1}$ and $J_{interp2}$ were selected experimentally. The result showed that the performance of J_{max} was slightly superior to that of J_{ave}. As mentioned in Sect. 4.2, J_{ave} and J_{max} have complementary characteristics. Both interpolated methods $J_{interp1}$ and $J_{interp2}$ yielded lower error rate than J_{ave} and J_{max} because they could play a complementary role between J_{ave} and J_{max}. Tables 2 and 3 showed WER for different control parameters of $J_{interp1}$ and $J_{interp2}$, where α for $J_{interp1}$ varied between 0 and 1, and m for $J_{interp2}$ varied between 1 and 100. The results indicated that the optimal values of control parameters of $J_{interp1}$ and $J_{interp2}$ were 0.6 and 16 on the CENSREC-3 database, respectively. The results showed that $J_{interp2}$ gave better performance than that of $J_{interp1}$. This is because that $J_{interp2}$ can reduce classification error of several class pairs with large overlaps, as m is a large value, while $J_{interp1}$ reduces that of only one class pair with the maximum overlap between class distributions.

6. Conclusions

To improve speech recognition performance, we propose a dimensionality reduction method that minimizes the maximum classification error, instead of the average classification error. In addition, we also propose interpolated methods that can describe the maximum classification error and the average one. Experimental results show the effectiveness of the proposed methods.

Future work includes choosing the control parameters for interpolated methods to obtain optimal performance.

References

[1] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process., vol.34, no.1, pp.52–59, 1986.
[2] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," Speech Commun., pp.283–297, 1998.
[3] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Commun. Technol., vol.15, no.1, pp.52–60, 1967.
[4] H.P. Decell and J.A. Quirein, "An iterative approach to the feature selection problem," Conf. Machine Processing of Remotely Sensed Data, pp.3B1–3B12, 1973.
[5] G. Saon and M. Padmanabhan, "Minimum Bayes error feature selection for continuous speech recognition," Advances in Neural Information Processing Systems, pp.800–806, 2001.
[6] M. Basseville, "Distance measures for signal processing and pattern recognition," Signal Process., vol.18, no.4, pp.349–369, 1989.
[7] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1990.
[8] D.E. Boekee and J.C.A.V. der Lubbe, "Some aspects of error bounds in feature selection," Pattern Recognit., vol.11, pp.353–360, 1979.
[9] J.R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, John Wiley & Sons, 1999.
[10] 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.
[11] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," Proc. ICASSP, 1998.