USING MINIMUM CLASSIFICATION ERROR TRAINING IN DIMENSIONALITY REDUCTION

Xuechuan Wang and Kuldip K. Paliwal
School of Microelectronic Engineering, Griffith University, QLD 4111, Australia
wang@me.gu.edu.au K.Paliwal@me.gu.edu.au

Abstract: **Dimensionality reduction is an important problem in pattern recognition. In a speech recognition system, the size of the feature set is normally large in the order of 40. Therefore, it is necessary to reduce the dimensionality of the feature space for efficient and effective speech recognition. Two popular methods to reduce the dimensionality of the feature space are Linear Discriminat Analysis (LDA) and Principal Component Analysis (PCA). This paper uses the Minimum Error Classification (MCE) training algorithm for dimensionality reduction and presents an alternative MCE training algorithm that performs better on testing data than the conventional MCE training algorithm. The effects of the initial value of the transformation matrix on the performance of MCE have also been studied.**

1. INTRODUCTION

In a speech recognition system, one tries to use larger feature set to enhance the speech recognition performance. Although the addition of new features does improve the accuracy of speech recognition, not every newly added feature has the same contribution to the improvement of the speech recognizer. Some of them may not contribute at all. The increase in the number of the speech features has caused other problems. For example, the recognizer using higher dimension feature set requires more parameters to characterize the classifier and requires more storage. Thus, it will increase the complexity of computation and make its real-time implementation more difficult and costly. Furthermore, a larger amount of data is needed for training such a recognizer. To avoid these problems, a number of dimensionality reduction algorithms have already been proposed to obtain compact feature set.

The dimensionality reduction methods can be grouped into two categories: feature selection methods and feature extraction methods. The feature selection methods select the features by devising a figure of merit that reflects the goodness of an individual feature in the recognition task. The F-ratio (ratio of between-class and within-class variances) is often used in the feature selection methods [8]. In this paper, we are mainly concerned with the feature extraction methods. Feature extraction methods reduce the dimensionality by projecting the original feature space into a smaller subspace through a transformation. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are two major methods used to extract new features [9, 10, 13, 15]. In practice, LDA performs better than PCA. Recently Minimum Classification Error (MCE) training algorithm has been studied in improving the speech recognition performance [2, 4, 5, 12, 14]. In this paper, we use the MCE algorithm for dimensionality reduction. We study its performance and propose some alternatives to improve its performance.

2. LINEAR DISCRIMINANT ANALYSIS

The basic idea of LDA is to find a transformation that projects the raw data to a low

0-7803-6278-0/00$10.00 (C) IEEE

dimensional space and makes the ratio of between-class variation and within-class variation the large. In Fisher's LDA, the transformation is defined in terms of a set of D-dimensional orthogonal vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ... $\mathbf{u}_d$, where $d < D$ and D is the dimension of the feature space. The first vector, $\mathbf{u}_1$, gives the maximum value of the ratio of between-class and within-class variances:

$$\gamma = \frac{u^T B u}{u^T W u} \tag{1}$$

Where W is the within-class variance and B is the between-class variance. The second vector, $\mathbf{u}_2$, is chosen to be linearly independent to $\mathbf{u}_1$ and produce the second largest ratio. This process is repeated until the last vector, $\mathbf{u}_d$, is found. It can be shown that these vectors are the d eignvectors corresponding to the first d largest eignvalues of the matrix $W^{-1}B$. Thus the LDA transformation is given by: $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_d]$.

3. MCE TRAINING ALGORITHM

The framework of MCE algorithm was first proposed by Katagiri, et al.[1]. It is a type of discriminant analysis. It achieves the minimum classification error by employing the gradient descent method using a loss function which is a differentiable function of misclassification measure defined as a close approximation of the actual classification error. Thus, the MCE algorithm is a more direct way to achieve the minimum misclassification rate than the conventional discriminative training algorithms. The MCE algorithm can be described as a three-step procedure. A discriminant functions $g_i(G, x)$ is prescribed in the first step. We usually use a simple Euclidean distance to define the discriminant functions [3]. The distance from pth observation to class i is given by:

$$D_i^{(p)} = \left\| TX^{(p)} - \mu_i \right\|^2 \tag{2}$$

where T is the transformation matrix with rank d ($d \le D$), D is the dimension of the original data, μ_i is the mean vector of class i.

The second step is to introduce the misclassification measure in which we embed the classification criterion in the overall minimum classification error formulations. There are many ways to define the misclassification measure. The commonest one is as follows [2, 4, 5]:

$$d_k(x^{(p)}) = -g_k(x^{(p)}, \Lambda) + \sum_{for\ all\ i \neq k} \frac{1}{N-1} g_i(x^{(p)}, \Lambda) \tag{3}$$

where $g_i(x^{(p)}, \Lambda)$, $i = 1, 2, \ldots, N$, is a set of discriminant functions; $x^{(p)}$ is the pth observation vector; N is the number of classes; Λ is the parameter set of each class; ζ is a positive number. One extreme case is when ζ approaches ∞, the misclassification measure becomes:

$$d_k(x^{(p)}) = -g_k(x^{(p)}, \Lambda) + g_i(x^{(p)}, \Lambda) \tag{4}$$

where class i has the largest discriminant value of all the classes other than class k. The classification criterion embedded in this definition is: $d_k(x^{(p)}) < 0$ means correct classification, $d_k(x^{(p)}) \ge 0$ implies misclassification. The loss function is then defined in the third step as a monotonic function suitable for gradient algorithms to smooth the misclassification measure. Sigmoid function is usually used in the definition of the loss function since it is a smooth zero-one function suitable for gradient algorithms. The loss function is given as:

$$L(x^{(p)}) = \frac{1}{1 + e^{-\alpha d(x^{(p)}, \Lambda)}} \tag{5}$$

The total loss function is defined as:

$$L = \sum_{p=1}^{P} L^{(p)} \tag{6}$$

The transformation matrix is obtained by minimizing this loss function through the steepest gradient descent algorithm in which the parameters at (k + 1)th iteration (such as transformation matrix, means and/or variance) are computed from the kth iteration:

$$T_{ij}(k+1) = T_{ij}(k) - \eta \frac{\partial L}{\partial T_{ij}} \tag{7}$$

$$m_i^{(C^{(p)})}(k+1) = m_i^{(C^{(p)})}(k) - \eta \frac{\partial L}{\partial m_i^{(C^{(p)})}} \tag{8}$$

$$m_i^{(N^{(p)})}(k+1) = m_i^{(N^{(p)})}(k) - \eta \frac{\partial L}{\partial m_i^{(N^{(p)})}} \tag{9}$$

where η ($\eta > 0$) is the adaptation constant.

4. AN ALTERNATIVE MCE TRAINING ALGORITHM

The MCE algorithm usually provides fairly good classification results on small data sets with few classes and small dimensionality of observation vectors. However, when it is operated on a data set with a large number of classes and high dimension, its performance is not satisfactory. The reason of failure of the MCE algorithm is that its classification criterion tries to minimize the misclassification measure $g_k(x, \Lambda) - g_i(x, \Lambda)$. When using the steepest gradient descent algorithm to minimize this classification measure, we hope that $g_k(x, \Lambda)$ decreases while $g_i(x, \Lambda)$ increases. But things do not always work as we expect. Since there is no constriction on the joint behaviour of $g_k(x, \Lambda)$ and $g_i(x, \Lambda)$, in many cases when $g_k(x, \Lambda)$ decreases, $g_i(x, \Lambda)$ decreases too, which makes the training process inefficient or even ineffective. Therefore minimizing the misclassification measure does not always lead to good results of classification. To overcome this defect of conventional MCE algorithm, we present a new definition of the misclassification measure, which is defined as follows:

$$d_k(x^{(p)}, \Lambda) = \frac{g_k(x^{(p)}, \Lambda)}{\left[\frac{1}{N-1} \sum_{for\ all\ j \neq k} g_j(x^{(p)}, \Lambda)^{\zeta}\right]^{1/\zeta}} \tag{10}$$

When ζ approaches ∞, the new definition becomes:

$$d_k(x^{(p)}, \Lambda) = \frac{g_k(x^{(p)}, \Lambda)}{g_i(x^{(p)}, \Lambda)} \tag{11}$$

In this definition the embedded classification criterion becomes: $d_k(x^{(p)}) < 1$ means correct classification and $d_k(x^{(p)}) \geq 1$ means incorrect classification. When using the steepest gradient descent algorithm, the partial derivatives of the loss function with respect to the T and μ are:

$$\frac{\partial L}{\partial T_{ij}} = -\alpha \sum_{p=1}^{P} L(1-L) \frac{\frac{\partial g_k(x^{(p)}, \Lambda)}{\partial T_{ij}} g_i(x^{(p)}, \Lambda) - \frac{\partial g_i(x^{(p)}, \Lambda)}{\partial T_{ij}} g_k(x^{(p)}, \Lambda)}{[g_i(x^{(p)}, \Lambda)]^2} \tag{12}$$

$$\frac{\partial L}{\partial \mu_k} = -\alpha \sum_{p=1}^{P} L(1-L) \frac{1}{g_i(x^{(p)}, \Lambda)} \frac{\partial g_k(x^{(p)}, \Lambda)}{\partial \mu_k} \tag{13}$$

and

$$\frac{\partial L}{\partial \mu_i} = \alpha \sum_{p=1}^{P} L(1-L) \frac{g_k(x^{(p)}, \Lambda)}{[g_i(x^{(p)}, \Lambda)]^2} \frac{\partial g_i(x^{(p)}, \Lambda)}{\partial \mu_i} \qquad (14)$$

5. CLASSIFICATION EXPERIMENTS

5.1 THE RESULTS OF THE MCE ALGORITHMS AND LDA

In this section, we study the performance of LDA, conventional MCE and our new alternative MCE algorithm. Deterding vowel database is used in this classification experiment, which has 11 vowel classes and each vowel is represented by a 10 dimension vector. The 11 vowels used in this database are listed in table 1.

Table 1. Vowels and Words Used in Recording Vowels

Vowel	As in word	Vowel	As in word	Vowel	As in word
I	heed	a:	hard	U	hood
I	hid	Y	hud	u:	who'd
E	head	O	hod	3:	heard
A	had	C:	hoard		

Figures 1 and 2 show the results of the three algorithms employed on training and testing data respectively. We denote the alternative MCE in these figures as new MCE or

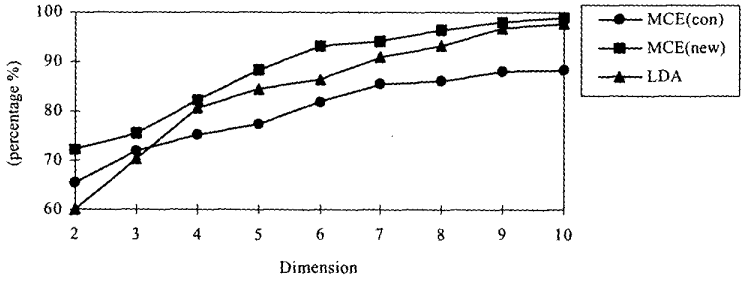

Figure 1. Recognition Rate of the Three Algorithm on Training Data

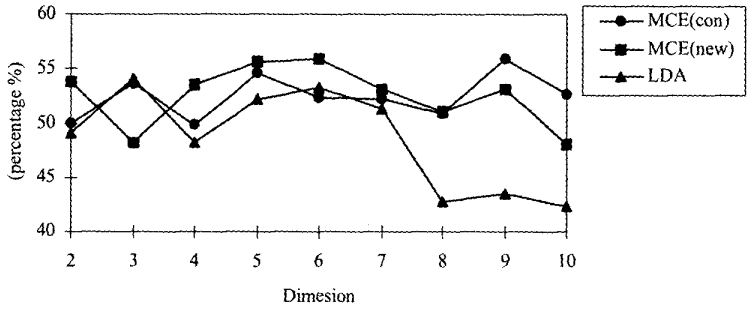

Figure 2. Recognition Rate of the Three Algorithm on Testing Data

MCE(new) and conventional MCE as MCE(con).

The above figures show that the performance of new MCE in training process is better than the other two, while the conventional MCE does worse than LDA. The performance of both the conventional MCE and the new MCE on testing data is better for most sub-space dimensions than LDA, except for dimension 3. Generally speaking, MCE can get a more generalized model than LDA. Figure 3 shows how new MCE works better in training process than conventional MCE. The tracks of a single observation in the g_i - g_k decision plane throughout the whole training process conducted by conventional MCE and new MCE respectively are shown in this figure. The Y-axis is $g_i(x, \Lambda)$ and X-axis is $g_k(x, \Lambda)$. O is the starting point of the observation at the beginning of the iterative process. The shaded part on the picture is the failure area of classification and the other part is correct classification area. The dividing line is $g_k(x, \Lambda) - g_i(x, \Lambda) = 0$, i.e. the embedded classification criterion. If the points of a class are far from the failure area, it means that this class is far from the other classes, i.e. the separation between it and any other class is good. From the figure we can see that after 1000 iterations the conventional MCE training moves the point closer to the failure area, which means the training process is ineffective and makes the separation between the classes worse. On the contrary, the new MCE training process moves the point further to the failure area after iteration. Naturally this training process will lead to a better classification result.

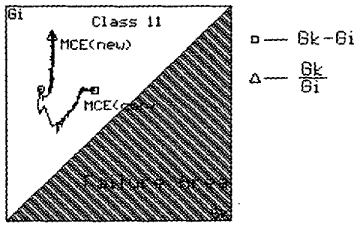

Figure 3. The Tracks of a Single Observation in the g_i - g_k Decision Plane Moving Throughout the Whole Training Process by Conventional MCE and New MCE

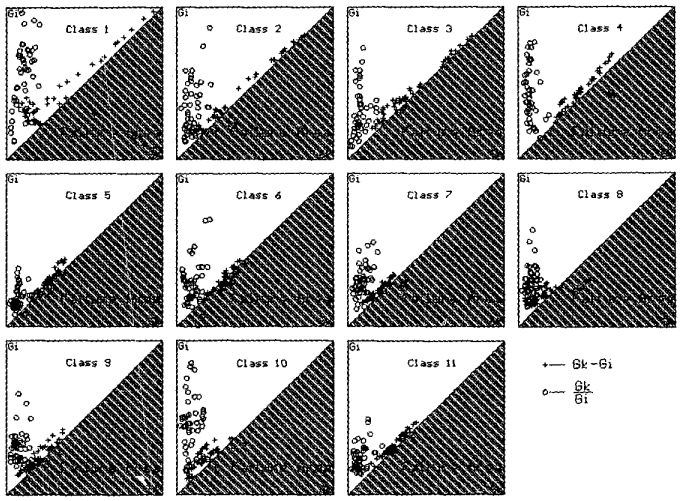

Figure 4. Distributions of the observations of each class on the $g_k(x, \Lambda) - g_i(x, \Lambda)$ Decision Plane After 1000 iterations by conventional MCE and new MCE respectively

Figure 4 shows the distributions of the observations of each class on the $g_k(x, \Lambda) - g_i(x, \Lambda)$ decision plane after 1000 iterations by conventional MCE and new MCE respectively. This figure clearly shows that many observations are still in the failure area with the conventional MCE training algorithm, while most observations move to the correct classification area with the new MCE training algorithm.

5.2 THE EFFECTS OF THE INITIAL VALUE OF TRANSFORMATION MATRIX

Despite the good results of MCE algorithms we achieved, there are still some defects in MCE training algorithms which make MCE not effective in some sub-spaces, such as the dimension 3 sub-space. One reason is that MCE algorithm is sensitive to the initial value of the transformation matrix. In the above experiments we use the unity matrix or its sub-matrix as the initial transformation matrix. Although the models we get from the training process beginning with this initial value work very well on training data, they only reach the local maximum since the number of training data is limited. That is why the MCE algorithms do not obtain much better results on testing data than LDA. In this paper we also tried another initial transformation matrix which is obtained from the LDA training process. The results are shown in Figures 5 and 6. In these figures, we denote the training

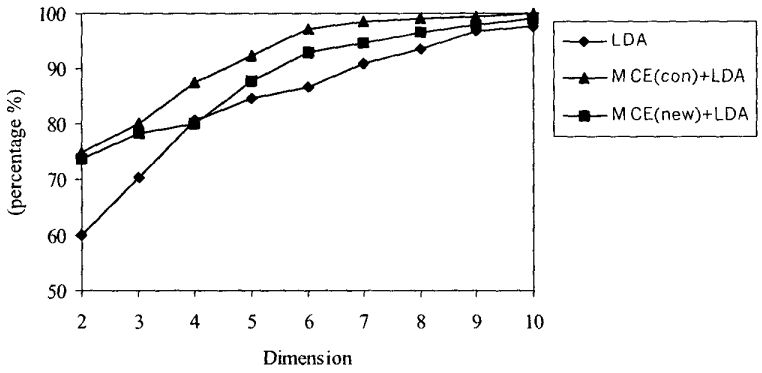

Figure 5. Recognition Rate of the LDA and MCE Algorithms with New Initial Transformation Matrix on Training Data

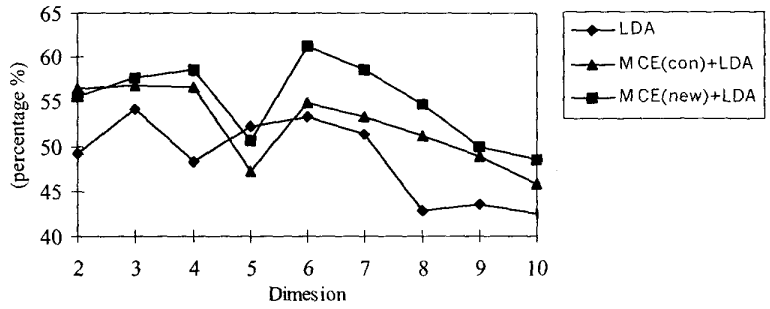

Figure 6. Recognition Rate of the LDA and MCE Algorithms with New Initial Transformation Matrix on Testing Data

process beginning with the initial transformation matrix obtained from the LDA as "MCE+LDA".

From these two figures, we can see that when using the new initial transformation matrix, the performance of the conventional MCE algorithm on training data gets better than the alternative MCE algorithm and the results of both MCE algorithms are much better than LDA in all lower dimensional sub-spaces. Interestingly the alternative MCE algorithm gets better results on testing data than conventional MCE algorithm. This means that the property of generalization of alternative MCE algorithm is better than conventional MCE algorithm.

6. SUMMARY

In this paper, we have investigated the use of the MCE algorithm for dimensionality reduction. We have also proposed an alternative criterion for improving the classification performance of the MCE algorithm. Although MCE algorithms achieve better results than LDA, their recognition scores on testing data are not satisfactory. The MCE algorithms are not very effective on certain sub-space(s) and they are not very stable or robust. Also, the MCE algorithm is sensitive to the initial value of the transformation matrix. How to choose the initial transformation matrix to make the MCE algorithm more effective is still a question for future research.

REFERENCES

[1] S. Katagiri, C.H.Lee and B.H. Juang., "A Generalized Probabilistic Descent Method", *Proceeding of the Acoustic Society of Japan*, Fall Meeting, pp. 141-142, 1990

[2] B.H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification", *IEEE Trans. On Signal Processing*, vol. 40, pp. 3043-3054, Dec. 1992

[3] K.K. Paliwal, M. Bacchiani and Y. Sagisaka, "Simultaneous Design of Feature Extractor and Pattern Classification Error Training Algorithm", *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, Boston, USA, pp. 67-76, September, 1995

[4] E. McDermott, "New Results for the Prototype-Based Minimum Error Classifier", *Preliminary Report*, ATR Human Information Processing Research Laboratories, 1994

[5] E. McDermott and S. Katagiri, "Prototype-Based Minimum Classification Error/Generalized Probabilistic Descent Training for Various Speech Units", *Computer Speech and Language*, Oct. 1994, pp. II-291 to II-294

[6] K.V. Mardia, J.T. Kent and J.M. Bibby, "Multivariate Analysis", *Academic Press Inc.*, San Diego, 1979

[7] S. Amari, "A Theory of Adaptive Pattern classifiers", *IEEE Trans. On Electronic Compution*, vol. 16, pp. 299-307, Jun. 1997

[8] K.K. Paliwal, "Dimensionality Reduction of the Enhanced Feature Set for the HMM-Based Speech Recognizer", *Digital Signal Processing*, no. 2, pp. 157-173, 1992

[9] D.X. Sun, "Feature Dimension Reduction Using Reduced-Rank Maximum Likelihood Estimation for Hidden Markov Models", *Proceedings of International Conference on Spoken Language Processing*, Philadelphia, USA, pp. 244-247, 1996

[10] W.L. Poston and Marchette, D. J., "Recursive Dimensionality Reduction Using Fisher's Linear Discriminant", *Pattern Recognition*, vol. 31, no. 7 pp881-888, 1998

[11] N.A. Campbell, "Shrunken Estimators in Discriminant and Canonical Variate Analysis", *Apply Statistics*, vol. 29, no. 1, pp5-14, 1980

[12] B.H. Juang and S. Katagiri, "Discriminative Training", *Journal of Acoustical Society of Japan*, vol.13, no. 6, pp333-339, 1992

[13] A. Mkhadri, "Shrinkage Parameter for the Modified Linear Discriminant Analysis", *Pattern Recognition Letters*, vol.16, March, pp267-275, 1995
[14] A. Biem and S. Katagiri, "Feature Extraction Based on Minimum Classification Error/Generalized Probabilistic Descent Method", *Computer Speech and Language*, Oct. 1994, pp II-275 to II-278
[15] N. Kambhatla, "Local Models and Gaussian Mixture Models for Statistical Data Processing", *PhD Thesis*, Oregon Graduate Institute of Science & Technology, 1996
[16] P.A. Lachenbruch, "Discriminant Analysis", *Hafner Press*, New York, 1975
[17] I.T. Jolliffe, "Principal Componant Analysis", *Springer-Verlag*, New York, 1986