International Symposium on Chinese Spoken
Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

COMPARATIVE STUDY OF LINEAR FEATURE TRANSFORMATION TECHNIQUES FOR MANDARIN DIGIT STRING RECOGNITION

Jian Shan, Yuanyuan Shi, Jia Liu, Runsheng Liu

Department of Electronic Engineering, Tsinghua University, Beijing 100084, P. R. China
shanjian00@mails.tsinghua.edu.cn

ABSTRACT

Linear feature transformation technique is widely used to improve feature discriminability. It can reduce the dimensionality of the feature space, un-correlate the feature components, hence more discriminative model can be obtained. In this paper we compare three discriminative linear transformation approaches in Mandarin digit string recognition (MDSR) system. Compared with the conventional Linear Discriminant Analysis (LDA), two other discriminative linear transformation methods derived from LDA, that is Confusion Discriminant Analysis (CDA) and Heteroscedastic Discriminant Analysis (HDA), are studied on the basis of state-specific confusable class definition and its class-dependent linear transformations.

1. INTRODUCTION

Several researches have verified the effectiveness of LDA for the task of speech recognition, especially for the confusing vocabulary recognition whose performance can be improved significantly. In this paper we investigate the discriminative linear transformation techniques, except for LDA, in order to achieve high recognition accuracy for Mandarin digit string recognition (MDSR) system, a well-known task of confusable speech recognition also.

Confusion Discriminant Analysis (CDA) performs the similar functions as LDA to find the most discriminative features. The main difference between CDA and LDA is how to compute the scatter matrices. The scatter matrices of the CDA are computed from the data collected to be confusable for any given state of HMM, so the CDA must be applied on the confusion class specific basis while LDA can be performed on both class-dependent and class-independent bases.

Heteroscedastic Discriminant Analysis (HDA) is also used to handle the heteroscedastic unequal covariance classes. It is a model-based generalization of LDA by the criterion of maximum likelihood. We apply a confusion class specific HDA for MDSR system. This approach is also compared with other methods.

Each mandarin digit has a special group of confusing digits or rival digits. The states of these rival digit HMMs are defined as the confusing classes of the discriminant analysis. The utterances in the training data are recognized by each digit model, and the candidates of recognized digits having likelihood within a threshold of the correct model are considered as the rival digits. Then the speech frames of the alignments to the states are used as the confusion data.

This paper is organized as follows. The base-line Mandarin digit speech recognition system is described in section 2. The linear discriminative analysis methods are compared and discussed in section 3. The experimental results are provided in section 4. The last section is the conclusion.

2. BASE-LINE MANDARIN DIGIT STRING RECOGNITION SYSTEM

In the Base-line Mandarin digit string recognition system 19 CHMMs based on the phonemes (17 phonemic CHMMs, a background silence CHMM and a within-syllable pause CHMM) are established. These phonemic models include 8 initials and 9 finals correspond to the consonants and vowels of the Mandarin digit pronunciations respectively. Each digit model consists of the sequence of initial and final models. The within-syllable context-dependency is not considered here in order to reduce the number of the models. The speech frame length is 24ms with 12ms overlap. The feature vector include 12 MFCC coefficients, 12 ΔMFCC coefficients, normalized energy and Δenergy, that is, a 26-dimensional feature vector. The recognition models are trained using the segmental K-Means training procedure.

Each phonemic model may have 3 to 6 states with the left to right topology. Using different number of states for the phonemic models will make the recognition models have different discriminability and computing complexity.

3. COMPARISONS OF LINEAR DISCRIMINATIVE ANALYSIS METHODS

For baseline MDSR system, it is found that some digits are easily confused with others because of the phonetic similarity between their initials or finals. So we use linear feature transformation to improve the discriminability for these easy confusable phoneme models.

The 26-dimensional feature vector is used as the input features of linear transformation. The states of CHMMs are defined as the classes in respect that the phonemes of these digit pronunciations are modeled by the several state distribution. The class-dependent discriminative transformation is used. In this way improved discrimination for the phonemes can be obtained. In order to derive the discriminative transformation for the given state, the data modeled by the state distribution and the confusion data to that state must be gathered before computing the within-class and between-class scatter matrices.

The confusing data gathering approach is based on the Viterbi search. In this approach, we use following three steps. (1) Each training utterance is aligned against the correct initial-final digit models to allocate the feature vector of each frame for the correct state in the initial and final. (2) Each training utterance is aligned against all the other incorrect digit models and the likelihood is compared with the correct likelihood that is determined in step (1). Then top N candidates can be selected simply as confusion data (top N threshold). These frames of the feature vectors that are allocated to the state of the different initial or final model in the confusion alignment are pooled and used as the confusion data for that state. (3) After the training utterances are all processed in step (1) and (2), the in-class data and between-class data (or confusion data) for each state are collected.

3.1 LDA

Firstly we define the state class as C_q, the confusion class as $C_{\overline{q}}$. For LDA the within-class and between-class scatters can be computed as below:

$$W_{C_q} = \frac{1}{N_x} \sum (x_i - m_{C_q})(x_i - m_{C_q})^t$$

$$W_{C_{\overline{q}}} = \frac{1}{N_y} \sum (y_i - m_{C_{\overline{q}}})(y_i - m_{C_{\overline{q}}})^t$$

$$S_W = p_{C_q} W_{C_q} + p_{C_{\overline{q}}} W_{C_{\overline{q}}} \qquad (1)$$

$$S_B = \sum_{i=C_q, C_{\overline{q}}} p_i (m_i - m)(m_i - m)' \qquad (2)$$

where x_i is the in-class feature vector of one state; y_i is the confusion-class feature vector of the same state; N_x and N_y are the number of in-class and confusion-class frames, respectively.

In LDA, by the discriminative transformation matrix T, the original feature vectors are transformed to the new feature vectors, which involves that S_w and S_B are diagonalized simultaneously.

The optimal criterion in the linear discrimination analysis usually is the ratio of between-class scatter to the within-class scatter. Equation (3) presents the criterion. The optimal solution, which is obtained by maximizing Equation (3), defines the axes of the transformed space. If the class-dependent transformation is used in the linear discrimination analysis, then L different optimal transformations can be obtained for L-class by Equation (3).

$$J = \frac{|S_B|}{|S_W|} \qquad (3)$$

After applying the class-dependent transformation, the original state feature space is projected into a new space in which each element of the feature vector is uncorrelated for both the in-class and confusion data of the given state, and each feature element of the in-class data has a unit variance. The larger the variance of the feature element of the confusion data is, the more discriminative the element is. The elements with the variance smaller than one may not discriminate the in-class and confusion class well, which should be discarded. Therefore, the feature elements can be selected according to the descending order of the variances of the confusion distributions. Although the dimensionality of features is reduced, the discriminability can be increased in the lower dimensional space.

3.2 CDA

The Confusion Discriminant Analysis (CDA) is similar to class-dependent LDA, which rotate the original feature space to the new feature space that has the best discriminability of the features based on the CDA criterion. The main difference between CDA and class-dependent LDA is the definition of within-class and between-class scatter matrices. For CDA two scatter matrices are respectively expressed as below:

$$S_{W,CDA} = \frac{1}{N_x} \sum (x - m_x)(x - m_x)^T \qquad (4)$$

$$S_{B,CDA} = \frac{1}{N_y} \sum (y - m_x)(y - m_x)^T \qquad (5)$$

In order to compare CDA with LDA, we have to analyze the relationship between the scatter matrices of LDA and CDA. Let the scatter matrices of LDA be $S_{W,LDA}$ and $S_{B,LDA}$ respectively. According to their definitions, the scatter matrices of LDA can respectively be expressed as:

$$\begin{aligned} S_{W,LDA} &= \frac{1}{N_x + N_y}\left[\sum (x - m_x)(x - m_x)^T + \sum (y - m_y)(y - m_y)^T\right] \\ &= \frac{N_x}{N_x + N_y} S_{W,CDA} + \frac{N_y}{N_x + N_y} W_y \\ &= p_x S_{W,CDA} + p_y W_y \qquad (6) \end{aligned}$$

$$\begin{aligned} S_{B,LDA} &= p_x (m_x - m)(m_x - m)^T + p_y (m_y - m)(m_y - m)^T \\ &= p_x p_y (m_x - m_y)(m_x - m_y)^T \qquad (7) \end{aligned}$$

The scatter matrix of CDA, Equation (5), can also be expressed as:

$$\begin{aligned} S_{B,CDA} &= \frac{1}{N_y} \sum (y - m_y + m_y - m_x)(y - m_y + m_y - m_x)^T \\ &= W_y + (m_y - m_x)(m_y - m_x)^T \qquad (8) \end{aligned}$$

where $W_y = \frac{1}{N_y}\sum (y - m_y)(y - m_y)^T$; and $m = p_x m_x + p_y m_y$. The optimal criterions are respectively:

$$J_{LDA} = \frac{| p_x p_y (m_x - m_y)(m_x - m_y)' |}{| p_x S_{W,CDA} + p_y W_y |} \quad (9)$$

$$J_{CDA} = \frac{| W_y + (m_x - m_y)(m_x - m_y)' |}{| S_{W,CDA} |} \quad (10)$$

Comparing Equation (9) with Equation (10), we find that W_y, which is a part of the denominator of J_{LDA}, is moved to the numerator of J_{CDA} . This change means that CDA only compresses the within-class scatter of a class, but don't guarantee the increase of the between-class scatter of this class. It may expand the within-class scatter of the other class.

In our experiments, C_x is defined as the class that is to be recognized, while C_y is defined as the confusable class that is easily confused with C_x. The training and recognition procedures of CDA are similar to those of the class-dependent LDA.

Because of the phonetic characteristic of Mandarin, the confusion data for the states of some phonemes, for example the states in initial [p], are quite limited. In this case, the discriminant transformation is difficult to be robustly estimated by using the confusion data. Our experiments show that at least 200 frames of the confusion data for each state are needed for the efficient estimation of the discriminant transformation. So Equation (2) is used in the transformation computation when the frame number of confusion data is smaller than 200.

3.3 HDA

LDA is related to the maximum-likelihood estimation of parameters for a Gaussian model, if two a priori assumptions on the structure of the model were kept. The first assumption is that all the class-discrimination information resides in a p-dimensional subspace of the n-dimensional feature space. The second assumption is that the within-class variances are equal for all the classes. Heteroscedastic discriminant analysis (HDA)[3] is able to handle heteroscedasticity by dropping the assumption of equal variance in the parametric model.

Let's have two Gaussian models. If q be a nonsingular n×n matrix which defines a linear transformation, mapping the feature vector x into new feature vector y , $q : y = qx$. Assume that only the first p components of y carry the class-discrimination information. Reference [3] has given an estimate of q :

$$\hat{q} = \arg\max_{q} \left\{ -\frac{N}{2} \ln\left| q_{n-p} \Sigma q'_{n-p} \right| - \sum_{j=1}^{2} \frac{N_j}{2} \ln\left| q_p \Sigma_j q'_p \right| + N \ln|q| \right\} \quad (11)$$

Since the closed-form solution $\hat{q}$ cannot be directly obtained by the maximum likelihood estimate method, the optimal process has to be performed numerically. But when $\Sigma_1 = \Sigma_2$, the transformation matrix of LDA is the solution of Equation (11). According to Reference [3], LDA can be considered the maximum-likelihood estimate of a constrained model, while HDA can be used for any Gaussian model.

The training and recognition procedures of class-dependent HDA are similar to those of class-dependent LDA. The only difference is that the transformation matrix is estimated by Equation (11).

4. EXPERIMENTAL RESULTS

4.1 Speech Database

Two databases are used in our experiments. Database one is Mandarin digit database that includes 160 speakers (80 males, 80 females). Each speaker pronounces the 11 digits one time. Database two is Mandarin digit string database that includes 40 female speakers. Each speaker pronounces 80 digit strings one time. And the average length of these digit strings is 4. The digit " 1" has two different pronunciations, one is " yi" , the other is " yao" that is denoted as "a".

"Leave-One-Out"[4] method is applied to training and recognition. We divide these speakers into 4 groups. Then in each testing procedure, 3 groups are used for training set and 1 group is used for testing set. The recognition rate will be tested four times with no testing set overlap. The averaging recognition rate of four times is used as the final recognition rate.

For Database one the recognition error rate of baseline system is 2.84, and for Database two it is 13.50.

4.2 Comparison

Here we compare the performances of the different discriminative transformation methods. The class-dependent LDA, CDA and HDA are used in the experiments.

The recognition error rate curves are shown on Figure 1 and Figure 2. The different dimension of the transformed feature vector is used. The dimension varies from 10 to 26 or 14 to 26.

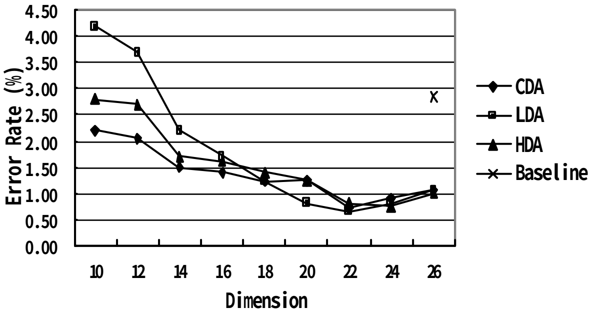

Figure 1 Error rate comparison of different discriminative transformations with different dimension in transformed space using the Database one

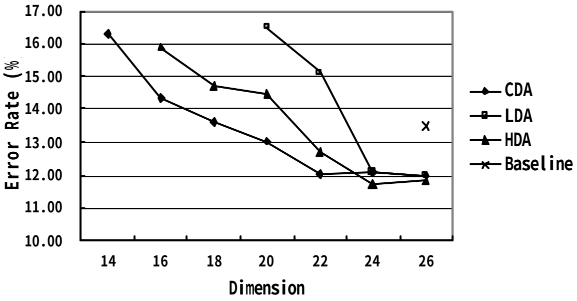

Figure 2 Error rate comparison of different discriminative transformations with different dimension in transformed space using the Database two

Figure 1 is the recognition error rate comparison of the different discriminative transformation methods for the Database one; and Figure 2 is the recognition error rate comparison for the Database two. The experiments show the similar results can be obtained based on two different speech databases.

The experiments show these three kinds of linear transformation methods can all reduce the error rate of the baseline recognition system, and their optimal performances are similar. Although CDA loosens the condition in optimizing the model and HDA drops the assumption of equal variance, both methods cannot reduce the error rate further more, compared with LDA.

The sensitivity of these three methods to the reduction of dimension is different. From the figures above, we can see that CDA achieves the best results in two different databases. This is also shown in Table 1, where we list the dimension number of transformation at which the error rate is the same as our 26-dimension baseline.

	Baseline	LDA	CDA	HDA
Database One	26	13	<10	10
Database Two	26	23	18	21

Table 1 The dimension of transformation at which the error rate is the same as our 26-dimension baseline

Comparing computational procedure of three kinds of transformation matrices that are described in section 3, from Figure 1 and Figure 2 we can find that CDA has the fewest computation complexity and the best performance when fewer features are used.

5. CONCLUSIONS

In this paper several class-dependent linear feature transformation methods are investigated based on the speaker independent Mandarin digit recognition system. It increases effectively the discriminability of the confusable digits, which exists in the consonants and vowels of Mandarin digits. The recognition rate is increased greatly. Compared with the baseline system, the error rate is reduced from 2.84 to 0.65 for database one and from 13.50 to 11.70 for database two by using the discriminative transformation.

LDA assumes that each class has equal within-class covariance. It is not the optimal transform when the class distributions are heteroscedastic. Then HDA are proposed to generalize LDA to handle heterosecdasticity by dropping the assumption of equal variance in the parametric model. Confusion Discriminant Analysis performs much the same function as state-dependent LDA. The main difference between CDA and LDA is how to compute the scatter matrices. This difference makes CDA more effective when applied on the confusion class specific basis.

6. REFERENCES

[1] Y. Y. Shi. Research on Methodologies for Mandarin Digit Speech Recognition: [Ph.D. Dissertation]. Tsinghua, 2002

[2] C. J. Leggetter. Improved acoustic modeling for HMMs using linear transformations: [Ph.D. Dissertation]. Cambridge, 1995

[3] Nagendra Kumar, Andreas G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Communication, 1998, 26(4): 283-297

[4] R. O. Duda, P. E. Hart. Pattern Classification and Scene Analysis. New York: John Wiley. 1973

[5] K. Fukunaga. Introduction to statistical pattern recognition. Academic Press, NY, 1990