GMM-BASED BHATTACHARYYA KERNEL FISHER DISCRIMINANT ANALYSIS FOR SPEAKER RECOGNITION

Yi-Hsiang Chao[1,2], *Hsin-Min Wang*[1] *and Ruei-Chuan Chang*[1,2]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2] Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan
{yschao, whm}@iis.sinica.edu.tw, rc@cc.nctu.edu.tw

ABSTRACT

Clearly, the linear discriminant classifier is not robust enough to cope with most real-world data classification problems. Kernel Fisher Discriminant Analysis (KFDA) tries to increase the expressiveness of the discriminant based on the high order statistics of the data set. In this paper, we propose the GMM-based KFDA with the Bhattacharyya kernel to obtain a transformation, or called a speaker eigenspace, based on which the transformed MFCC features are more discriminative for speaker recognition. In our approach, the eigenspace is directly constructed from the complete GMM parameter set, rather than the supervectors considering mean vectors only as the eigenvoice approach. Moreover, FDA, which is believed to be more appropriate for classification accuracy than Principal Component Analysis (PCA), is applied for eigenspace construction. The speaker identification experiments show that the new features outperform the MFCC features, in particular when the amount of enrollment data for each speaker is very small.

1. INTRODUCTION

Speech recognition and speaker recognition are two different things. The objective in speech recognition is to minimize the inter-speaker variation while maximizing the intra-speaker variation among acoustic units, but vice versa in speaker recognition. Therefore, these two tasks should better use different signal traits as input. However, the most widespread feature parameters used to date in both tasks are Mel-Frequency Cepstral Coefficient (MFCC) features, which were originally designed to fulfill the demand of speech recognition. Though the MFCC-based Gaussian Mixture Model (GMM) [1] has been applied to speaker recognition in recent years, this approach performs well only when a large amount of enrollment data for each client speaker is available. In other words, in the training phase, the distribution of MFCC features from each client speaker should be wide enough to cover all possible pronunciations, in particular when the speaker recognition is conducted under the text-independent mode. In theory, speaker characteristics should be invariant to the size of enrollment data and different pronunciations of the same speaker. It is crucial to develop more reliable features that magnify the inter-speaker variation while reducing the intra-speaker variation for speaker recognition.

Fisher Discriminant Analysis (FDA) [2] has been applied to feature transformation in many pattern classification problems. This technique is used to seek directions that maximize the between-class scatter while minimizing the within-class scatter. However, for most real-world data (e.g., speech frames) the linear discriminant is not complex enough. Therefore, Kernel Fisher Discriminant Analysis (KFDA) [3] tries to increase the expressiveness of the discriminant based on the high order statistics of the data set.

In this paper, we want to find a speaker space that can better discriminate the speakers from each other. We propose the GMM-based KFDA with the Bhattacharyya kernel (BKFDA) to obtain a transformation, or called a speaker eigenspace, based on which the transformed MFCC features are more discriminative for speaker recognition. The rest of this paper is organized as follows: FDA and KFDA are briefly introduced in Section 2. The GMM-based BKFDA is presented in Section 3. Then, the application of GMM-based BKFDA to speaker identification is described in Section 4. Finally, the experimental results are discussed in Section 5, and concluding remarks are made in Section 6.

2. FISHER DISCRIMINANT ANALYSIS WITH KERNELS

2.1. Fisher Discriminant Analysis (FDA)

Suppose that there are C classes and each class i has n_i d-dimensional data samples, $\mathbf{X}_i = \{\mathbf{x}_1^i,..,\mathbf{x}_{n_i}^i\}$. We want to find a linear transformation matrix $\mathbf{W}$ for the original data samples such that the following Fisher's criterion function J($\mathbf{W}$) is maximized,

$$\mathrm{J}(\mathbf{W}) = \frac{\left|\mathbf{W}^T \mathbf{S}_b \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_w \mathbf{W}\right|}, \tag{1}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are, respectively, the between-class and the within-class scatter matrices defined as follows,

$$\mathbf{S}_b = \sum_{i=1}^{C} n_i \left(\mathbf{m}_i - \mathbf{m}\right)\left(\mathbf{m}_i - \mathbf{m}\right)^T, \tag{2}$$

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \mathbf{m}_i)(\mathbf{x}_j^i - \mathbf{m}_i)^T. \tag{3}$$

$\mathbf{m}$ and $\mathbf{m}_i$ are, respectively, the overall sample mean vector and the sample mean vector of the ith class computed by,

$$\mathbf{m} = \frac{1}{n} \sum_{i=1}^{C} n_i \mathbf{m}_i, \tag{4}$$

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j^i, \tag{5}$$

where $n=n_1+n_2+\ldots+n_C$. The column vectors of an optimal **W** maximizing J(**W**) are the generalized eigenvectors of $\mathbf{S}_b$ and $\mathbf{S}_w$ (or the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ if $\mathbf{S}_w$ is nonsingular) corresponding to the largest K eigenvalues ($K \le \min\{C\text{-}1, d\}$).

2.2. Kernel Fisher Discriminant Analysis (KFDA)

Let Φ be a nonlinear mapping from the input feature space R^d into the implicit higher dimensional (maybe infinite) feature space F. Then, we can find Fisher's discriminant in F by maximizing

$$\mathrm{J}(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b^{\Phi} \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^{\Phi} \mathbf{w}}, \tag{6}$$

where **w** is a column vector of **W** in F, and $\mathbf{S}_b^{\Phi}$ and $\mathbf{S}_w^{\Phi}$ are the between-class and the within-class scatter matrices in F, i.e.,

$$\mathbf{S}_b^{\Phi} = \sum_{i=1}^{C} n_i \left(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi}\right)\left(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi}\right)^T, \tag{7}$$

$$\mathbf{S}_w^{\Phi} = \sum_{i=1}^{C} \sum_{j=1}^{n_i} (\Phi(\mathbf{x}_j^i) - \mathbf{m}_i^{\Phi})(\Phi(\mathbf{x}_j^i) - \mathbf{m}_i^{\Phi})^T, \tag{8}$$

with $\mathbf{m}_i^{\Phi} = (1/n_i)\sum_{j=1}^{n_i} \Phi(\mathbf{x}_j^i)$. Let $\mathbf{X} = \mathbf{X}_1 \cup \ldots \cup \mathbf{X}_C = \{\mathbf{x}_1, .., \mathbf{x}_n\}$, $\mathbf{m}^{\Phi}$ can be computed by

$$\mathbf{m}^{\Phi} = \frac{1}{n}\sum_{i=1}^{C} n_i \mathbf{m}_i^{\Phi} = \frac{1}{n}\sum_{i=1}^{C} n_i \left(\frac{1}{n_i}\sum_{j=1}^{n_i} \Phi(\mathbf{x}_j^i)\right) = \frac{1}{n}\sum_{j=1}^{n} \Phi(\mathbf{x}_j). \tag{9}$$

Usually, it is impossible to directly compute $\Phi(\mathbf{x})$. We can introduce a kernel function $k(\mathbf{x},\mathbf{y})=(\Phi(\mathbf{x})\cdot\Phi(\mathbf{y}))$, which is the inner product of two vectors $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ in F, to solve the maximization problem. The function $k(\ )$ must be symmetric positive and obey the Mercer's condition [4]. In this way, hopefully, the data, which is not linearly separable in the input space R^d, can be linearly separable in the space F.

From the theory of reproducing kernels, we know that any solution $\mathbf{w} \in F$ must lie in the span of all data samples in F, i.e.,

$$\mathbf{w} = \sum_{j=1}^{n} \alpha_j \Phi(\mathbf{x}_j). \tag{10}$$

Let $\boldsymbol{\alpha}^T = [\alpha_1, \ldots, \alpha_n]_{1\times n}$, we can find Fisher's discriminant in F by maximizing

$$\mathrm{J}(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{N} \boldsymbol{\alpha}}. \tag{11}$$

M and **N** are computed by

$$\mathbf{M} = \sum_{i=1}^{C} n_i (\boldsymbol{\eta}_i - \boldsymbol{\eta}_0)(\boldsymbol{\eta}_i - \boldsymbol{\eta}_0)^T, \tag{12}$$

$$\mathbf{N} = \sum_{i=1}^{C} \mathbf{K}_i (\mathbf{I}_{n_i} - \mathbf{1}_{n_i}) \mathbf{K}_i^T, \tag{13}$$

where $\boldsymbol{\eta}_i$ is an n×1 vector with $(\boldsymbol{\eta}_i)_j = (1/n_i)\sum_{k=1}^{n_i} k(\mathbf{x}_j, \mathbf{x}_k^i)$, and $\boldsymbol{\eta}_0$ is an n×1 vector with $(\boldsymbol{\eta}_0)_j = (1/n)\sum_{k=1}^{n} k(\mathbf{x}_j, \mathbf{x}_k)$, $\mathbf{K}_i$ is an $n \times n_i$ matrix with $(\mathbf{K}_i)_{pq} = k(\mathbf{x}_p, \mathbf{x}_q^i)$, $\mathbf{I}_{n_i}$ is an $n_i \times n_i$ identity matrix, and $\mathbf{1}_{n_i}$ is an $n_i \times n_i$ matrix with all entries equal to $1/n_i$. This maximization problem can be solved (analogous to the prime problem in the input space as defined in Eq. (1)) by finding the leading eigenvectors of $\mathbf{N}^{-1}\mathbf{M}$. The projection of a new pattern **x** onto **w** is given by

$$(\mathbf{w} \cdot \Phi(\mathbf{x})) = \sum_{j=1}^{n} \alpha_j (\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x})) = \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \mathbf{x}). \tag{14}$$

There are several implementation issues. First, **N** is obvious a singular matrix because we are estimating an n dimensional covariance structure from n data samples. In order to make **N** a positive matrix and at the same time preserve its eigenvalues, we can simply add a small constant μ to its diagonal components, i.e., we use $\mathbf{N}_{\mu} = \mathbf{N} + \mu\,\mathbf{I}$ to replace **N** in Eq. (11) [3]. Second, **N** and **M** are both of size n×n, which in practice can be very big. We need to solve an n×n eigen-decomposition problem, which might be intractable for large n. One possible solution is the sampling scheme, i.e., to restrict **w** to lie in a subspace spanned by l examples instead of all training samples in F [5],

$$\mathbf{w} = \sum_{j=1}^{l} \alpha_j \Phi(\mathbf{z}_j), \tag{15}$$

where $l \ll n$. The examples $\mathbf{z}_j$, $j = 1,\ldots,l$, could either be a subset of data samples or be estimated by some clustering algorithms.

3. GMM-BASED BHATTACHARYYA KFDA

Since the GMM is a very good representation of the training samples, its model parameters in fact fit the requirement of $\mathbf{z}_j$ in Eq. (15) very well. Suppose that each class GMM has R mixture components, $\mathbf{X}_i$, the training samples of the ith class, can be represented by $G_i(p_r^i, \mathbf{g}_r^i, r = 1,\ldots,R)$, the ith class GMM. $\mathbf{g}_r^i \sim N(\boldsymbol{\mu}_r^i, \boldsymbol{\Sigma}_r^i)$ is the rth component of the ith class GMM, and p_r^i, $\boldsymbol{\mu}_r^i$ and $\boldsymbol{\Sigma}_r^i$ are, respectively, the mixture weight, the mean vector, and the covariance matrix associated with $\mathbf{g}_r^i$. The basic idea here is that all mixture components of a GMM are regarded as representative examples of the corresponding class. Therefore, the GMM-based KFDA is derived as follows,

$$\mathbf{S}_b^{\Phi} = \sum_{i=1}^{C} \left(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi}\right)\left(\mathbf{m}_i^{\Phi} - \mathbf{m}^{\Phi}\right)^T, \tag{16}$$

$$\mathbf{S}_w^{\Phi} = \sum_{i=1}^{C} \sum_{r=1}^{R} p_r^i (\Phi(\mathbf{g}_r^i) - \mathbf{m}_i^{\Phi})(\Phi(\mathbf{g}_r^i) - \mathbf{m}_i^{\Phi})^T, \tag{17}$$

where $\mathbf{m}^{\Phi} = (1/C)\sum_{i=1}^{C} \mathbf{m}_i^{\Phi}$ and $\mathbf{m}_i^{\Phi} = \sum_{r=1}^{R} p_r^i \Phi(\mathbf{g}_r^i)$. Notice here the weight associated with $\Phi(\mathbf{g}_r^i)$ is p_r^i, the mixture weight associated with $\mathbf{g}_r^i$. We let the vector **w** lie in the span of C×R mixture components of GMMs in F, i.e.,

$$\mathbf{w} = \sum_{i=1}^{C} \sum_{r=1}^{R} \alpha_{ir} p_r^i \Phi(\mathbf{g}_r^i). \tag{18}$$

Then, we can apply the Bhattacharyya kernel [6] in computing the inner product of $\Phi(\mathbf{g}_r^i)$ and $\Phi(\mathbf{g}_s^j)$ in F. We define the kernel function as $k(\mathbf{g}_r^i,\ \mathbf{g}_s^j)=\exp(-\ \gamma\ \|\mathbf{g}_r^i - \mathbf{g}_s^j\|^2)$, and apply the Bhattacharyya distance [2] in computing $\|\mathbf{g}_r^i - \mathbf{g}_s^j\|$, i.e.,

$$\|\mathbf{g}_r^i - \mathbf{g}_s^j\|^2 = \frac{1}{8}(\boldsymbol{\mu}_r^i - \boldsymbol{\mu}_s^j)^T \left[\frac{\boldsymbol{\Sigma}_r^i + \boldsymbol{\Sigma}_s^j}{2}\right]^{-1} (\boldsymbol{\mu}_r^i - \boldsymbol{\mu}_s^j) + \frac{1}{2}\ln \frac{\left|\frac{\boldsymbol{\Sigma}_r^i + \boldsymbol{\Sigma}_s^j}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_r^i|}\sqrt{|\boldsymbol{\Sigma}_s^j|}}. \tag{19}$$

Therefore, we term the proposed approach as the GMM-based Bhattacharyya KFDA (BKFDA).

4. APPLICATION TO SPEAKER RECOGNITION

4.1. Eigenspace Construction

Suppose that there are C training speakers (classes) and each speaker has his/her well-trained GMM. In the eigenvoice approach [7], the mean vectors of each speaker's GMM are concatenated to form a supervector first. A speaker eigenspace is then constructed by performing PCA on these C supervectors. Considering the fact that PCA seeks directions that are efficient for representation whereas FDA seeks directions that are efficient for discrimination [8], for the speaker recognition task, it is believed that FDA is more appropriate for classification accuracy than PCA. However, it is impossible to calculate the within-class scatter matrix when FDA is performed on the above supervectors because each speaker (class) only has one supervector. The proposed GMM-based BKFDA is more appropriate for eigenspace construction than the standard eigenvoice approach because the eigenspace is directly constructed from the complete GMM parameter set rather than the supervectors considering mean vectors only.

We first pool all training speakers' data to train a Universal Background Model (UBM) [9] with R mixture components $\mathbf{g}_r$, $r=1, \ldots, R$, i.e., $\mathbf{X}_1 \cup \ldots \cup \mathbf{X}_C = \mathbf{X} \sim G(p_r, \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r, r=1,\ldots,R)$. Since the UBM is a large GMM covering the distribution of all possible pronunciations from all speakers, we can let the vector $\mathbf{w}_k$ lie in the span of R mixture components of UBM in F, i.e.,

$$\mathbf{w}_k = \sum_{r=1}^{R} \alpha_{rk} p_r \Phi(\mathbf{g}_r). \tag{20}$$

We then apply the Bayesian adaptation [10] to train the speaker GMMs from the UBM using the speaker specific training data. Since all speaker GMMs are adapted from the UBM, they have similar intra-speaker variation structures. The situation fits the requirement of making good use of FDA that all classes have similar within-class scatter structures. Let $\boldsymbol{\alpha}_k{}^T = [\alpha_{1k}, \ldots, \alpha_{Rk}]_{1\times R}$, same as Eq.(11), we need to maximize J(α_k). Here, the matrices $\mathbf{M}$ and $\mathbf{N}$ are defined as,

$$\mathbf{M} = \mathbf{P}\left(\sum_{i=1}^{C} (\boldsymbol{\eta}_i - \boldsymbol{\eta}_0)(\boldsymbol{\eta}_i - \boldsymbol{\eta}_0)^T\right)\mathbf{P}, \tag{21}$$

$$\mathbf{N} = \mathbf{P}\left(\sum_{i=1}^{C} \mathbf{K}_i (\mathbf{P}_i - \mathbf{p}_i \mathbf{p}_i{}^T) \mathbf{K}_i{}^T\right)\mathbf{P}. \tag{22}$$

$\boldsymbol{\eta}_i$ is a $R\times1$ vector with $(\boldsymbol{\eta}_i)_j = \sum_{r=1}^{R} p_r^i \cdot k(\mathbf{g}_j, \mathbf{g}_r^i)$ and $\boldsymbol{\eta}_0$ is a $R\times1$ vector with $(\boldsymbol{\eta}_0)_j = (1/C)\sum_{i=1}^{C} (\boldsymbol{\eta}_i)_j$, $\mathbf{K}_i$ is a $R\times R$ matrix with $(\mathbf{K}_i)_{pq} = k(\mathbf{g}_p, \mathbf{g}_q^i)$, $\mathbf{P}$ and $\mathbf{P}_i$ are diagonal matrices of size $R\times R$ whose rth diagonal elements are p_r and p_r^i, respectively, and $\mathbf{p}_i$ is a $R\times1$ vector whose rth element is p_r^i. Again, the maximization problem can be solved by finding the leading eigenvectors of $\mathbf{N}^{-1}\mathbf{M}$, as described in Section 2.2.

4.2. Feature Transformation

In the eigenvoice [7] or eigen-MLLR [11] approaches, the coordinate in the speaker eigenspace can be found and used to construct a model for a speaker based on a small amount of enrollment data. In the extreme case, the coordinate with respect to each speech frame can be obtained and regarded as a new feature (i.e., the so-called EMC features in [12]). The similar idea can be applied here. However, we can not obtain the projection of the feature vector $\mathbf{x}$ by computing $y_k = (\mathbf{w}_k \cdot \Phi(\mathbf{x}))$, $k=1,\ldots,K$, because the speaker eigenspace is constructed by models instead of features. To apply the Bhattacharyya kernel for feature transformation, we need to extend the feature vector $\mathbf{x}$ to a Gaussian $\mathbf{g}(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We can use the adaptation trick in the extreme case, which adapts a reference model from UBM with a single feature vector. Because UBM represents a distribution over a large space, a single feature vector will be close to only few mixture components of UBM. The likelihood values can be approximated well using only few best scoring mixture components [9]. For simplicity purpose, we choose the best mixture component for a feature vector $\mathbf{x}$ in adaptation, set the relevance factors to heavy emphasis this feature vector in the Maximum a Posteriori (MAP) formulation [9], and keep the covariance matrix unchanged, i.e., $\mathbf{g}(\mathbf{x}) \sim N(\mathbf{x}, \boldsymbol{\Sigma}_{r*})$, where $r*$ is the mixture component that has the maximal likelihood with respect to $\mathbf{x}$. Therefore, the projection of a feature vector $\mathbf{x}$ can be expressed as the projection of $\mathbf{g}(\mathbf{x})$ onto $\mathbf{w}_k$, i.e.,

$$\begin{aligned} y_k &= (\mathbf{w}_k \cdot \Phi(\mathbf{g}(\mathbf{x}))) \\ &= \sum_{r=1}^{R} \alpha_{rk} p_r (\Phi(\mathbf{g}_r) \cdot \Phi(\mathbf{g}(\mathbf{x}))) = \sum_{r=1}^{R} \alpha_{rk} p_r k(\mathbf{g}_r, \mathbf{g}(\mathbf{x})). \end{aligned} \tag{23}$$

In this way, we can obtain a K-dimensional new feature vector $\mathbf{y}^T = [y_1, \ldots, y_K]$ from a feature vector $\mathbf{x}$ using the GMM-based BKFDA. If C training speakers are available for eigenspace construction, K must be less or equal to C-1 because $\mathbf{N}^{-1}\mathbf{M}$ has at most C-1 eigenvectors. However, since the feature vector dimension d is usually much less than C, we can obtain at most d rather than C-1 eigenvectors when applying FDA in eigenspace construction. This is why FDA is widely used in reducing the dimension of feature vectors. In the proposed GMM-based BKFDA approach, the dimension of the new feature vector can be higher than that of the original feature vector.

4.3. GMM-based Speaker Identification

In this study, we apply the GMM-based BKFDA in GMM-based speaker identification. In the training phase, the MFCC features of each client speaker's enrollment data are first transformed into the BKFDA features, and then used to train the speaker GMM. In the test phase, the BKFDA features transformed from the MFCC features of test utterances are used for speaker identification evaluation.

5. EXPERIMENTS

5.1. Experimental Setup

The NIST 2001 cellular speaker recognition evaluation database [13] was used in the following experiments. We divided this database (including the development data and the evaluation data) into two subsets. The first subset consists of 90 female and 84 male speakers. It was used to train the UBM first. Then, the 174 speaker GMMs were adapted from the UBM using the speaker specific training data, respectively. Finally, the 174 speaker GMMs were applied in eigenspace construction. The second subset consisting of the remaining 22 females and 28 males was used for speaker identification evaluation. Each speaker has about 2 minutes of training data and 10 test segments on average.

Enrollment data	24-dim MFCC		12-dim BKFDA		24-dim BKFDA		40-dim BKFDA		50-dim BKFDA		60-dim BKFDA		70-dim BKFDA	
3 sec	31.44	4	**40.64**	2	40.13	2	39.97	4	39.13	8	38.13	8	35.79	8
6 sec	46.15	4	51.84	8	52.68	8	**53.68**	16	49.50	8	47.66	8	49.83	8
9 sec	56.86	8	53.68	8	58.03	16	**59.53**	8	59.03	16	59.20	8	58.03	16
15sec	62.88	8	61.04	16	65.39	16	**66.56**	16	65.89	16	66.05	16	65.55	8
30sec	67.22	16	63.55	64	68.56	16	71.74	32	72.07	32	**72.41**	32	71.91	32
45sec	69.40	32	66.89	32	71.41	64	73.58	32	**73.75**	32	73.24	32	72.41	64
60sec	72.58	32	67.73	64	73.08	32	**74.75**	16	74.58	64	**74.75**	32	**74.75**	16
ALL	73.91	64	70.40	64	74.58	32	75.08	64	**75.59**	64	75.08	64	**75.59**	64

Table 1*: Speaker identification accuracy (%) for MFCC-based GMMs and BKFDA-based GMMs.*

All the speech data were processed by a Voice Activity Detector (VAD) to discard silence-noise frames [14]. After passing VAD, about 4.22 hours of speech from 174 speakers in the first subset were used to train the UBM with 1024 mixture components, the speaker GMMs and the eigenspace. In the second subset, the training data of each client speaker lie within the range of 90~122 seconds and the duration of each test segment is 6~45 seconds. There are 598 test segments in total.

For both subsets, the speech was sampled at 8 kHz. Spectral analysis was applied to a 32 ms frame of speech waveform every 10 ms. For each speech frame, 12 MFCCs along with the first time derivatives were combined together to form a 24-dimensional feature vector.

5.2. Experimental Results and Discussions

For performance comparison, a baseline system built upon GMMs and MFCC features was evaluated first. Extensive experiments with respect to the number of Gaussian mixture components used in a GMM and the amount of enrollment data used for training a GMM have been run. For the sparse enrollment data cases ($\leq$15 seconds), GMMs with 2, 4, 8, and 16 mixture components were evaluated, while for the abundant enrollment data cases (>15 seconds), GMMs with 8, 16, 32, and 64 mixture components were evaluated. The accuracies associated with the empirically most accurate configuration (i.e., the most appropriate mixture number with respect to the amount of enrollment data) are summarized in the first column of Table 1, where "3 sec" means only the first 3 seconds of the training utterance were used for training the speaker GMM, while "All" means the complete training utterance was used. The remaining columns show the results of the BKFDA features with different K (i.e., the feature vector dimension). It is obvious that the performance of our approach is better than the conventional MFCC-based GMMs, in particular when the amount of enrollment data for each client speaker is very small (<9 seconds). When the enrollment data from a client speaker are relatively sufficient, say 60 seconds, our approach still performs better than the conventional MFCC-based GMM approach, though the improvement is not significant. The experimental results show that by using our approach, the amount of enrollment speech for a client speaker can be reduced.

6. CONCLUSIONS

In this paper, we want to find a speaker space that can better discriminate the speakers from each other. We propose the GMM-based KFDA with the Bhattacharyya kernel to obtain a transformation, or called a speaker eigenspace, based on which the transformed MFCC features are more discriminative for speaker recognition. The speaker identification experiments show that the BKFDA features outperform the MFCC features, in particular when the amount of enrollment data for each speaker is very small.

7. ACKNOWLEDGEMENTS

This work is financially supported by the National Science Council, Taiwan, under Grant NSC93-2413-E-001-017.

8. REFERENCES

[1] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, vol.17, pp. 91-108, 1995.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd. ed., Academic Press, 1990.

[3] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher Discriminant Analysis with Kernels," *Neural Networks for Signal Processing IX*, pp. 41-48, 1999.

[4] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol.2, pp. 121-167, 1998.

[5] S. Mika, "Kernel Fisher Discriminants," PhD thesis, University of Technology, Berlin, 2002.

[6] T. Jebara and R. Kondor, "Bhattacharyya and Expected Likelihood Kernels," *Proc. COLT 2003*.

[7] O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua, "Speaker Identification and Verification using Eigenvoices," *Proc. ICSLP2000*.

[8] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd. ed., John Wiley & Sons, New York, 2001.

[9] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[10] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Trans. on Speech Audio Process.*, vol.2, pp. 291-298, 1994.

[11] K.T. Chen, W.W. Liau, H.M. Wang and L.S. Lee, "Fast Speaker Adaptation using Eigenspace-based Maximum Likelihood Linear Regression", *Proc. ICSLP2000*.

[12] Nick J.-C. Wang, W.H. Tsai, and L.S. Lee, "Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification", *Proc .EUROSPEECH2001*.

[13] *http://www.nist.gov/speech/tests/index.htm*

[14] The VIMAS speech codec. *http://www.vimas.com*