A Discriminative Method for Speaker Identification with Limited Data

Lin lin, Chen jian, Sun xiaoying
College of Communication Engineering
Jilin University
Changchun, China
chenjian@jlu.edu.cn

***Abstract*—Speaker recognition system needs sufficient data to discriminate speaker well. In case of limited data, especially when the amount of available training and testing data were few seconds, the system performance decreased significantly. It proposed a discriminative weighted fuzzy kernel vector quantization method for speaker identification with limited data. By non-linear mapping, it quantized the input data in the high-dimensional feature space, and used the cluster centers to form the speaker's model. In the matching phase, it took into account the relationship between the reference models in feature space, and assigned the larger weights for code vectors with high discriminative power. Experimental results show that when the training data and testing data is limited, this method can provide good performance.**

Keywords-speaker recognition, discriminative weighted method, kernel method, fuzzy kernel vector quantization

I. INTRODUCTION

Speaker recognition is the process of automatically recognizing a speaker by using speaker-specific information included in speech waves. Usually, speaker recognition system needs sufficient data to discriminate speaker well, and yields good performance. In the present work, few minutes speech data (≥ one minute) is used to symbolize the speech data. When the amount of data available is about few seconds (≤10 seconds), the classic methods such as Gaussian mixture models (GMM) can't perform well due to sparse data. Although, the concept of Universal Background Model (UBM) has been used to mitigate the sparseness under limited data conditions, it requires additional speech data to train the GMM-UBM model [1, 2]. H. S. Jayanna [3] used Fuzzy Vector Quantization (FVQ) for speaker recognition with limited data, and FVQ showed significant improved performance compared to Direct Template Matching (DTM), and Crisp Vector Quantization (CVQ). Lin.L used kernel-based fuzzy c-means (KFCM) algorithm to design vector quantizer, and using the entropy function in the kernel mapping feature space to improve system performance [4].

It is well known that different phonemes have unequal discriminative power between speakers [5]. That is, the inter-speaker variation of certain phonemes is clearly different from other phonemes. In the methods correlative with VQ, different regions (clusters) represent acoustically different units. Thus, different code vector should have the different importance. However, the above methods don't consider the influence of each code vector. Therefore, in this paper, it proposed a discriminative weighted method for kernel-based speaker identification. By non-linear mapping, it quantized the input data in the high-dimensional feature space, and used the cluster centers to form the speaker's model. In the identification phase, it developed a maximum overall average weighted membership classifier. This matching method took into account the relationship between the reference models in the feature space, and assigned the larger weights for code vectors with high discriminative power. It does not require any a priori knowledge about the nature of the feature vectors, or any phonetic knowledge about the discrimination powers of the different phonemes. Instead, the method adapts to the statistical properties of the feature vectors in the given database.

II. FUZZY KERNEL VECTOR QUANTIZATION

Fuzzy kernel vector quantization (FKVQ) uses the fuzzy clustering and kernel method to design the vector quantizer. FKVQ can quantize data with diversiform structures, and provide an efficient way to describe the distribution of speakers' features.

Given a dataset, $X = \{x_1, x_2, \ldots x_N\} \in R^d$ define a nonlinear map as $\Phi : x_i \to \Phi(x_i), i = 1,2,\ldots, N$. FKVQ minimizes the following objective function

$$J_m^{\Phi}(X;U,V) = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^m \| \Phi(x_k) - \Phi(v_i) \|^2 \qquad (1)$$

where $U = [u_{ik}], i = 1,\ldots, c; k = 1,\ldots, N$ is the membership matrix, and $u_{ik} \in [0,1]$ represents the membership degree of x_k in cluster *i*. $V = \{v_1, v_2, \ldots v_c\}$, $v_i \in R^d$ is a *c*-tuple of cluster prototypes which have to be determined, and *m*>1 is a fuzzy index which determines the fuzziness of the clusters. According to the characteristic of Mercer kernel, the Euclidian distance between x_k and v_i in the feature space of the kernel *K* can be definition

$$\begin{aligned} d_K^2(x_k, v_i) &= \| \Phi(x_k) - \Phi(v_i) \|^2 \\ &= K(x_k, x_k) - 2K(x_k, v_i) + K(v_i, v_i) \end{aligned} \qquad (2)$$

It considers the objective function can be

$$J_m^{\Phi}(X;U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m} d_K^2(x_k, v_i)$$

$$= \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m}[K(x_k,x_k) - 2K(x_k,v_i) + K(v_i,v_i)] \qquad (3)$$

The membership function in high feature space should satisfy

$$u_{ik} = (1/d_K^2(x_k,v_i)^{1/(m-1)}) / \sum_{l=1}^{c}(1/d_K^2(x_k,v_l)^{1/(m-1)}) \qquad (4)$$

The new cluster prototypes in the feature space R_q can be computed

$$\Phi(\hat{v}_i) = \sum_{k=1}^{N} u_{ik}^{m}\Phi(x_k) / \sum_{k=1}^{N} u_{ik}^{m} \qquad i=1,2,\ldots,c \qquad (5)$$

Then

$$K(x_k,\hat{v}_i) = \Phi(x_k)\bullet\Phi(\hat{v}_i) = \sum_{j=1}^{N} u_{ij}^{m} K(x_j,x_k) / \sum_{j=1}^{N} u_{ij}^{m} \qquad (6)$$

$$K(\hat{v}_i,\hat{v}_i) = \Phi(\hat{v}_i)\bullet\Phi(\hat{v}_i) = \sum_{k=1}^{N}\sum_{l=1}^{N} u_{ik}^{m}u_{il}^{m} K(x_k,x_l) / |\sum_{j=1}^{N} u_{ij}^{m}|^2 \quad (7)$$

From (6) and (7), the new membership function $\hat{u}_{ik}$ in the feature space can be rewritten as

$$\hat{u}_{ik} = \frac{(1/d_K^2(x_k,\hat{v}_i))^{1/(m-1)}}{\sum_{l=1}^{c}(1/d_K^2(x_k,\hat{v}_l))^{1/(m-1)}}$$

$$= \frac{(1/(K(x_k,x_k) - 2K(x_k,\hat{v}_i) + K(\hat{v}_i,\hat{v}_i)))^{1/(m-1)}}{\sum_{l=1}^{c}(1/(K(x_k,x_k) - 2K(x_k,\hat{v}_l) + K(\hat{v}_l,\hat{v}_l)))^{1/(m-1)}} \qquad (8)$$

For the speaker recognition, after clustering, compute the new centers as the speakers' model.

$$\hat{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \qquad (9)$$

The FKVQ algorithm to design the codebook in the high dimensional feature space is following:

1） Choose c , termination criterion $\varepsilon \in (0,1)$;

2） Choose kernel function K and its parameters;

3） Initialize cluster centers $v_i, i = 1,2,...,c$;

4） Use (6) and (7) to calculate the kernel functions $K(x_k,\hat{v}_i)$ and $K(\hat{v}_i,\hat{v}_i)$;

5） Use (8) to compute the degree of membership of all feature vectors in all the clusters $\hat{u}_{ik}, i = 1,...,c; k = 1,...,N$;

6） If $\max_{j,i}|u_{ij}(t-1) - u_{ij}(t)| < \varepsilon$ continue, otherwise, go to step 3） ;

7） Use (9) to compute the speaker's codebook and stop.

III. Speaker Discriminative Matching in Feature Space

A. Maximum overall average weighted membership classifier

A classifier for speaker recognition must decide to which speaker the sequence X belongs, therefore a decision made for the sequence X can't be fuzzy [5]. We can describe the belonging that a test vector x_k belongs to several codebooks by a fuzzy membership function. Let $X = \{x_1, x_2, \ldots x_N\} \in R^d$ be the sequence of test vectors and $v_i(j)$, i= 1,…,c, be the code vectors in the jth codebook, j=1,...,M, where M is the number of speaker. A classifier of N test vectors x_k in the sequence X into M codebooks can be described by a $M \times N$ matrix U, whose k, jth entry, $\mu_k(j)$ is the fuzzy membership of the vector x_k with the codebook V_j and satisfies

$$0 \le \mu_k(j) \le 1 \ ; \ \sum_{j=1}^{M}\mu_k(j) = 1 \qquad (10)$$

The membership is used to identify the concentration of the test vector x_k in the codebook V_j. If we define a fuzzy conditional risk function $h_k(j)$=1-$\mu_k(j)$, then in order to achieve the minimum error rate, we should make a decision that the speaker j is correct if the membership function $\mu_k(j)$ is a maximum. Define the overall average of fuzzy membership function of the codebook V_j with the degree of fuzziness m as follows

$$F(j) = \frac{1}{N}\sum_{k=1}^{N}\mu_k{}^{m}(j) \ , \quad j = 1,...,M \qquad (11)$$

The fuzzy membership of the vector x_k with the codebook V_j is

$$\mu_k(j) = (1/d^2(x_k,V_j)^{1/(m-1)}) / \sum_{l=1}^{M}(1/d^2(x_k,V_l)^{1/(m-1)}) \qquad (12)$$

where the distortion $d(x_k, j)$ between x_k and the jth codebook should be

$$d(\mathbf{x}_k,\mathbf{V}_j) = \min_{1\le i\le c} d(\mathbf{x}_k,\mathbf{v}_i(j)) \qquad (13)$$

To take into account the relationship between the reference models in the feature space, the overall average weighted membership function can be defined as

$$F_w(j) = \frac{1}{N}\sum_{k=1}^{N} w_{NN[x_k]}^{\Phi}\mu_k{}^{m}(j) \ , \quad j = 1,...,M \qquad (14)$$

Here $w_{NN[x_k]}^{\Phi}$ is the weight associated with the nearest code vector in the feature space. The product $w_{NN[x_k]}^{\Phi}\mu_k{}^{m}(j)$ can be viewed as a local operator that moves the decision surface towards more significant code vectors.

Using $F_w(j), j = 1,...,M$ in (14) as discriminant functions, the decision rule can be named the maximum overall average weighted membership classifier. For the sequence X, a decision rule is stated as follows

$$result = \arg\max_{1\le j\le M}(F_w(j)) \qquad (15)$$

B. Assigning the weights

Let $v_i(j) \in V_j$ be a code vector of the *j*th speaker, and in the high dimensional feature space, it is denoted by $\Phi(v_i(j))$. The codebooks are post-processed to assign weights for the code vectors, and the result of the process is a set of weighted codebooks $(V_j, W_j), j = 1,...,M$, where the weights assigned for the *j*th codebook are $W_j = \{w(\Phi(v_1(j))),...,w(\Phi(v_c(j)))\}$. In this way, the weighting approach does not increase the computational load of the matching process, as it can be done in the training phase when creating the speaker database.

In the paper, the weights of a code vector depend on the minimum distances generated from the code vectors of the other classes in the feature space. Let us denote the index of its nearest neighbour in the *l*th codebook simply by *NN* $^{(l)}$. The weight of $\Phi(v_i(j))$ is then assigned as follows

$$\begin{aligned} w(\Phi(v_i(j))) &= \frac{1}{\sum_{l\neq j}(1/d(\Phi(v_i(j)),\Phi(v_{NN^{(l)}})))} \\ &= \frac{1}{\sum_{l\neq j}(1/d_K(v_j(j),v_{NN^{(l)}}))} \qquad (16) \end{aligned}$$

where $d_K(v_j(j), v_{NN^{(l)}})$ can use (2) to calculate.

In other words, nearest code vector from all other classes are found, and the inverse of their distances reciprocal sum is taken. If some of the distances equal to zero, set $w(\Phi(v_i(j))) = 0$ for mathematical convenience [6]. The algorithm is looped over all the code vectors and all codebooks.

IV. Experimental Results

Experiments are based on the PKU-SRSC [7]. It chooses 40 speakers randomly from the database, including 20 males and 20 females. Choose 6 sessions to do the experiments. Some utterances in first session are used to train the speaker models and the other sessions are as the test data for the identification evaluation. The test utterances of each speaker are about one second. Speech utterances are parameterized with *Mel* frequency cepstral coefficients (MFCC). Speech signal is pre-emphasized using a coefficient of 0.95. Each frame of speech is windowed by a Hamming window and represented by a 39 dimensional feature vector, including 20 MFCC and its first differentials, in which the first coefficient is discarded. The initial codebook is designed by split method.

To demonstrate the validity of the discriminative weighted method in feature space, compare the error rate of FKVQ (Fuzzy Kernel Vector Quantization) and DFKVQ (Discriminative Fuzzy Kernel Vector Quantization) algorithm proposed in this paper. It uses five seconds training data of each speaker. Speaker model is trained with codebook volume of 8, 16, 32, and 64. The results are shown in Fig.1.

In Fig.1, the error rate of DFKVQ is less than FKVQ for different codebook volume. It can be seen that the discriminative weighted method in feature space increases the dissimilarity among different speakers and improves the system performance.

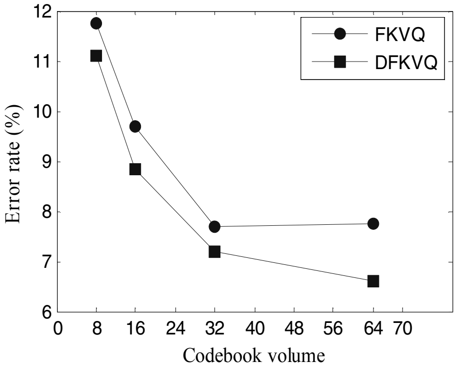

Figure 1. Error rate of FKVQ and DFKVQ

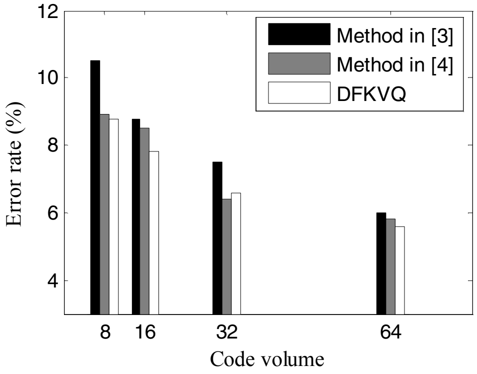

Figure 2. Error rate of DFKVQ and two methods in [3] and [4]

Compare the performance of DFKVQ and the two methods in [3] and [4]. The data is the same as test 1. Speaker model is trained with codebook volume of 8, 16, 32, and 64. The results are shown in Fig.2. In Fig.2, the error rate of DFKVQ is less than method in [3] for different codebook volume. Compared with method in [4], except codebook volume of 32, DFKVQ can get the lower error rate for other three codebook volume. From the result we can see that the discriminative weighted method adapts to the statistical properties of the feature vectors in the given database, and improves the system performance.

In order to discuss the effectiveness of the DFKVQ method with limited data, it uses 3 to 8 seconds training data to train speaker's model, and here we set codebook volume to be 8,16,32,64. The results are shown in Fig.3.

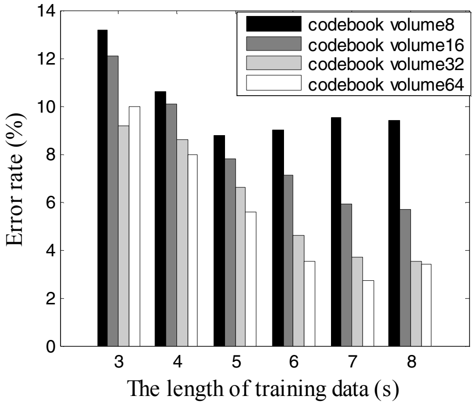

Figure 3. Error rate of DFKVQ with different training data length

From the Fig.3, we can see that when the length of the training data is constant, the error rate decreases with the increase of the codebook volume. However, in case of training data of 3s, the large codebook volume can't describe the small amount of the training data accurately. So when the training data is 3s, the error rate of codebook volume 32 is less than that of codebook volume 64. When it used codebook volume of 64, seven seconds training data, and one second testing data, the error rate of system is 2.7%. From the result, it can be seen that when the length of training data is less 8 seconds, and testing data is one second, DFKVQ can get good identification performance.

V. CONCLUSION

In this paper, it proposes a discriminative weighted method for kernel-based speaker identification. Because the kernel mapping, it develops a discriminative weighted matching method in the high dimensional feature space, which takes into account the relationship between the known models in the database and assigned the larger weights for code vectors with high discriminative power. Compared with FKVQ, and the two methods in [3] and [4], the DFKVQ algorithm proposed in this paper can get best performance for different codebook volume From the result it can be seen that discriminative weighted method can increase the dissimilarity among different speakers. To discuss the effectiveness of DFKVQ algorithm with limited data, it uses 3 to 8 seconds training data to train speaker's model, and 1 second to test. The results show that when the codebook volume is 64, the error rate of DFKVQ algorithm is 2.7%. From this study we may conclude that when the training and testing data is limited, the DFKVQ algorithm yields good recognition performance.

REFERENCES

[1] P. Angkititrakul and J. H. L. Hansen, "Discriminative In-Set/Out-of-Set Speaker Recognition," IEEE Trans. Audio Speech Language Processing, vol. 15(2), pp. 498-508, Feb. 2007.

[2] V. Prakash and J. H. L. Hansen, "In-Set/Out-of-Set Speaker Recognition Under Sparse Enrollment ," IEEE Trans. Audio Speech language Processing, vol. 15(7), pp. 2044-2051, Sep. 2007.

[3] Jayanna, H.S., Prasanna, S.R. Mahadeva. "Fuzzy Vector Quantization for Speaker Recognition under Limited Data Conditions". IEEE Region 10 Conference , TENCON, pp. 1-4, 2008

[4] Lin lin, Wang shuxun, Chen jian, "Speaker Recognition with Little Data Based on Fuzzy Kernel Entropy". Journal of System Simulation, vol. 20(16), pp. 4368-4372,Aug. 2008.

[5] D. Tran, M. Wagner, T. Van Le, "A Proposed Decision Rule for Speaker Recognition Based on Fuzzy C-Means Clustering", in Robert H Mannell and Jordi Robert-Ribes (ed), 5th International Conference on Spoken Language Processing, ICSLP '98, pp 755-758, 1998

[6] Kinnunen, T., Franti, P. "Speaker discriminative training algorithm for VQ-based speaker identification". Proc.3rd International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA).Halmstad, Sweden, pp. 150-156, 2001.

[7] Xihong Wu, "A Chinese speech corpora for speaker recognition", http://nlprweb.ia.ac.cn/english/irds/chinese/SinobiometricsPDF/Wuxihong.pdf, 2002.