Speaker Verification Using MFCC and Support Vector Machine

Shi-Huang Chen and Yu-Ren Luo

***Abstract* —This paper proposes a study on the use of mel-frequency cepstral coefficients (MFCC) and support vector machine (SVM) for text-dependent speaker verification. The MFCCs used in this paper are extracted from the voiced password spoken by the user. These MFCCs will be normalized and then can be used as the speaker features for training a claimed speaker model via SVM. Finally one could make use of the claimed speaker model to discriminate between the speaker and other impostors. Experiments were conducted on the Aurora-2 database with various orders of MFCCs. It follows from the experimental results that the proposed text-dependent speaker verification system based on the 22th-order MFCCs and SVM gives an equal error rate (EER) of 0.0% and average accuracy rate of 95.1%.**

***Keywords*— Speaker recognition, speaker verification, MFCC, SVM.**

I. Introduction

The speaker verification is regarded as a subcategory of automatic speaker recognition (ASR) system and can apply to determine whether a person is who he/she claims to be. Therefore, the problem of speaker verification is a true-false (accept-reject) question [1-6]. The speaker verification is desirable widely in many speech related applications, such as banking by telephone, voice dialing, and biometric security system [1-6]. Meanwhile, depended on the differences of recognition target, the systems of speaker verification fall into two types: text-dependent and text-independent. The former one requires that the speaker should provide keywords or sentences of the same text for both training and recognition, while the latter one dose not depend on the specific text being spoken [1-6]. For security consideration, this paper will focus on the problem of the text-dependent speaker verification.

Several methods have been proposed for speaker verification. It follows from [1-6] that a typical speaker verification system consists of two tasks: enrollment and verification as shown in Fig. 1. Enrollment is the task to construct a speaker model. This step will capture the speaker characteristics or features. Most of the present-day systems use the speaker-specific vocal tract information like MFCCs or linear prediction coefficients (LPCs) as speaker features for speaker verification. Then these speaker features are used to build a model that could authenticate the speaker during the verification phase. In the speaker verification task, the speaker features of the input speech from test subject will be extracted and matched against the speaker model. A likelihood ratio will evaluate the similarity between the model and the measured observations. The general approach is based on a threshold set for the acoustic likelihood ratio to decide the test speaker is accepted or rejected. Conventional speaker verification systems use hidden Markov models (HMM) or Gaussian mixture model (GMM) to perform the likelihood ratio test [1-6]. These systems make use of a generative model for all speaker models. This will result in over-fitting and maybe cannot maximize the discrimination of speaker and impostors.

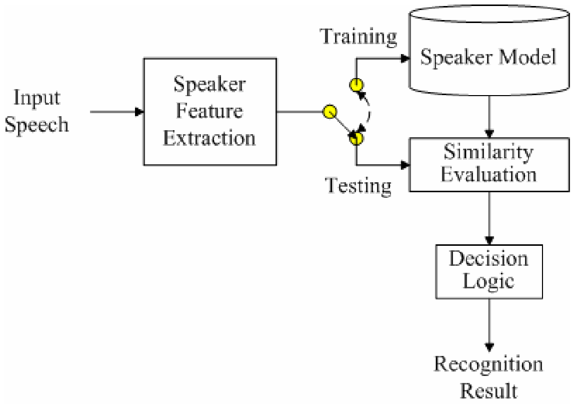

Fig. 1. The typical speaker verification system.

The choice of speaker features is another primary concern in the development of a speaker verification system. The ideal speaker features set should have higher inter-class variance and lower intra-class variability. In addition, the selected speaker features should be independent of each other as in order to minimize redundancy. Based on the above discussion, the goal of this paper is to develop a more efficient approach to the text-dependent speaker verification using MFCCs and SVM. Previous researches [4-6] have shown that MFCCs can represent detail characteristics of individual speakers and therefore are mostly usable features for speaker verification. On the other hand, SVM is a two-class classifier based on the principles of structural risk minimization. It is shown that SVM has well generalization ability when compared to hidden Markov model and neural network based classifier [7]. Furthermore, since speaker verification is basically a binary decision, SVM seems to be a promising candidate to perform this task.

In this paper, various orders of MFCCs are used as speaker features to perform speaker verification. In the beginning, the user has to provide a voiced password and the corresponding MFCCs will be extracted from this spoken password. Then the proposed text-dependent speaker verification system will

This work was supported in part by the National Science Council under Grant NSC 97 -2221 -E -366 -010 -MY3. Shi-Huang Chen* and Yu-Ren Luo are with the Department of Computer Science Information Engineering, Shu-Te University, Kaohsiung, 824, Taiwan. (*Corresponding author, Tel: +886-7-6158000 ext. 4607, Fax: +886-7-6152559, e-mail: shchen@mail.stu.edu.tw)

make use of SVM to train the speaker features from these MFCCs and generate a speaker model to discriminate between the speaker and other impostors. Using speech signals selected from the Aurora-2 database, experimental results shown the performance of the proposed speaker verification algorithm yields an equal error rate (EER) of 0% and average accuracy rate of 95.1% with 22-order MFCCs.

The remainder of this paper is organized as follows. The introductions to MFCCs and SVM are briefly reviewed in Sections II and III, respectively. Section IV will describe the proposed text-dependent speaker verification system. Section V illustrates the various experimental results with different orders of MFCCs. Finally, conclusions are given in Section VI.

II. MFCC

It is shown that MFCC can capture the acoustic characteristics for speech recognition, speaker recognition, and other speech related applications [4-8]. According to psychophysical studies, human perception of the frequency content of sounds follow a subjectively defined nonlinear scale called the "mel" scale [9] defined as,

$$f_{\text{mel}} = 1125\ln(1+\frac{f}{700}) \qquad (1)$$

where f is the actual frequency in Hz. This leads to the definition of MFCC and its calculation process is given as follows.

Let $s(n)$, n = 1~N, be a speech frame that is pre-emphasized and Hamming-windowed [8, 9]. First, the time domain signal, $s(n)$, is transferred into frequency domain by an M point discrete Fourier transform (DFT). The resulting energy spectrum can be represented as

$$|S(k)|^2 = \left|\sum_{n=1}^{M} s(n)\cdot e^{\left(\frac{-j2\pi nk}{M}\right)}\right|^2 \qquad (2)$$

where $1 \le k \le M$. Then, the triangular filter banks, whose frequency bands are linearly spaced in the mel scale, are imposed on the spectrum obtained in (2). The outputs $e(l)$, l = 1~Q, of the mel-scaled band-pass filters can be calculated by a weighted summation between respective filter response $H_i(k)$, i = 1~M, and the energy spectrum $|S(k)|^2$ as

$$e(i) = \sum_{k=1}^{M} |S(k)|^2 \cdot H_i(k) \qquad (3)$$

where $H_i(k)$ is defined as

$$H_i(k) = \begin{cases} 0, & \text{for} \quad k < f_{b(i-1)} \\ \dfrac{(k - f_{b(i-1)})}{(f_{b(i)} - f_{b(i-1)})}, & \text{for} \quad f_{b(i-1)} \le k < f_{b(i)} \\ \dfrac{(f_{b(i+1)} - k)}{(f_{b(i+1)} - f_{b(i)})}, & \text{for} \quad f_{b(i)} \le k < f_{b(i+1)} \\ 0, & \text{for} \quad k > f_{b(i+1)} \end{cases} \qquad (4)$$

In (4), $f_{b(i)}$ are the boundary points of the filters and are depended on the sampling points Fs and the number of points N in DFT:

$$f_{b(i)} = \left(\frac{N}{F_s}\right)\cdot f_{\text{mel}}^{-1}\left(f_{\text{mel(low)}} + i\,\frac{f_{\text{(mel)high}} - f_{\text{(mel)low}}}{M+1}\right). \qquad (5)$$

Here, $f_{\text{mel(low)}}$ and $f_{\text{mel(high)}}$ are respectively the low and high boundary frequencies for the entire filter bank. f_{mel}^{-1} is the inverse to (1) transformation, formulated as

$$f_{\text{mel}}^{-1} = 700\left[e^{\left(\frac{f_{mel}}{1125}\right)} - 1\right] \qquad (6)$$

Fig. 2 show the mel space filter bank with M=40 [10].

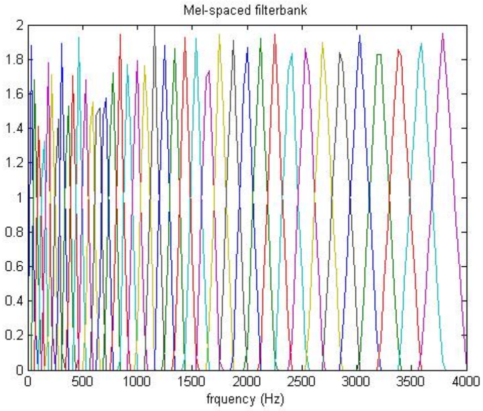

Fig. 2. Mel-space filter bank (M=40) [10].

Finally, discrete cosine transform (DCT) is taken on the log filter bank energies, log[$e(l)$], and the MFCC coefficients C_m can be written as,

$$C_m = \sqrt{\frac{2}{Q}}\sum_{p=0}^{Q-1}\log[e(p+1)]\cdot\cos\left[m\cdot\left(\frac{2p-1}{2}\right)\cdot\frac{\pi}{Q}\right] \qquad (7)$$

where $0 \le m \le M$-1. Fig. 3 shows the summary of MFCC calculation process.

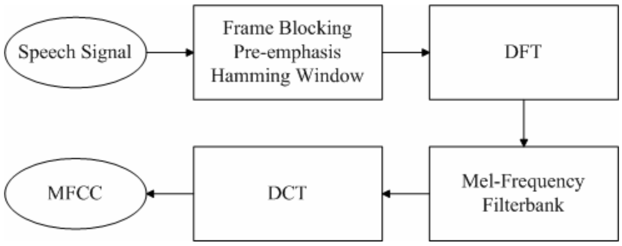

Fig. 3. The block diagram of MFCC calculation process.

III. Support Vector Machine

An SVM is a two-class classifier constructed from sums of a known kernel function $K(\cdot,\cdot)$ to define a hyperplane.

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \qquad (8)$$

where $y_i \in \{1,-1\}$ are the target values, $\sum_{i=1}^{N}\alpha_i y_i = 0$, and $\alpha_i > 0$. The vector $\mathbf{x}_i \subseteq R^n$ are support vectors and obtained from the training. This hyperplane will separate given points into two predefined classes. Suppose a training set $S = \{(x_1, y_1), \cdots, (x_l, y_l)\}_{i=1}^{l} \subseteq (X \times Y)^l$ and a kernel function $K(x_i, x_j) = <\phi(x_i), \phi(x_j)>$ on $X \times X$ is given, where $<\cdot,\cdot>$ denotes the inner product and ϕ maps the input space X to another high dimensional feature space F. With suitably chosen ϕ, the given nonlinearly separable samples S may be linearly separated in F, as shown in Fig. 4. An improved SVM called soft-margin SVM can tolerate minor misclassifications [4] and use in this paper.

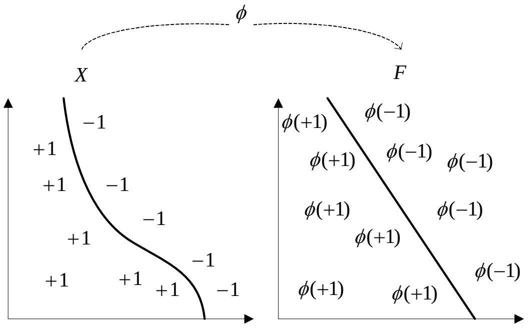

Fig. 4. A feature map simplifies the classification task.

Many hyperplanes can achieve the above separation purpose but the SVM used in this paper is to find the one that maximizes the margin (the minimal distance from the hyperplane to each points). The soft-margin SVM, which includes slack variables $\xi_i \geq 0$, is proposed to solve non-separable problems. Fig. 5 shows the slack variables, where ξ_i is defined as

$$\xi_i = \max(0, \gamma - y_i(< w, x_i > + b)). \tag{9}$$

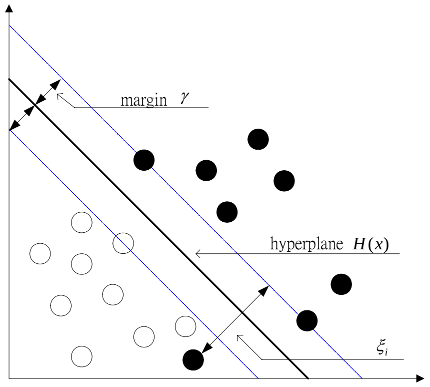

Fig. 5. The margin and the slack variable for a classification problem.

In (9), the parameter ξ_i can measure the amount by which the training set fails to have margin γ, and take into account any misclassification of the training data. Consequently, the training process tolerates some points misclassified and is suitable in most classification cases.

There are three common kernel functions for the nonlinear feature mapping, shown in Fig. 3: (1) exponential radial basis function (ERBF) $K(x,\bar{x}) = \exp(-|x-\bar{x}|/2\sigma^2)$, (2) Gaussian function $K(x,\bar{x}) = \exp(-|x-\bar{x}|^2/2\sigma^2)$, where parameter σ is the width of the Gaussian function, and (3) polynomial function $K(x,\bar{x}) = (< x,\bar{x} > +1)^d$, where parameter *d* is the degree of the polynomial. It is shown in [7] the ERBF has better results in speaker classification task. This paper will select ERBF as SVM kernel.

IV. The Proposed Speaker Verification System

Fig. 6 shows the block diagram of the text-dependent speaker verification system proposed in this paper. Before performing speaker verification, one has to build a claimed speaker model and an imposter model via SVM training. The training procedure is described as follows. Assume that n_T is the number of the obtained MFCC vectors. The training set *T* is then defined to be the $n_T \times p$ array with row vectors being these *p*-order MFCC vectors. In this paper, 13 different settings of *p* are evaluated, they are 2-, 4-, 6-, …, 22-, 24-, and 26-order MFCC. The next section will discuss the performances of these 13 settings of *p*. Let *T*(*i*, *j*)denote the (*i*, *j*)-position of *T*. Use this array *T* to construct another $n_T \times p$ array *T'* whose (*i*, *j*) position *T'*(*i*, *j*) is defined to be $T'(i,j) = T(i,j) - \mu_j$, where $\mu_j = \sum_i T(i,j)/n_T$ is the mean of column *j*. Next, one can normalize *T'* by computing $T^N(i,j) = T'(i,j)/m_j$, where m_j is the maximum of the absolute value of elements in column *j*. Thus, each MFCC feature will have similar weights after the normalization process.

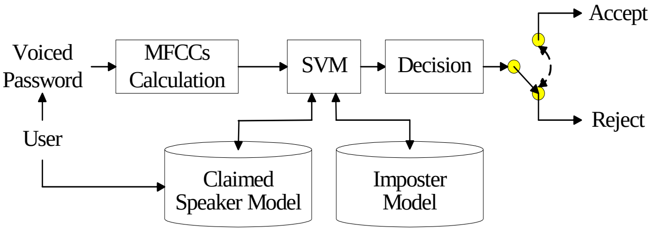

Fig. 6. The block diagram of the proposed text-dependent speaker verification system.

To train a model for a specific speaker, this paper utilizes a SVM classification method called one-against-all strategy. That is the speaker data are trained to an SVM target value of +1 whereas the imposter data are trained to an SVM target value of −1. Finally SVM will find a linear hyperplane that can separate speaker and imposter MFCCs features. The selection of MFCC orders will be discussed in the next section.

V. Experimental Results

The experimental results of the proposed text-dependent speaker verification system are achieved by using 20 male and 20 female speakers selected from the Aurora 2 database [11]. All of the test speech signals are noisy-free and are sampled at 8000 Hz with 16-bit resolution. Each test speech signal consists of 2~8 English digital numbers or English alphabets. Speaker verification performance will be reported using the false acceptance rate (FAR), the false rejection rate (FRR), and the equal error rate (EER).

The definitions of FAR and FRR are given as follows:

$$\text{FAR} = \frac{\#\,\text{acepted imposter claims}}{\#\,\text{imposter accesses}} \times 100\% \tag{10}$$

$$\text{FRR} = \frac{\#\,\text{rejected genuine claims}}{\#\,\text{genuine accesses}} \times 100\% \tag{11}$$

Once the receiver operating characteristic (ROC) curve of FAR vs. FRR is obtained, one can determine the EER, which FAR and the FRR at this point is the same for both of them.

In this paper, the different settings of MFCC order are studied experimentally for speaker verification. It follows from [8] that the higher-order MFCC does not further reduce the error rate in comparison with the lower-order MFCC. Hence, this paper compared the results obtained on the SVM based speaker verification system with 13 settings of MFCC order, namely $p= 2q$, $q = 1\sim13$. An impostor model was trained on all the MFCCs in the impostor data set while the speaker model was built using the corresponding speaker data set. During speaker verification task, a likelihood ratio was computed between the speaker model and the impostor model. The likelihood ratio was defined as:

$$LR = \log P(x \mid \text{speaker model}) \\ - \log P(x \mid \text{impostor model}) \qquad (12)$$

where x is the input test MFCCs vector. Table 1 shows a summary of the experimental results of the proposed text-dependent speaker verification systems. It follows from Table 1 that the better performance could be obtained when MFCC order $p = 22$. An EER of 0% and average accuracy rate of 95.1% are achieved using the proposed system. The ROC plots of FRR and FAR with MFCC order = 10 and 22 are shown in Figs. 7 and 8, respectively.

Table 1. Comparison of SVM based text-dependent speaker verification system with different MFCC orders.

MFCC order	Average accuracy rate	EER
2	72.1%	12.2%
4	83.9%	5.8%
6	86.7%	2.2%
8	87.7%	2.7%
10	90.7%	2.0%
12	92.5%	0.7%
14	93.1%	1.3%
16	94.0%	0.4%
18	94.4%	0.0%
20	94.7%	0.2%
22	95.1%	0.0%
24	95.0%	0.0%
26	94.8%	0.4%

VI. VI. CONCLUSIONS

In this paper, the mel-frequency cepstral coefficients (MFCC) and support vector machine (SVM) are applied to the task of text-dependent speaker verification system. First, the MFCCs will be extracted from the voiced password provided by user. Then the proposed algorithm will make use of SVM to train the speaker characteristics model from these MFCCs and result in a claimed speaker model that can discriminate between the speaker and other impostors. Various experiments were conducted on the Aurora-2 database and shown the performance of the proposed algorithm yields an equal error rate (EER) of 0.0% and average accuracy rate of 95.1% with 22-order MFCCs.

REFERENCES

[1] Peter Day and Asoke K. Nandi, "Robust Text-Independent Speaker Verification Using Genetic Programming," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 1, pp. 285-295, 2007.

[2] Minho Jin, Frank K. Soong, and Chang D. Yoo, "A Syllable Lattice Approach to Speaker Verification," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2476-2484, 2007.

[3] A.E. Rosenberg, "Automatic speaker verification: A review," *IEEE Proceedings*, Vol. 64, pp. 475-487, 1976.

[4] Guiwen Ou and Dengfeng Ke, "Text-independent speaker verification based on relation of MFCC components," 2004 International Symposium on Chinese Spoken Language Processing, pp. 57-60, Dec. 2004.

[5] A. Mezghani and D. O'Shaughnessy, "Speaker verification using a new representation based on a combination of MFCC and formants," 2005 Canadian Conference on Electrical and Computer Engineering, pp. 1461-1464, May 2005.

[6] M.M Homayounpour and I. Rezaian, "Robust Speaker Verification Based on Multi Stage Vector Quantization of MFCC Parameters on Narrow Bandwidth Channels," ICACT 2008, vol 1, pp.336-340, Feb. 2008

[7] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 5, pp. 644-651, Sept. 2005.

[8] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993

[9] S. B. Davis and P. Mermelstein, "Comparison of Parametric Repre- sentation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. On ASSP, vol. ASSP 28, no. 4, pp. 357-365, Aug. 1980.

[10] http://www.softwarepractice.org/wiki/Team_D_Speaker_Recognition (MediaWiKi, Term D Speaker Recognition)

[11] http://www.elda.org/article52.html. **(**Aurora Database 2.0**)**

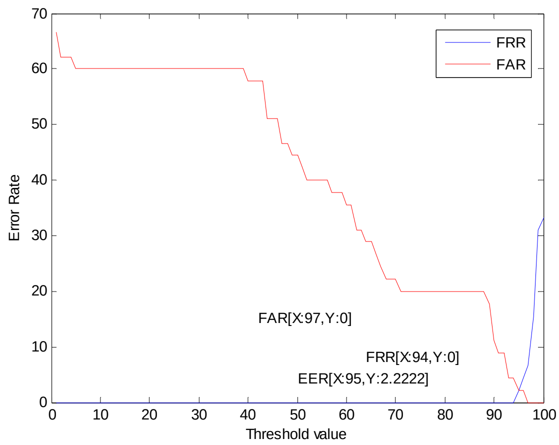

Fig. 7. A ROC plot of FRR and FAR with MFCC order = 10

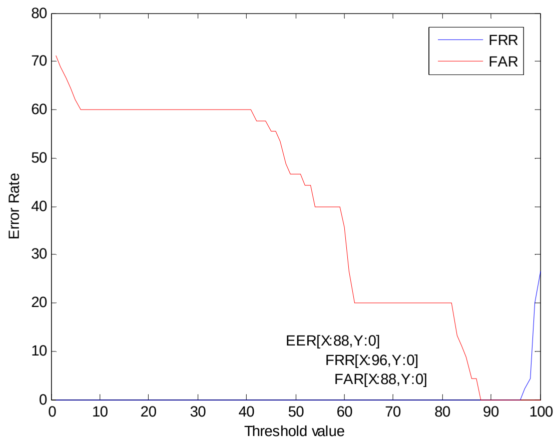

Fig. 8. A ROC plot of FRR and FAR with MFCC order = 22.