Efficient Speaker Recognition based on Multi-class Twin Support Vector Machines and GMMs

Hanhan Cong,Chengfu Yang,Xiaorong Pu
Computational Intelligence Laboratory
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, P. R. China.
E-mail: {Candicecong, ycfwsm}@gmail.com,puxiaorg@uestc.edu.cn

***Abstract*— This paper proposes a new approach for text-independent speaker recognition using Twin Support Vector Machines (TWSVMs) and feature extraction based on Gaussian Mixture Models (GMMs). Because of the perfect discriminability and the ability of managing large scale dataset, the proposed approach performs better than the traditional Support Vector Machines (SVMs) on Ahumada Biometric Database and Gaudi Biometric Database.**

I. Introduction

The goal of the speaker recognition is to decide which person is talking from a group of known speakers. The generative model and the discriminnative model are two main models for speaker recognition. The generative models such as Gaussian Mixture Model (GMM) have become the dominant modeling approaches in text-independent speaker recognition for its scalability and robustness. On the other hand, because discriminnative models such as Support Vector Machine (SVM)[3] have perfect discriminability, they have excellent performances in text-independent speaker recognition[7]. SVMs are model-free methods that do not make any distributional assumptions about the data. And at the same time, SVMs offer a discriminative solution to classification problems with strong bounds on error minimization. The last decade has winessed SVMs as a powerful paradigm for speaker recognition[5]. And it can achieve a generalisation performance that is better than or equal to other classifiers[11]. SVMs have been applied to speaker recognition in several instances. For example, in 2001, Campbell et al.[13] demonstrated an SVM-based approach analog to the traditional approach of modeling cepstral features with Gaussian mixture models (GMMs). However, SVMs become inefficient when the the number of training patterns is large[6].

In order to overcome the limitation of the traditional SVMs, various approaches have been suggested in recent years. One of the approaches is combining the advantages of the SVMs and that of the GMMs for speaker recognition. There are three main methods. Firstly, Fisher kernel which is based on GMMs is used in SVMs[16]. Secondly, SVMs are used as postprocessing modules for GMMs scores[14]. The Last, an SVM classifier was employed as an advisor to the GMM classifier in uncertain cases[15].

Recently, Jayadeva and R. Khemchandani proposed a nonparallel plane classifier for binary data classification. They termed it as Twin Support Vector Machines (TWSVMs)[1]. This algorithm aims at generating two nonparallel planes such that each plane is closer to one of the two classes while as far as possible from the other. The formulation of TWSVMs is similar with the traditional SVMs. The TWSVMs solve a pair of Quadratic Programming Problems (QPPs)[1], while the traditional SVMs solve a single QPP[9]. In SVMs, the QPP has all data points in the constraints, but, in TWSVMs, they are distributed in the sense that patterns of one class give the constraints of the other QPP and vice versa. In this way, TWSVMs can deal with larger number of patterns more rapidly.

In this paper, a new approach that uses TWSVMs is proposed for text-independent speaker recognition. It is also a combination of the generative model and the discriminnative model. But it is different from above hybrid approaches. Firstly, the proposed method extracts features from the training data based on GMMs[2]. There are two advantages. One is reducing the size of the training database. The other is that the final features reflect the human speech production process. Then TWSVMs models are trained with the features extracted by GMMs. In our approach, one TWSVMs model corresponds to one speaker. The number of the TWSVMs models is same to the number of speaker in the training database. Since the approach reduces the number of features, it's more efficient for large scale dataset than traditional SVMs. Excellent experimental results also show the success of our method for speaker recognition.

The rest of the paper is organized as follows: Section 2 outlines two traditional methods: the GMMs and the SVMs for speaker recognition. Section 3 describes the proposed approach in detail. The proposed approach is formulated in two steps. Firstly, features are extracted based on GMMs. Secondly, TWSVMs models are trained. In this section, the process of our method is also outlined. In Section 4 experimental results of the approach are compared with the traditional SVMs and GMMs approaches. The results indicate efficiency of the proposed method. Section 5 contains concluding remarks.

II. Traditional Methods for Speaker Recognition

A brief outline of two main methods for speaker recognition are given in this section. The Gaussian Mixture Model (GMM) is a statistic model and the Support Vector Machine (SVM) is

978–1–4244–1676–9/08/$25.00 © 2008 IEEE

a discriminative model.

A. Gaussian Mixture Model

GMM has become the dominant approach for modeling in text-independent speaker recognition applications over the past several years[4]. This is a probability density function that itself consist of a sum of multivariate Gaussian density functions. In practice, each speaker has a GMM that is trained for them individually and the likelihoods generated from the GMM form the basis for generating the speaker scores from which a decision is made with regard to identity of a speaker.

The probability density function of k Gaussian probability density function is given by

$$p(x_t|\lambda) = \sum_{i=1}^{k} \omega_i p[x_t|\mu_i, \Sigma_i], \tag{1}$$

where x_t is a n dimensional vector, $\omega_i (i = 1, 2, \ldots, k)$ is the mixture weight and $p[x_t|\mu_i, \Sigma_i]$ is the component density. Each component density is n-variate Gaussian function of the form

$$p[x_t|\mu_i, \Sigma_i] = \frac{exp\left[-\frac{1}{2}(x_t - \mu_i)^T \Sigma_i^{-1}(x_t - \mu_i)\right]}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \tag{2}$$

where μ_i is the mean vector and Σ_i is the covariance matrix.

B. Support Vector Machine

SVM is the classifier based on the principle of structural risk minimisation[3]. A typical SVM is a two-class classifier. In order to achieve speaker recognition, four methods are used commonly. They are one-against-one, one-against-all, top-down binary tree and bottom-up binary tree. For example in one-against-all method, there are L SVMs needed altogether when there are L speakers in the dataset. When the ith SVM is trained, the patterns of the ith speaker are one class and the rest patterns are the other class. After all the SVMs are trained, the support vectors of each SVM are stored and then the speaker models are constructed. In test phase, a test pattern is inputed into all SVM models. If the value of decision function for the ith SVM model is largest, the test pattern is assigned to the ith speaker.

III. Speaker Recognition Based on TWSVMs

In this section, the proposed approach for text-independent speaker recognition is demonstrated in detail.

Fig 1 shows the framework of the proposed approach. In the training process, there are two steps. Firstly, features are extracted from the GMMs. Secondly, the extracted features are used to train TWSVMs models. The obtained TWSVMs models are stored as reference models. In the test process, the extracted features are used in to reference models and get the recognition results.

Given a speech dataset of L speakers, speech of each speaker is divided into H segments. Let $X = [x_1^T; x_2^T; \ldots; x_m^T] \in R^{m \times n}$ be the n dimensional data set of m data points. It is the MFCC (Mel Frequency Cepstrum Coefficient) feature of each speech segment. $z_j^{(i)} \in R^{2n \times 1} (i = 1, 2, \ldots, L; j = 1, 2, \ldots, H)$ is the final feature of the speech segment extracted by GMM. The dimension n of GMM models is the same as that of X. The n dimensional mean vectors and covariance matrixes of GMMs form the $2n$ dimensional feature $z_j^{(i)}$ together. $Z^{(i)} = [(z_1^{(i)})^T; (z_2^{(i)})^T; \ldots; (z_H^{(i)})^T]$ is the final features belong to ith person. $Z = [Z^{(1)}; Z^{(2)}; \ldots; Z^{(L)}]$ is all the final features.

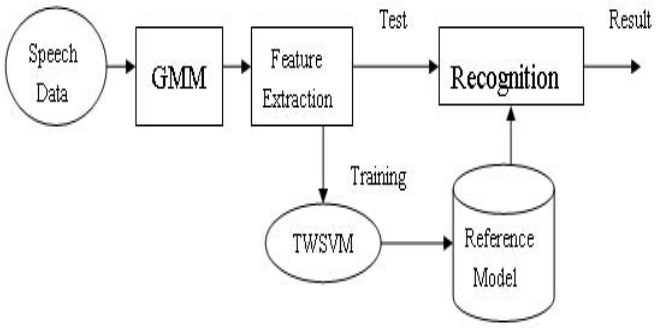


Fig. 1. The framework of the proposed method.

A. Feature Extraction based on GMMs

Firstly, the MFCC feature X is extracted from each speech segment. Then X is used as the input set to computed GMM models.

The ω_i, μ_i and Σ_i in GMMs can be computed by EM arithmetic. The reestimation formulas as follows

$$\tilde{\omega_i}^{(d)} = \frac{1}{m} \sum_{t=1}^{m} P_i \tag{3}$$

$$\tilde{\mu_i}^{(d)} = \frac{\sum_{t=1}^{m} P_i x_t}{\sum_{t=1}^{m} P_i} \tag{4}$$

$$\tilde{\Sigma_i}^{(d)} = \frac{\sum_{t=1}^{m} P_i [x_t - \tilde{\mu_i}^{(d-1)}][x_t - \tilde{\mu_i}^{(d-1)}]^T}{\sum_{t=1}^{m} P_i} \tag{5}$$

where d is the number iterations. And P_i is the posteriori probability $P[i|x_t, \tilde{\mu_i}^{(d-1)}, \tilde{\Sigma_i}^{(d-1)}]$. It is given by

$$\begin{aligned} & P[i|x_t, \tilde{\mu_i}^{(d-1)}, \tilde{\Sigma_i}^{(d-1)}] \\ = & \frac{\omega_i p[x_t|\tilde{\mu^i}^{(d-1)}, \tilde{\Sigma_i}^{(d-1)}]}{\sum_{j=1}^{n} \tilde{\omega_i}^{(d-1)} p[x_t|\tilde{\mu_i}^{(d-1)}, \tilde{\Sigma_i}^{(d-1)}]} \end{aligned} \tag{6}$$

Because GMMs is a means of quantizing the speech space in a way that closely reflects the speech production process[2], the feature extraction method can increase recognition rate.

Then the final feature of each segment is computed as

$$\begin{aligned} z_j^{(i)} = & \left[\sum_{t=1}^{k} \omega_t \mu_t; \sum_{t=1}^{k} \omega_t \Sigma_t\right] \\ & i = 1, 2, \ldots, L; j = 1, 2, \ldots, H. \end{aligned} \tag{7}$$

where k is the number of Ganssian in GMMs.

Because the number of feature vectors is very large in speaker recognition, the size of the training dataset need to be reduced. As several times decrease in the number of final features by the above feature extraction, the proposed approach perform well in practice.

B. Model train based on TWSVMs

The TWSVMs classifier for the ith speaker is obtained by solving the following pair of quadratic programming problems[1]

$$\begin{aligned}&\text{(TWSVM1)}\\ &\min_{\mu^{(i)},b^{(i)},q} \frac{1}{2} \parallel K[Z^{(i)}, Z^T]\mu^{(i)} + e_i b^{(i)} \parallel^2 + c_i \bar{e_i}^T q\\ &s.t. \quad -\{K[Z^{\bar{(i)}}, Z^T]\mu^{(i)} + \bar{e_i} b^{(i)}\} + q \geq \bar{e_i},\\ &\qquad q \geq 0, \end{aligned} \tag{8}$$

and

$$\begin{aligned}&\text{(TWSVM2)}\\ &\min_{\mu^{\bar{(i)}},b^{\bar{(i)}},q} \frac{1}{2} \parallel K[Z^{\bar{(i)}}, Z^T]\mu^{\bar{(i)}} + \bar{e_i} b^{\bar{(i)}} \parallel^2 + \bar{c_i} e_i^T q\\ &s.t. \quad K[Z^{(i)}, Z^T]\mu^{\bar{(i)}} + e_i b^{\bar{(i)}} + q \geq e_i,\\ &\qquad q \geq 0, \end{aligned} \tag{9}$$

where Z is all the final features, $Z^{(i)} \subset Z$ is the features belong to ith person, $Z^{\bar{(i)}} \subset Z$ is the other features. $c_i, \bar{c_i} > 0$ are parameters. A larger value of c_i or $\bar{c_i}$ emphasizes the classification error, while a smaller one places more importance on the classification margin. And e_i and $\bar{e_i}$ are vectors of ones of appropriate dimensions, $K[Z^{(i)}, Z^T]$ and $K[Z^{\bar{(i)}}, Z^T]$ are appropriate kernels as the dataset of speaker recognition may not be linearly separable[10].

The algorithm finds two hyperplanes, i.e., one for each class, and classifies data points according to which hyperplane a given point is closest to. The first term in the objective function of (8) and (9) is the sum of squared distances from the hyperplane to data points of one class. The second term of the objective function is the sum of error variables.

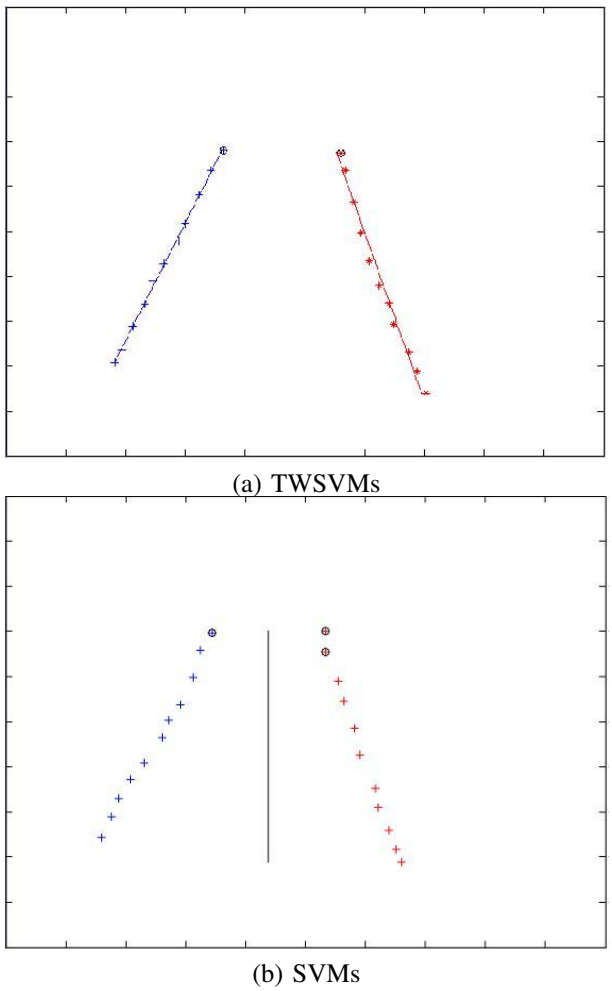
The difference between the separating plane of TWSVMs and that of traditional SVMs is showed on Fig 2 and Fig 3.

Based on the above discussion, all patterns belong to the ith speaker lie closest to the plane get from (8). This separating plane for the ith person is expressed as

$$K(z^T, Z^T)\mu^{(i)} + b^{(i)} = 0 \tag{10}$$

Assuming there are L speakers in the dataset. The separating plane for each speaker in the recognition dataset can be obtained by the above approach. The number of separating planes is the same as the number of speakers in the dataset. And each plane is closest to all patterns of its corresponding speaker. This can be illustrated on Fig 4

If a new pattern $z_s \in R^{2n \times 1}$ lies closest to the rth separating plane, it is assigned to the rth speaker,



(a) TWSVMs

(b) SVMs

Fig. 2. The separating planes of TWSVMs and traditional SVMs for linearly separable data points.

$$K(z_s^T, Z^T)\mu^{(r)} + b^{(r)} = \min_{i=1,2,\ldots,L} |K(z_s^T, Z^T)\mu^{(i)} + b^{(i)}| \tag{11}$$

where $|\cdot|$ is the perpendicular distance of point z_s from the separating plane.

C. Summary of The Algorithm

The detailed steps of the proposed approach for speaker recognition are as follows:

1)The MFCC features of each speech segment for the ith speaker is extracted.

2)Use X as input set. From equation (3), (4) and (5), the GMM model is obtained. Then the final feature $z_j^{(i)}$ of the speech segment is formed as (7).

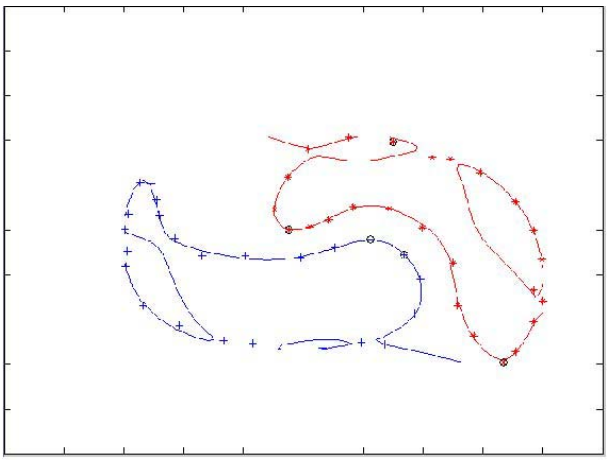
3)Do the same work to all speech segments of ith speaker.

4)Repeat 1), 2) and 3), the final feature Z of all speakers in the dataset are obtained.

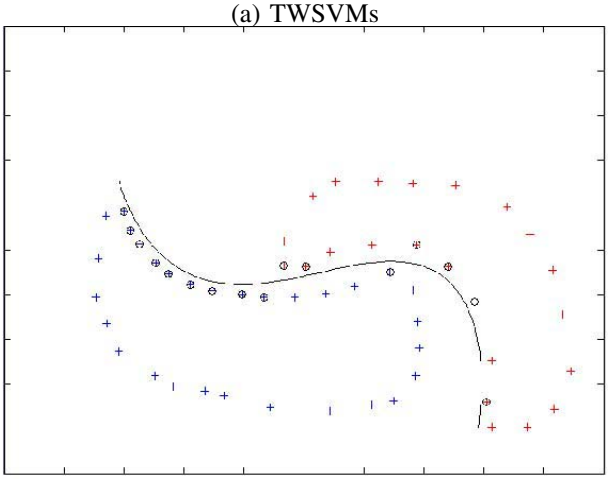
5)The separating plane for all speakers is obtained as (10).

6)Get final feature z_s of the test pattern s. The process is similar with 1) and 2).

7)Assign s to a speaker according with (11).



(a) TWSVMs



(b) SVMs

Fig. 3. The separating planes of TWSVMs and traditional SVMs for nonlinearly separable data points.

An extra restriction can be added in the sixth step. It aims at influencing the final result. If the closest distance is larger than a threshold, the segment isn't considered to be any speaker in the dataset.

IV. Experimental results

In our text-independent speaker recognition, the database used are Ahumada Biometric Database (AHUMADA) and Gaudi Biometric (GAUDI)[12]. There are 25 male speakers in AHUMADA and 25 female speakers in GAUDI. Each of the 50 target speakers has about one minute speech for training and another one minute speech for testing. All speeches are PCM form with 16 kHz sampling rate. In the experiment, each train speech is divided into six segments and each segment is about nine seconds. Each test speech is divided into ten segments and each segment is considered as an unit to be identified. Firstly the 12-dimensional MFCC feature vector is extracted from each train segment. Then the MFCC features are computed to a 12-dimensional GMM model. Mean vectors and covariance matrixes of the GMM are calculated to form the 24-dimensional final feature vector.

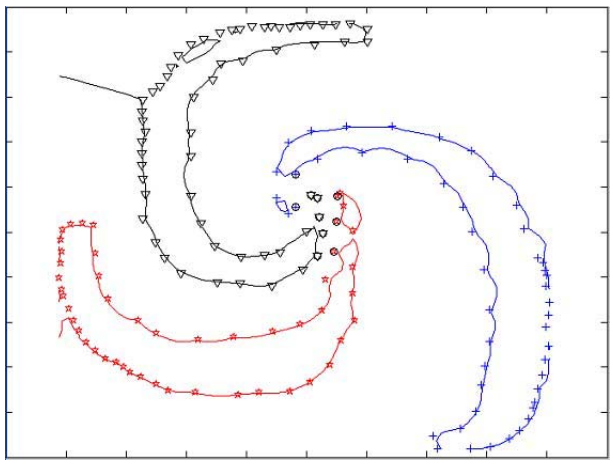


Fig. 4. Multiple classification of TWSVMs

TABLE I
RECOGNITION RATE WITH AN RBF KERNEL

	TWSVMs with GMM	SVMs	GMMs
AHUMADA	93.6	83.2	99.2
GAUDI	92.4	84.8	97.2
AHUMADA and GAUDI	92.2	84.0	98.4

Using this database, the TWSVMs with feature extraction based on GMMs, the traditional SVMs [8] and normal GMMs [4] data classification methods are implemented by running uncompiled Matlab code on a PC with an Intel P4 processor (2.40 GHz) with 512MB RAM.

Table I compares the performance of the TWSVMs classifier with that of SVM and GMM on AHUMADA and GAUDI. The kernals used are RBF kernel. Optimal values of c_i and $\bar{c}_i$ in TWSVMs are set to 10, and optimal values of c in SVM is set to 1. This setting is based on experimental experience. The table indicates that the proposed approach performs efficiently on AHUMADA and GAUDI. At the same condition, it performs better than SVMs and almost as effective as GMMs in terms of recognition rate. The recognition rate of GMMs reduces quickly with the decreasing of the number of the training patterns. But the proposed method has not this disadvantage as it combines the discriminnative model. In other word, the proposed approach can get the stable recognition rate using fewer training patterns. This advantage makes the proposed method more efficient and suitable for situations with small sample numbers.

In addition, the proposed approach also has the ability to deal with large dataset. When the number of patterns in database is increasing, the recognition rate of TWSVMs reduce slowly. Fig 5 shows the recognition rate of TWSVMs for different number of test segments. In every experiment, half of test segments come from AHUMADA and the other half are from of GAUDI[12]. It indicates that TWSVMs has ability to deal with large database.

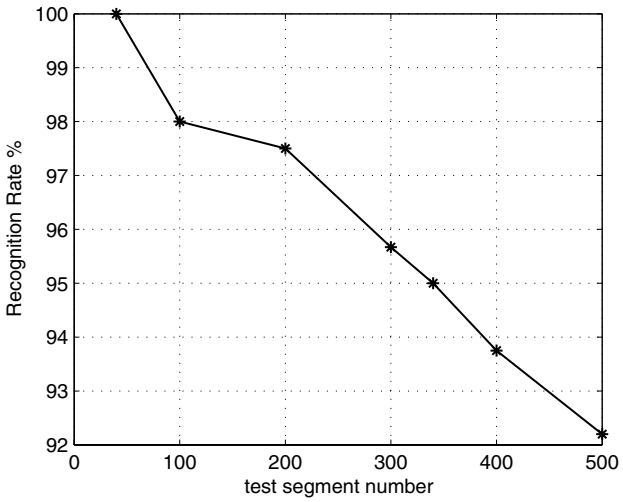


Fig. 5. The recognition rate of TWSVMs for different number of test segments.

V. Conclusion

This paper presents a new approach that combines TWSVMs with GMMs for text-independent speaker recognition . GMMs are used to extra feature in the approach. The proposed method is compared with the traditional SVMs and GMMs. The TWSVMS show more excellent performance than the standard SVMs on AHUMADA and GAUDI. The experimental results indicate the proposed method is efficient. And it has the ability to deal with large dataset.

One of the largest challenges in speaker recognition is to deal with variations between the training situations and testing situations. Since noises impose different characteristics on the acoustic signal, the spectrum-based features extracted for enrollment and recognition are different and hence may result in low match scores. The future work will focus on extending the recognition algorithm to different speech conditions, for example, telephone and noisy speech.

Acknowledgment

This work was supported by Chinese 863 High-Tech Program under Grant 2007AA01Z321.

References

[1] Jayadeva, R. Khemchandani and S. Chandra, *Twin Support Vector Machines for Pattern Classification*. IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 29, no. 5, pp. 905-910, May 2007.

[2] M. Liu, Y. Xie, Z. Yao and B. Dai, *A New Hybrid GMM/SVM for Speaker Verification*. International Conference on Pattern Recognition(ICPR'06), 2006.

[3] C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery, vol. 2, pp. 1-43, 1998.

[4] D. A. Reynolds, T. F. Quatieri and R. Dunn, *Speaker verification using adapted Gaussian mixture models*. Digital Signal Process, vol. 10, no. 1-3, pp. 19-41, 2000.

[5] W. M. Campbell, D. E. Sturim and D. A. Reynolds, *Support Vector Machines Using GMM Supervectors for Speaker Verification*. IEEE SIGNAL PROCESSING LETTERS, vol. 13, NO. 5, MAY 2006.

[6] R. Collobert and S. Bengio, *SVMTorch: Support vector machines for large-scale regression problems*. J. Mach. Learn. Res, vol. 1, pp. 143-160, 2001.

[7] G. Guo and S. Z. Li, *Content-Based Audio Classification and Retrieval by Support Vector Machines*. IEEE Trans. on Neural Networks, vol. 14, no. 1, pp. 209-215, Jan. 2003.

[8] P. Ding, Z. Chen, Y. Liu, and B. Xu, *Asymmetrical support vector machines and applications in speech processing*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), vol. 1, pp. 1-73 - 1-76, May 2002.

[9] C. J. C. Burges, *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowl. Discov, vol. 2, no. 2, pp. 1-47, 1998.

[10] V.Wan and S. Renals, *Evaluation of kernel methods for speaker verification and identification*. Proc. ICASSP, vol. 1, pp. 669-672, 2002.

[11] W. M. Campbell and K. T. Assaleh, *Polynomial Classifier Techniques for Speaker Verification*. Proc. ICASSP, vol. 1, pp. 321-324, 1999.

[12] J. Ortega Garcła,J. Gonzlez Rodrłguez and V. Marrero-Aguiar, *AHUMADA: A large speech corpus in Spanish for speaker characterization and identification*. Speech Communication, vol. 31, pp. 255-264, June 2000.

[13] W. M. Campbell, *A Sequence Kernel and its Application to Speaker Recognition*. Advances in Neural Information Processing Systems, vol. 14, 2001.

[14] M.H. Liu, B.Q. Dai, Y.L. Xie and Z.Q. Yao, *Improved GMM-UBM/SVM for speaker verification*. Proc. IEEE ICASSP, 2006.

[15] S. Fine, J. Navratil and R. A. Gopinath, *Enhancing GMM scores using SVM "hints"*. Proceedings of Eurospeech, pp. 1757-1761, 2001.

[16] V. Wan and S. Renals, *SVMSVM: Support Vector Machine speaker verification methodology*. Proc. IEEE ICASSP, 2003.