International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:1, No:4, 2007

A Hybrid GMM/SVM System for Text Independent Speaker Identification

Rafik Djemili, Mouldi Bedda, and Hocine Bourouba

***Abstract*—**This paper proposes a novel approach that combines statistical models and support vector machines. A hybrid scheme which appropriately incorporates the advantages of both the generative and discriminant model paradigms is described and evaluated. Support vector machines (SVMs) are trained to divide the whole speakers' space into small subsets of speakers within a hierarchical tree structure. During testing a speech token is assigned to its corresponding group and evaluation using gaussian mixture models (GMMs) is then processed. Experimental results show that the proposed method can significantly improve the performance of text independent speaker identification task. We report improvements of up to 50% reduction in identification error rate compared to the baseline statistical model.

***Keywords*—**Speaker identification, Gaussian mixture model (GMM), support vector machine (SVM), hybrid GMM/SVM.

I. Introduction

RECOGNIZING speakers from their voices are one of the important applications of speech technology in real world environment. Speaker recognition refers to two fields: Speaker Identification (SI) and Speaker Verification (SV) [1], [2]. In speaker identification, the goal is to determine which one of group of known voices (closed set) best matches the input voice sample. There are two tasks: text-dependent and text-independent speaker identification. In text-dependent identification, the spoken phrase is known to the system whereas in the text-independent case, the spoken phrase is unknown. Success in both identification tasks depends on extracting and modelling the speaker dependent characteristics of the speech signal, which can effectively distinguish between talkers. In the past years, several modelling techniques have been addressed. These cover pattern matching approaches (dynamic time warping), statistical modelling (Hidden Markov Models HMM or Gaussian Mixture Models GMM), and connectionist methods (multilayer perceptrons) [3]-[5].

Gaussian Mixture Models (GMM) represent the state-of-the-art technique in text independent speaker identification [6]. However GMMs trained with maximum likelihood (ML) criterion suffer from lack of discrimination. Recently, a new classification method called Support Vector Machines (SVM) [7], [8] based on the principle of structural risk minimization has found a great attention in the speech community. SVMs are attractive because they discriminate between classes and could be used to train non-linear decision boundaries in an efficient manner. So one can hope to increase the efficiency of standard generative models like GMMs and HMMs with the discriminative power of SVMs. Some researchers in last few years proposed methods following this way in different tasks of speaker recognition with much success [9]-[11].

In this paper we propose a new combination scheme using the SVM ability in discrimination between two classes and the classification power of a GMM, we argue and we will particularly show that our combination method brings a significant performance applied in a text independent speaker identification task over the standard approach (baseline system) using only GMMs.

The remainder of this paper is organized as follows: in section II we review the basics of a GMM system and its application in a speaker identification task. In section III we present SVM theory, we also describe our combination scheme. Experimental results that lead us to construct and choose some crucial parameters are given in section IV. Finally conclusions and perspectives are drawn in section V.

II. Gaussian Mixture Speaker Model

This section describes the form of a Gaussian mixture model (GMM) and its use as a representation of speaker identity for text independent speaker identification. Prior to construct a GMM for each speaker, speech signal is first transformed to a set of spectral vectors, which is a convenient representation of a person's vocal tract structure and would constitute an important factor distinguishing one person's voice from another. Details of this transformation are given later in section IV. Description of the GMM system herein uses the same notation as in [6].

A. The Gaussian Model

A Gaussian mixture density is a weighted sum of M component densities given by:

$$p(x/\lambda)=\sum_{i=1}^{M} p_i b_i(x) \quad (1)$$

Where x is a d-dimensional vector, $b_i(x)$ are the component densities and p_i the mixture weights. Each component density is a d-variate Gaussian function having the form:

Manuscript received April 7, 2007.

R. Djemili is with the Electronics Department, University Badji Mokhtar of Annaba, Algeria (phone: +213 074 061 068; fax: +213 038 876 565; e-mail: djemili_rafik@ yahoo.fr).

M. Bedda is with the Electronics Department, University Badji Mokhtar of Annaba, Algeria (e-mail: mouldi_bedda@yahoo.fr).

H. Bourouba is with the Electronics Department, University of Constantine, Algeria (e-mail: bourouba2004@yahoo.fr).

$$b_i(x)=\frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \quad (2)$$

With mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$.

The Gaussian mixture density is parameterised by the mean vectors, covariance matrices and mixture weights from all component densities. The parameters are represented by a single notation:

$$\lambda=(p_i,\mu_i,\Sigma_i) \qquad i=1...M \quad (3)$$

For speaker identification each speaker is modelled by a GMM and is referred to by his model λ.

B. Parameter Estimation

Given training speech (transformed to spectral vectors) from a speaker's voice, the goal of speaker model training is to estimate the parameters of the GMM λ, which in some sense best matches the distribution of the training feature vectors. The most popular method for training GMMs is a maximum likelihood (ML) estimation[12]. The aim of ML estimation is to find the model parameters, which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors $X=(x_1,...,x_T)$ the GMM likelihood can be written as:

$$p(X/\lambda)=\prod_{i=1}^{T} p(x_i/\lambda) \quad (4)$$

Maximization of the quantity in (4) is accomplished through running the expectation-maximization (EM) algorithm [13]. The idea is beginning with an initial model λ, to estimate a new model $\bar{\lambda}$ satisfying $p(X/\bar{\lambda})\geq p(X/\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. Following formulas are used on each EM iteration.

Mixture weights: $$\bar{p}_i=\frac{1}{T}\sum_{t=1}^{T} p(i/x_t,\lambda) \quad (5)$$

Means: $$\bar{\mu}_i=\frac{\sum_{t=1}^{T} p(i/x_t,\lambda)x_t}{\sum_{t=1}^{T} p(i/x_t,\lambda)} \quad (6)$$

Variances: $$\bar{\Sigma}_i=\frac{\sum_{t=1}^{T} p(i/x_t,\lambda)x_t^2}{\sum_{t=1}^{T} p(i/x_t,\lambda)}-\bar{\mu}_i^2 \quad (7)$$

The *a posteriori* probability for acoustic class is given by

$$p(i/x_t,\lambda)=\frac{p_i\, b_i(x_t)}{\sum_{k=1}^{M} p_k\, b_k(x_t)} \quad (8)$$

C. Speaker Identification

For speaker identification, a group of S speakers S=(1,2,…,S) is represented by GMM's $\lambda_1,\lambda_2,...,\lambda_S$. The objective is to find the speaker model, which has the maximum *a posteriori* probability for a given observation sequence.

$$\hat{S}=\underset{1\leq k\leq S}{\text{Argmax}}\, P(\lambda_k/X)=\underset{1\leq k\leq S}{\text{Argmax}}\frac{p(x/\lambda_k)P(\lambda_k)}{p(X)} \quad (9)$$

Where the second equation is due to Bayes's rule. Assuming equally likely speakers $(P(\lambda_k)=1/S)$ and noting that p(X) is the same for all speaker models, the classification becomes:

$$\hat{S}=\underset{1\leq k\leq S}{\text{Argmax}}\, p(x/\lambda_k) \quad (10)$$

Finally with logarithms, the speaker identification system gives:

$$\hat{S}=\underset{1\leq k\leq S}{\text{Argmax}}\sum_{t=1}^{T}\log p(x_t/\lambda_k) \quad (11)$$

In which $p(x_t/\lambda_k)$ is given in (1).

D. Performance Evaluation

Evaluation of a speaker identification experiment is conducted as follows. The test speech is first processed by the front-end analysis to produce a sequence of spectral vectors $(x_1,...,x_T)$. Different test utterances of length 2, 5 and 10 seconds were used each having a number of T feature vectors. Performance evaluation is then computed using the Identification Error Rate (IER) given by:

$$\text{IER}(\%)=\frac{\text{Num. Incorrect Identified Vectors}}{\text{Total Num. of Vectors}}*100 \quad (12)$$

The IER is calculated for each test utterance of length T vectors.

III. Support Vector Machines

Support Vector Machines (SVM) are binary classifiers based on the principle of structural risk minimization [14]. Experimental results indicate that SVMs can achieve a generalisation performance greater than or at least equal to traditional classifiers. SVMs use a known kernel function to define a hyperplane in order to separate given data points into two predefined classes. Within this separation, the soft-margin SVM can tolerate minor misclassifications [15]. It is considered to be more suitable for classification and therefore is used in our work.

A. Description of SVMs

We will give below a brief description of SVMs and how to use them in a pattern categorization. More details can be found in Vapnik's book [16] and in Burges' tutorial [17].

An SVM is a binary classifier that makes its decisions by constructing a linear decision boundary or hyperplane that optimally separates two classes.

The hyperplane is defined by $x.w+b=0$ where w is the normal to the plane. For linearly separable data labelled

(x_i, y_i), $x_i \in \Re^n$, $y_i \in \{-1,1\}$, i=1...N. The optimal hyperplane is chosen according to the maximum margin criterion (the minimal distance from the hyperplane to each points) i.e. by choosing the separating plane that maximises the Euclidean distance to the nearest data points on each side of that plane. The problem can be formulated as

$$\begin{cases} \text{minimise} & \frac{1}{2}\|w\|_2^2 \\ \text{subject to} & (x_i.w+b)y_i \geq 1 \end{cases} \quad (13)$$

The solution for the optimal hyperplane w_0, is a linear combination of a small subset of data, $x_s, s \in \{1...N\}$ known as support vectors. These support vectors also satisfy the equality

$$(x_s.w+b)y_s = 1 \quad (14)$$

When the data are not linearly separable then no hyperplane exists for which all points satisfy the inequality (13). In this case, we may include slack variables ξ_i shown in Fig.1 into the inequalities relaxing them slightly so that some points are allowed to be misclassified. The objective function becomes:

$$\begin{cases} \frac{1}{2}\|w\|_2^2 + C\sum_i L(\xi_i) \\ \text{subject to } (x_i.w+b)y_i \geq 1-\xi_i \text{ for all } i \end{cases} \quad (15)$$

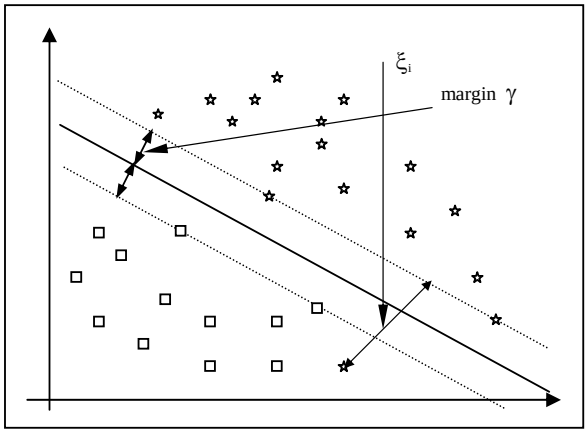

Fig. 1 Margin and slack variables for a classification problem

The second term of (15) is the empirical risk associated with those points that are misclassified, L is the loss function (cost function) and C is a hyperparameter that trades off the effects of minimizing the empirical risk against maximizing the margin. The first term can be thought as a regularization term, which gives the SVM its ability to generalize well on sparse data.

The linear error cost function is the most commonly used since it is robust to outliers. The dual formulation which is more conveniently solved, of (15) with $L(\xi_i)=\xi_i$ is:

$$\begin{cases} \alpha^* = \underset{\alpha}{\text{Max}}\left(\sum_i \alpha_i + \sum_{i,j}\alpha_i\alpha_j y_i y_j x_i.x_j\right) \\ \text{subject to } 0 \leq \alpha_i \leq C \text{ and } \sum_i \alpha_i y_i = 0 \end{cases} \quad (16)$$

In which α_i i=1...N is the set of Lagrange multipliers of the constraints in the primal optimisation problem. The dual can be solved using standard quadratic programming techniques. The optimum decision boundary w_0 is given by:

$$w_0 = \sum_i \alpha_i y_i x_i \quad (17)$$

And is a linear combination of all points in feature space that have $\xi_i > 0$ and lying on the margin $(\alpha_i \neq 0)$. The extension to non-linear boundaries is achieved through the use of kernel functions that satisfy Mercer's condition [18]. In essence, a non-linear mapping is defined from the input space, in which the data are observed, to a manifold in higher dimensional feature space, which is defined implicitly by the kernel functions.

The hyperplane is constructed in the feature space and intersects with the manifold creating a non-linear boundary in the input space. In practice, the mapping is achieved by replacing the value of dot products between two vectors in input space with the value that results when the same dot product is carried out in the feature space. The dot product in the feature space is expressed conveniently by the kernel as some function of the two vectors in input space. The polynomial and radial basis function (RBF) kernels are commonly used, and take the form:

$$K(x_i, x_j) = (x_i . x_j + 1)^n \quad (18)$$

And

$$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2\right] \quad (19)$$

Respectively, where n is the order of the polynomial and σ is the width of the radial basis function. The dual for the non-linear case is thus:

$$\begin{cases} \alpha^* = \underset{\alpha}{\text{Max}}\left(\sum_i \alpha_i + \sum_{i,j}\alpha_i\alpha_j y_i y_j K(x_i, x_j)\right) \\ \text{subject to } 0 \leq \alpha_i \leq C \text{ and } \sum_i \alpha_i y_i = 0 \end{cases} \quad (20)$$

The use of kernels means that an explicit transformation of data to the feature space is not required.

B. SVMs and Speaker Identification Systems

Although GMMs are the most widely used in speaker identification systems [19], lack of discrimination of such generative models incited researchers to find out

discrimination based learning procedures in order to obtain or to outperform GMM's performance. SVM classifiers are well suited to separate complex regions between two classes through an optimal non-linear decision boundary.

The first approach in using SVM classifiers in the framework of speaker identification was implemented in [20] where SVMs were trained directly on the acoustic space, which characterize the client data and the impostor data, during testing the segment score is obtained by averaging the scores of the SVM output for each frame. Other applications of SVMs used kernels sequences [21].

Another approach became recently more popular, consists of making a combination of GMMs and SVMs. Several types of combination were proposed. In [22] a discriminative training of GMMs is performed by continuous density SVM. Another from of combination used SVMs as a post treatment of the GMMs by Fischer mapping [23]. This mapping allows obtaining vectors of high dimensions where the number of dimensions is equal to the number of the GMM parameters. These vectors are then used by SVMs to achieve discrimination and decision.

The work presented in [24] exploits the advantages of the GMM models and SVMs in a single system by deriving a probabilistic distance kernel computed using the divergence of Kullback-Leibler (KL) between GMMs.

C. The Proposed Hybrid GMM/SVM System

The work presented here belongs to the category of combining the benefits of GMMs in training and SVMs in discrimination. SVMs used in this paper are binary classifiers between two groups of speakers giving a hierarchical tree structure. Identification errors from the baseline system, as we will see in next section often occurs when a speaker is taken for another speaker belonging to the same gender, i.e. a male speaker ms_i is confused with another male speaker ms_j and a female speaker fs_k unrecognized as another female speaker fs_l. Since SVMs had proved their effectiveness in separating two given classes, we applied them in dividing very confusable speakers prior to the identification system using GMM speaker's models. The overall structure of our hybrid system is depicted in Fig. 2.

Following feature extraction of the input speech signal, SVM 1 is aimed at finding the gender of the input voice. SVM 2 and SVM 3 are trained to cluster the male group (respectively the female group) into two subsets of speakers. Finally at the last level identification is carried out using only a subset of GMM speaker models. In comparison with the standard approach GMMs are still used in evaluation but with much less computational load, since the initial S speakers are divided nearly by a factor of 4, and especially achieving higher identification accuracy. Thus, our hybrid system involves two main steps:

Training: where the individual GMM speakers' models are constructed along with the support vector classifiers SVM 1, SVM 2 and SVM 3.

Testing: the identification process of the GMM/SVM system proposed follows the hierarchical structure of Fig. 2. It should be noted here that errors for the baseline system and the hybrid one are not correlated. If the GMM system fails at identification one speaker among the whole set of S speakers, it is unlikely to happen when only a subset of the S speakers are in competition.

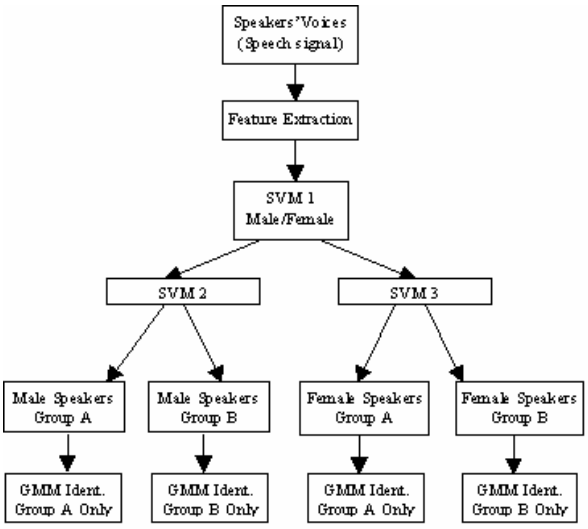

Fig. 2 Hierarchical tree structure of the hybrid GMM/SVM system

IV. Experimental Results

A. Database and Speech Analysis

Experiments in this study were performed using Arabic data sets built for the purpose. Clean speech signals using close talk high quality microphone was recorded under quiet laboratory conditions. Data sets were collected from 16 speakers, 10 were female and 6 male. For each speaker, there were 3 separate sessions, two sessions were used for training data and one session for testing.

Speech signals were sampled at 11025 Hz. Mel scale Frequency Cepstral Coefficients (MFCC) were employed as feature analysis [25],[26]. A pre-emphasis filter $H(z) = 1 - 0.95z^{-1}$ is used before framing. Each frame is multiplied with a 23.2ms Hamming window shifted by 11.6ms. From the windowed frame FFT is computed and the magnitude spectrum is filtered with a bank of 27 triangular filters spaced on the Mel-scale. The log-compressed filter outputs are converted into cepstral coefficients by DCT giving twelve coefficients. The zeroth cepstral coefficient is not used in the cepstral feature vector and replaced with log of energy of the frame calculated in the time domain. Hence, the feature vector is formed by thirteen coefficients.

This processing occurs every 11.6ms producing approximatively 2586 and 5172 vectors in 30 and 60 seconds of speech, these quantities were used in training the baseline system (GMM only). For testing all systems carried out in this work, durations of 2s, 5s and 10s were utilised since the emphasis in speaker identification tasks is to capture the identity of a speaker with the minimum material in hand, here

the speech signal. Vectors obtained are 172, 431 and 862 corresponding respectively to durations above.

B. Baseline System

By the baseline system, it meant the system in which only the GMM speakers' models are used in the evaluation process. GMMs were trained using the theoretical material given in section II. An important problem rises when constructing a GMM's speaker model in how to choose adequately the number of components M in a mixture. Fig. 3 shows the identification performance evaluated with the Identification Error Rate (IER) when training data size equals to 30s and 60s of speech versus different values of M varying from M=2 to M=256. For both experiments, IER is slightly decreased when M is increased by a factor of 2, this is true for different lengths of testing utterances.

TABLE I
GMM IDENTIFICATION ERROR RATE (IER) IN PERCENT FOR DIFFERENT AMOUNTS OF TRAINING DATA AND MODEL ORDERS

	Training with 30s			Training with 60s		
M	**2s**	**5s**	**10s**	**2s**	**5s**	**10s**
2	51.9	52.1	56.7	54.0	56.0	56.4
4	43.5	42.7	43.2	46.5	46.8	46.9
8	34.8	33.6	33.9	36.1	36.9	37.5
16	25.0	24.7	24.8	27.8	28.3	28.5
32	19.2	18.1	17.6	20.8	20.5	20.2
64	15.3	13.7	13.2	16.3	15.3	14.9
128	13.8	11.9	11.0	13.9	12.4	11.6
256	12.9	10.6	9.8	11.7	9.8	9.1

It is clear from the table that the identification errors are superior to 20% from 2 to 16 mixture components even if training material is large (equals to 60s). We could argue that for a value of M up to 16, there are still few components to produce an accurate model capable at distinguishing characteristics of a speaker's distribution. For M varying from 32 to 256 components per model, identification errors fall than 20% except for training data size equals to 60s of speech with M=32, identification error rates for this range of M appear rather stable. Greatest reductions in IER are reached with M=256 and are 24% with a training data size of 30s and 22.2% with a training data of 60s when utterances lengths are compared on their extremum values 2s and 10s. It is hence observed lowest identification errors occur with longest test utterance lengths. Our best result is achieved for 256 mixture components with an utterance test length of 10s and a training data size of 60s of speech. Choosing too many mixture components (M>256) could increase in some cases identification performance but only if we have an available training data larger enough relative to the number of mixture components. However this is not always what happens when dealing with speaker recognition tasks, where we could have only small amounts of some speakers' voices.

Own experiments not included here, had showed an increase in IER to 25% when using 512 mixture components, this result due to the overfitting effect was also reported in [27].

Based on these observations, next comparisons will be conducted with 32, 64, 128 and 256 components of the GMM speakers' models.

C. Choosing SVM parameters

Crucial parameters for training a SVM are the upper bound C allowing us how strictly we want the classifier to fit the training data and the variance σ^2 of the radial basis function (RBF).

To investigate the performance accuracy of SVM classifier some experiments in separating between males and females were applied and summarized in Table II. Training SVMs was done using 2s of speech from each speaker.

TABLE II
SUPPORT VECTORS AND NUMBER OF ITERATIONS IN TRAINING SVM1 FOR SEVERAL VALUES OF VARIANCES AND UPPER BOUND C

	C=1		C=10		C=100	
σ^2	**NSV (%)**	**Iter**	**NSV(%)**	**Iter**	**NSV**	**Iter**
0.2	36.9	167	22.4	560	12.7	1829
0.5	43.9	162	28.3	779	18.2	3128
0.8	47.7	132	33.0	898	21.4	3237
1	49.6	130	35.5	978	23.2	3131
2	54.0	97	42.0	458	29.0	2535
5	61.3	67	46.6	349	38.7	1714
10	69.3	67	49.6	323	43.2	1073
20	48.9	74	53.5	197	45.0	621
100	97.1	89	68.9	179	49.2	170

NSV: Number of Support Vectors in (%)
Iter: Number of iterations in training SVMs

Three values of upper bound C were compared over a range of preselected variances for the RBF. For each combination of C and σ^2, the number of support vectors (NSV) given in (%) and the number of iterations in training SVMs were computed. It is seen that small values for variances but with a linear increase in C give acceptable amount of support vectors, which in turn must be stored in memory for testing. However the algorithm seems to be faster with decreased upper bounds C. Evaluating the accuracy of the SVM classifier is shown in Fig. 3, where training was accomplished with 2s, 5s and 10s of speech per speaker. It is interesting to observe the insensitivity for the classifier regarding the amount of speakers' durations (Fig. 3a).

The average accuracy obtained in dividing the two genders, are respectively 94.08%, 94.7% and 95.5% for the 2s, 5s and 10s of speech. Only a 1.5% reduction in performance accuracy is noted when comparing the 10s and 2s durations while the number of vectors in training is multiplied by a factor of 5. Number of support vectors is also consistent when C is set to larger values.

A variance of 0.2 and an upper bound of 1000 appear to be good choices in training SVMs in this work.

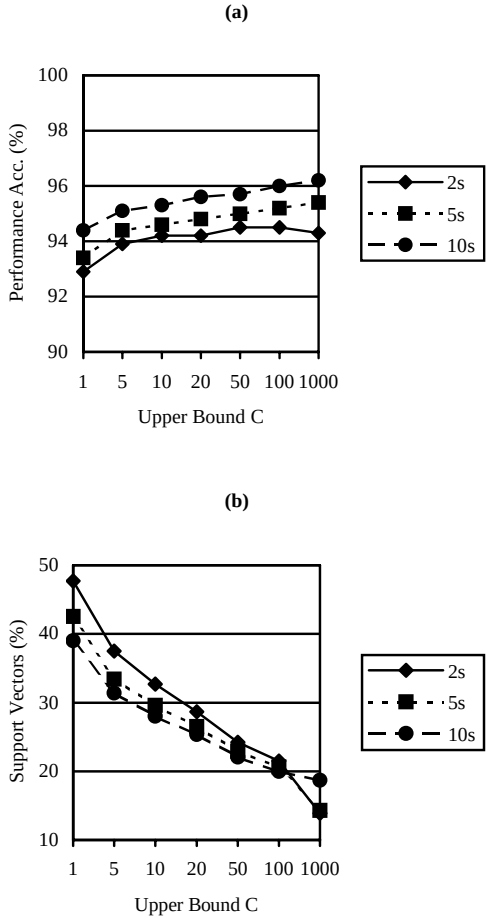

Fig. 3 Performance accuracy (a) and fraction of support vectors (b) in percent as a function of upper bound C obtained when training SVM1 with different speakers' durations

D. The GMM/SVM System

The focus of this paper is to compare the performance of the baseline GMM system and the proposed GMM/SVM described earlier in Fig. 2 applied on a text independent speaker identification task. The hybrid GMM/SVM uses GMMs in the identification process on a relative small subset of speakers given after classifying the unknown input speaker' voice into its corresponding group (subset) via the hierarchical splitting by SVMs namely SVM1, SVM2 and SVM3.

Training SVM1 in order to divide into male and female genders is straightforward, male and female speakers are known from the database, supervised training is directly implemented. For training SVM2 and SVM3 as binary classifiers on positive and negative examples from respectively male and female speakers, we used the following strategy: On the basis of experimental analysis of the GMM system and searching for sources of identification errors for each speaker s_i taken as another speaker s_j, we have noted that confusable identification errors occur often on the same gender group, speakers s_i and s_j are both males or females. This is indeed shown in Fig. 4 where male speakers are indexed as 1,6,9,10,14,16; the other index concerns female speakers. We could observe from the figure, let's take as an example for male speakers 1 and 10, they were confused with speakers 14 and 6; so the idea we adopted for training SVM2 was to put these confusable speakers on each side of the hyperplane. The same rule of reasoning was applied for training SVM3 corresponding to female speakers. It is worth noting the possibility of the evaluation step by GMMs to be processed after an individual use of SVM1, SVM1 + SVM2, SVM1 + SVM3, and SVM1 + SVM2 + SVM3. Identification performance for the hybrid GMM/SVM system is summarized in Table III.

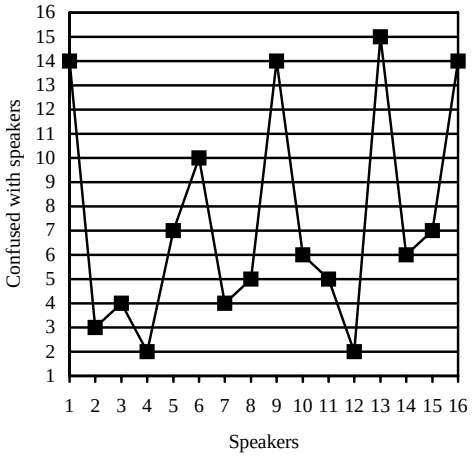

Fig. 4 Curve representing the confusable speakers averaged from experiments on the baseline system

TABLE III
IDENTIFICATION ERROR RATE (%) FOR THE HYBRID GMM/SVM SYSTEM FOR DIFFERENT MODEL ORDERS AND UTTERANCE LENGTHS USING: (A) TRAINING DATA WITH 30S OF SPEECH AND (B) TRAINING WITH 60S OF SPEECH PER SPEAKER

(A)

GMM Order	SVM1			SVM1+SVM2			SVM1+SVM3			All		
	2s	5s	10s	2s	5s	10s	2s	5s	10s	2s	5s	10s
32	17.9	16.9	16.4	14.2	13.8	13.0	13.0	11.8	11.7	9.30	8.60	8.30
64	14.6	12.8	12.3	11.7	10.1	9.6	10.7	9.0	8.6	7.8	6.2	5.9
128	12.9	11.0	10.2	10.1	8.4	7.8	9.5	7.7	7.0	6.7	5.1	4.7
256	12.3	10.0	9.1	10.1	8.1	7.2	8.9	6.7	6.2	6.6	4.9	4.4

(B)

GMM Order	SVM1			SVM1+SVM2			SVM1+SVM3			All		
	2s	5s	10s	2s	5s	10s	2s	5s	10s	2s	5s	10s
32	19.7	19.0	18.7	15.7	15.2	15.0	15.1	14.1	13.8	11.1	10.3	10.0
64	15.3	14.1	13.8	12.6	11.4	11.0	11.4	10.2	9.9	8.7	7.4	7.1
128	12.9	11.4	10.7	10.3	8.9	8.4	9.9	8.2	7.5	7.4	5.7	5.2
256	11.0	9.0	8.4	8.8	7.0	6.5	8.3	6.4	6.0	6.1	4.4	4.0

The purpose of the experiment is two manifold, we would evaluate the effect of training data size and of testing utterance length, using model sizes ranging from 32 to 256 and several combination of support vector machines trained on target subsets of speakers. It is clearly shown a significant improvement on speaker identification performance for the hybrid system compared to the baseline one over all the

variants explored. Although the two first raws of the two parts of table III give error rates a little bit greater than that of the last ones, this fact is due to the number of mixture components of the model, when this is large, better is the performance of the identification system. Relative improvements compared to the baseline system when using all the SVMs are 48.8%, 53.7% and 55.1% corresponding to testing utterance lengths of 2s, 5s and 10s respectively, with training data size equals 30s of speech per speaker. Further improvements are reached when training data size is augmented to 60s, reductions in IER are 47.9, 55.1 and 56% corresponding respectively to testing utterance lengths of 2s, 5s and 10s. As expected, larger testing utterance length provides smallest error rates. Hence, our best results are achieved with 256 mixture components and 10s of testing utterance length using the largest amount of training data.

V. Conclusion

In this paper, we have proposed a combination method which includes both the descriptive strength of the GMM system with the high performance classification capabilities of SVMs applied in a text independent speaker identification task. SVMs in this work are trained to divide the whole set of speakers into small subsets through a hierarchical tree structure. Next, GMMs would be used in the evaluation process. The highlights of the proposed hybrid system are:

1) A significant improvement compared to the baseline system is reported, a relative reduction in identification error rate up to 50% is reached, independently neither on the training data size nor on the testing utterances lengths.

2) A reduction in computational load, since for the hybrid system, testing is carried out on a limited GMM models depending on the size of speakers' subsets, whereas for the baseline system all the speakers' GMM models are evaluated.

Supervised training of SVMs in our work is based upon an objective analysis of identification errors of confusable speakers provided by the baseline system. Further work could include an automatic construction of the structured hierarchical tree avoiding any use of a priori knowledge about speakers.

Acknowledgment

The authors wish to thank the various members of the automatic and signals laboratory of Annaba (LASA) for their continuous support and help. I am also grateful to E. Chouiha and A. Abderahim for gathering data sets employed in this work.

References

[1] D. Reynolds, "An overview of automatic speaker recognition technology," *in Proc. Int. Conf. on Acoust. Speech and Signal Process. (ICASSP 2002)*, Orlando, FL, 2002, pp. 4072-4075.

[2] F. Bimbot, J.-F. Bonastre, G. Gravier, I. Chagnolleau, S. Meignier, T. Merlin, J. Garcia, D. Delacrètaz, and D. Reynolds, "A tutorial on text independent speaker verification," *Eurasip Journal on Applied Signal Process.*, vol. 4, pp. 430-451, 2004.

[3] T. Matsui and S. Furui, "Comparison of text independent speaker recognition methods using VQ distorsion and discrete/continuous HMM's," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 3, pp. 456-459, July 2004.

[4] K. Farell, R. Mammone, and K. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 1, pp. 194-205, 1994.

[5] J. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437-1462, Sep. 1997.

[6] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture models," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[7] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," *in Proc. of the 5th Annual ACM Workshop on Computational learning theory*, ACM press, pp. 144-152, 1992.

[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[9] J. Kharroubi, D. Petrovska, and G. Chollet, "Combining GMM's with support vector machines classifier," *in Proc. European Conf. Speech Communication and Technology (EUROSPEECH 2001)*, Aalborg, Denmark, 2001, pp. 1757-1760.

[10] S. Fine, J. Navratil, and R. Gopinath, "A hybrid GMM/SVM approach to speaker identification," *in Proc. Int. Conf. on Acoust. Speech and Signal Process. (ICASSP 2001)*, Salt Lake City, Utah, 2001, pp. 417-420.

[11] X. Dong, W. Zhaohui, and Y. Yingchun, "Exploiting support vector machines in hidden Markov models for speaker verification," *in Proc 7th Int. Conf. on Spoken Language Process. (ICSLP 2002)*, Denver, Colorado, 2002, pp. 1329-1332.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed., Wiley, New York, 2001.

[13] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1-38, 1977.

[14] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[15] V. Keckman, *Learning and Soft Computing*, MIT Press, Cambridge, MA, 2001.

[16] V. Vapnik, *Statistical Learning Theory*, John Wiley, New York, 1998.

[17] 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.

[18] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Wiley Interscience, New York, 1953.

[19] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspectives," *Speech Communication*, vol. 31, pp. 225-254, 2000.

[20] M. Schmidt and H. Gish, "Speaker identification via support vector machines," *in Proc. Int. Conf. on Acoust. Speech and Signal Process. (ICASSP 96)*, Atlanta, 1996, pp. 105-108.

[21] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 203-210, Mar. 2005.

[22] X. Dong and W. Zhaohui, "Speaker recognition using continuous density support vector machines," *Electronics Letters*, vol. 37, no. 17, pp. 1099-1101, 2001.

[23] V. Wan and S. Renals, "SVMSVM: Support vector machine speaker verification methodology," *in Proc. Int. Conf on Acoust. Speech and Signal Process. (ICASSP 2003)*, Hong Kong, 2003, vol. 2, pp. 221-224.

[24] P. Moreno and P. Ho, "A new approach to speaker identification and verification using probabilistic distance kernels," *in Proc. European Conf. Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, 2003, pp. 2965-2968.

[25] J. R. Della, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed., IEEE Press, New York, 2000.

[26] D. O'Shaughnessy, "Interacting with computers by voice: Automatic speech recognition and synthesis," *Proc. IEEE*, vol. 91, no. 9, Nov. 2003.

[27] R. Stapert and J. S. Mason, "Speaker recognition and the acoustic speech space", *in Proc. Speaker Odyssey: The Speaker recognition Workshop*, Crete, Greece, 2001, pp. 195-1999.