Novel Approach in Speaker Identification using SVM and GMM

H.Bourouba*, C.A.Korba, Rafik Djemili*****

** University of 8 May 1945 Guelma, Sciences and Technology Department*
ALGERIA (e-mail: bourouba2004@ yahoo.fr).
*** Souk Ahras University, Electrical Engineering Department*
ALGERIA (amara_korba_cherif@yahoo.fr).
**** University of 20 August 1955 Skikda, Electronics Engineering Department,*
ALGERIA ((djemili_rafik@yahoo.fr)}

Abstract: Conventional speaker Identification systems use Gaussian mixture models (GMM) and support vector machines (SVM) to model a speaker’s voice based on the speaker’s acoustic characteristics. Whereas GMM needs more data to perform adequately and is computationally inexpensive, SVM on the other hand can do well with less data and is computationally expensive. This paper proposes a novel approach that combines the power of generative Gaussian mixture models (GMM) and discriminative support vector machines (SVM). Due to its excellent expandability, GMM have been used to extract a small quantity of typical feature vectors from large numbers of speech data for SVM classifier. A hybrid system is described and experimentally evaluated on a text-independent speaker identification task. The results prove that the combination is beneficial in terms of performance and practical in terms of computation.

Keywords: GMM, SVM, Speaker identification, hybrid system.

1. INTRODUCTION

The speech signal conveys several levels of information. Primarily, the speech signal conveys the words or message being spoken, but on a secondary level, the signal also conveys information about the identity of the speaker. The area of speaker recognition is concerned with extracting the identity of the person speaking an utterance. As speech interaction with the computers become more pervasive in activities such as telephone transactions and information retrieval from speech databases, the ability to automatically recognize a speaker based on his vocal characteristics becomes more useful.

Speaker recognition refers to two fields: Speaker Identification (SI) and Speaker Verification (SV) (Beigi, 2011). In speaker identification, the goal is to determine which one of group of known voices 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 from the speech signal, which can effectively distinguish between talkers.

The problem of speaker identification can be formulated as a pattern classification problem and methods from statistics and machine learning are suitable. Two techniques are widely used, namely discriminative classifiers and generative model classifiers (Jaakkola, 1998). Discriminative models have to discriminate the information of different classes, while generative models use statistical information. In short, discriminative models use inter-class information and generative models uses intra-class information. Since discriminative model and generative model have both advantages in themselves, they also have disadvantages of lack using the other kind of information. Combining the discriminative model and the generative model can improve the performance in pattern recognition.

For speaker modeling and recognition many methods have been proposed. In text independent speaker recognition the most popular methods are: Vector Quantization (VQ) (Zhong *et al.,* 1999), artificial neural networks (ANN) (Fenglei, 2000), support vector machines (SVM) (Boujelbene *et al.,* 2009), and Gaussian Mixture Models (GMM) (Reynolds, 2002).

As example of discriminative models are Artificial Neural Networks (ANN) and Support Vector Machines (SVM), among others, and for Generative model are Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). Each one of them can construct speaker models for speaker recognition tasks, and previous studies show good performances.

Gaussian mixture models (GMM) represent a state-of-the-art technique in text-independent speaker recognition (Chaudhari, 2001) and many other tasks including detection and segmentation (Reynolds *et al.,* 2000; Weber and Beskin, 2000).

In recent years a new classification methodology based on Support Vector Machines (SVM) (Alex *et al.,* 2000) has found an increased interest in the speech community.

Favourable properties of the SVM such as their inherent class-discriminative model structure and the use of nonlinear kernel methods represent an attractive way of enhancing the standard method, mostly based on generative models (GMM) by complementary information and classification "power."

It is therefore attractive to explore methods that combine generative models GMM and discriminative models SVM. The hybrid text independent speaker identification based approach that incorporates both the SVM and the statistical models (GMM) is suggested in this paper. This is done in a way that the robustness advantage of the generative statistical models favourably combines with the discriminative power of the SVM.

The main idea of this proposed approach is to use a new feature representation based on GMM to construct the input vectors to train the SVM to discriminate the true-target speaker class from the non-target speaker class. The main idea is to use the mean vectors of the mixture components as a training set of the support vector machine, and then this SVM classifier is used in identification step. The new GMM/SVM system is tested and compared with baseline system in speaker identification using Gaussian Mixture Models with Mel Frequency Cepstral Coefficients.

The rest of the paper is organized as follows. In the next section, we introduce the speaker identification system and acoustic modeling used in our experiments. In section 3 and 4, we discuss some aspects of GMM and SVM respectively. In section 5, we discuss the hybrid approach in depth. In the next section we present the experiments carried out to examine the performance of GMM and GMM/SVM. Finally, we achieve our paper with conclusion.

2. FEATURE EXTRACTION

The voice signals naturally have a negative spectral slope (of approximately 20dB per decade) due to physiology of speech production (Picone, 1993). A pre-emphasis filter compensates this slope before signal analysis. A simple first order high-pass filter is used as:

$$s(t) = 1 - 0.97s(t-1) \tag{1}$$

The voice signal is divided into overlapping segments called frames. Each frame is multiplied by hamming window for smoothing the effect of using finite frame.

2.1 Cepstral coefficients

MFCCs are widely used features to characterize a voice signal. The signal is windowed in the time domain and converted into the frequency domain by FFT, which gives the amount of energy present within particular frequency range

Triangular Mel-frequency filters are then applied to reduce the amount of data by summing filtered FFT bin values to get the Mel filter bank outputs. Mel-scaling is performed to get higher resolution at low frequencies and lower resolution at high frequencies. This is based on the human perception, where a relationship between the real frequency scale (Hz) and the perceptual frequency scale (Mel) is logarithmic above 1000 Hz and linear below.

Finally, MFCCs are obtained by applying the discrete cosine transform (DCT) to the logarithm of Mel filter bank outputs (or energies). DCT represents the signal in terms of the first basis function (constant component) and the remaining basis functions (components of successively increasing frequency), which are uncorrelated. First *d* components of DCT represent a compacted MFCC vector of the corresponding frame (Fig.1).

Denoting the output of the filter bank by E_k ($k = 0,1,...,K$), the MFCCs are calculated *as*

$$C_n = \sum_{k=1}^{K} (\log E_k) \cos\left[n(k-0.5)\frac{\pi}{K} \right], n = 0,1,...,D \tag{2}$$

Where D is the number of MFCC coefficients, K is the number of Mel-scaled filters

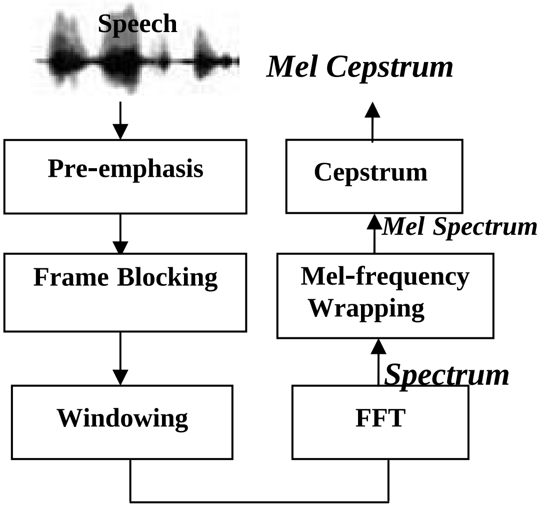


Fig. 1. MFCC calculation.

3. SUPPORT VECTOR MACHINE (SVM)

The SVM (Vapnik, 1998) is a useful statistic machine learning technique that has been successfully applied in the pattern recognition area (Jiang *et al.,* 2005; Guo and Li, 2003). If the data are linearly non separable but nonlinearly separable, the nonlinear support vector classifier will be applied. The basic idea is to transform input vectors into a high-dimensional feature space using a nonlinear transformation, and then to do a linear separation in feature space as shown in Fig. 2

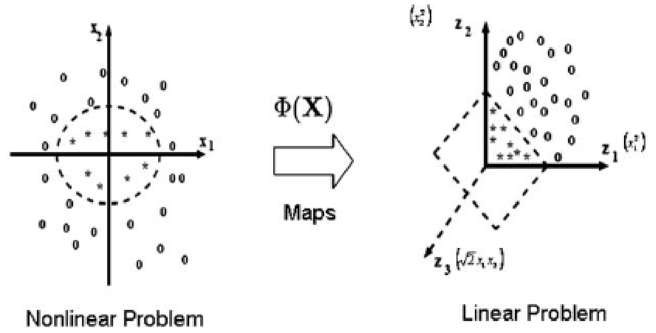
To construct a nonlinear support vector classifier, the inner product (x,y) is replaced by a kernel function K(x,y):

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \tag{3}$$

Where the α_i are the Lagrange Multipliers and b is the bias term.

The SVM has two layers. During the learning process, the first layer selects the basis $K(x_i, x)$, $i = 1,2,...,N$, from the given set of bases defined by the kernel; the second layer constructs a linear function in this space. This is completely equivalent to constructing the optimal hyper plane in the corresponding feature space.

The SVM algorithm can construct a variety of learning machines by use of different kernel functions. Three kinds of kernel functions are usually used. They are as follows:



$$\mathbf{Z} = \Phi(\mathbf{X}) = \{x_1^2, x_2^2, \sqrt{2}x_1x_2\}$$

Fig. 2. Principle of support vector machines.

1. Polynomial kernel of degree d:

$$K(X,Y) = (\langle X,Y \rangle + 1)^d \tag{4}$$

2. Radial basis function with Gaussian kernel of width C > 0:

$$K(X,Y) = \exp\left(\frac{-|X-Y|^2}{c}\right) \tag{5}$$

3. Neural networks with tanh activation function:

$$K(X,Y) = \tan h\left(K\langle X,Y \rangle + \mu\right) \tag{6}$$

where the parameters K and μ are the gain and shift.

3.1 Multiclass SVM

SVM was originally created for binary classification problems. Multiclass SVMs (when the number of classes $k \geq 3$) are usually implemented by combines several binary SVMs.

3.1.1 One-against-all Method

The problem of speaker identification can be formulated as a multiclass classification problem.

The earliest used implementation for SVM multiclass classification is probably the one-against-all (OAA) method (Christopher, 2006). It constructs k SVM models where k is the number of classes. The ith SVM is trained with all of the examples in the ith class with positive labels, and all other examples with negative labels.

3.1.2 One-against-one Method

Another major method is called the one-against-one method. It was first introduced in (Knerr *et al.,* 1990), and the first use of this strategy on SVM was in (Friedman, 1996; Kreßel, 1999). This method constructs $k(k-1)/2$ classifiers where each one trains data from two classes.

4. GAUSSIAN MIXTURE MODEL (GMM)

For speaker identification, each of S speakers is represented by GMMs $\lambda^1, \ldots, \lambda^{N_S}$, respectively.

Let λ^s be the stochastic model for the sth speaker derived from the training data of this speaker. We will have N_S stochastic models for the N_S speakers (one model for each speaker).

Let $Y(y_1, y_2, ..., y_L)$ be the sequence of the feature vectors representing the test utterance (having L frames). Our aim is to identify the speaker who has spoken this test utterance from the group of N_S speakers. This is done by computing the probability

$$p(Y \setminus \lambda^s) = p(y_1, y_2, ..., y_L \setminus \lambda^s) \tag{7}$$

for $s = 1,2,...,N_S$ and deciding the identity of the speaker on the basis of

$$s^* = \text{argmax}_{1 \leq s \leq N_S} p(Y \setminus \lambda^s) \tag{8}$$

If there is no correlation between the feature vectors of successive frames (i.e., they are independent), then Eq. (7) can be written as follows:

$$p(Y \setminus \lambda^s) = \prod_{i=1}^{L} p(y_i \setminus \lambda^s) \tag{9}$$

Thus, the task is to compute the probability of a test vector given the speaker model; i.e., $p(Y \setminus \lambda^s)$.

There are a number of methods recently proposed in the literature to compute this probability. The major one is the Gaussian Mixture Model (Reynolds, 1995a; Reynolds and Rose, 1995b; Reynolds, 1995c).

The motivation for the GMM comes from the need to model the acoustic space of a speaker in terms of a few acoustic classes a simple and reliable manner (Reynolds and Rose, 1995b). This is done by assuming the probability of a feature vector of the n^{th} frame $p(y_n \setminus \lambda^s)$ to be a linearly weighted mixture of M multidimensional Gaussian probability density functions (PDFs); i.e.,

$$p(y_n \setminus \lambda^s) = \sum_{i=1}^{M} p_i^s b_i^s (y_n) \tag{10}$$

Where $b_i^s(y_n)$ is the Gaussian PDF associated with the n^{th} mixture component (or, acoustic class) with mean n^{th} and covariance matrix Σ_i^s ; i.e.,

$$b_i^s(y_n) = \frac{1}{\sqrt{(2\pi)^d \Sigma_i^s}} e^{-\frac{1}{2}(y_n-\mu_i^s)^t (\Sigma_i^s)^{-1}(y_n-\mu_i^s)} \tag{11}$$

Where d is the dimensionality of the feature space. The mixture weights $b_i^s, i = 1,2,...,M$ in Eq. (10) satisfy the constraint $\Sigma_{i=1}^{m} p_s^i$. The covariance matrix used in Eq. (11) is assumed to be diagonal. This is done for the following two reasons:

1) It reduces the computational load (Reynolds, 1995c)

2) The cepstral features (normally used in speaker recognition systems) show a high degree of independence.

Collectively the s^{th} speaker's GMM model is represented by M components each consisting of $p_i^s, \mu_i^s, \Sigma_i^s$ (see Fig. 4); i.e.,

$$\lambda^s = \{p_i^s, \mu_i^s, \Sigma_i^s\}, \qquad 1 \le i \le M \tag{12}$$

The process of computing the probability of a feature vector given a GMM model is illustrated in Fig. 3.

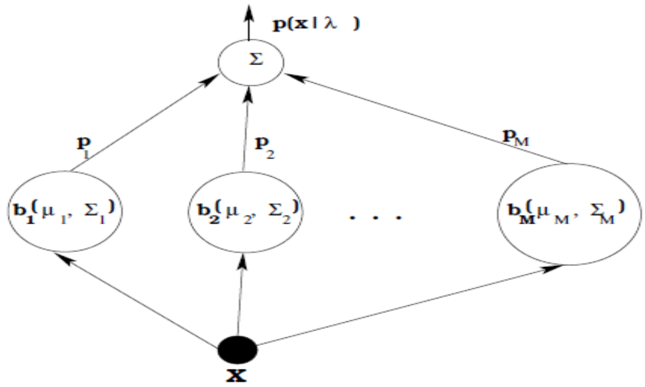


Fig. 3. The process of computing the probability of a feature vector given a GMM model

For estimating the speaker model parameters from the training data, the expectation-maximization (EM) algorithm is used. The EM algorithm uses a maximum likelihood procedure for computing the GMM model parameters. It consists of two steps: an E-step (Expectation) and an M-step (Maximization). Let us assume that we have I feature vectors $x_1, x_2, ..., x_I$ for a given speaker in the training data. The GMM model parameters for this speaker are initialized using a k-means algorithm, much like the one used in VQ. The EM algorithm for computing the GMM model parameters for the given speaker is given below. Note that we have dropped the speaker specific superscript s for clarity reasons.

The E-Step: Posterior probabilities are calculated for all the training feature vectors of the given speaker using

$$p(i \setminus x(n), \lambda) = \frac{p_i b_i(x(n))}{\sum_{k=1}^{M} p_k b_k(x(n))} \tag{13}$$

The M-Step: The M-step uses the posterior probabilities from the E-Step to estimate model parameters as follows:

$$\hat{p}_i = \frac{1}{I} \sum_{n=1}^{I} p(i \setminus x(n), \lambda) \tag{14}$$

$$\hat{\mu}_i = \frac{\sum_{n=1}^{I} p(i \setminus x(n), \lambda)\, x(n)}{\sum_{n=1}^{I} p(i \setminus x(n), \lambda)} \tag{15}$$

$$\hat{\Sigma}_i = \frac{\sum_{n=1}^{I} p(i \setminus x(n), \lambda)(x(n) - \mu_i)(x(n) - \mu_i)^t}{\sum_{n=1}^{I} p(i \setminus x(n), \lambda)} \tag{16}$$

Set $p_i = \hat{p}_i$, $\mu_i = \hat{\mu}_i$, $\Sigma_i = \hat{\Sigma}_i$ and iterate the sequence of E-step and M-step a few times till convergence is reached. The iterative process is normally carried out 10 times, at which point the model is assumed to converge to a local maximum

5. HYBRID APPROACH GMM/SVM

As mentioned before, in this paper we propose a hybrid classifier based on the combination of a generative model GMM and discriminative classifier SVM to achieve better classification and computation performances.

Generative GMM and discriminative SVM approaches have been successfully applied to many different problems and both have their own advantages and disadvantages.

Conventional speaker recognition systems use Gaussian mixture models (GMM) to model a speaker's voice based on the speaker's acoustic characteristics. GMM approaches try to find an optimal representation of the original data by keeping as much information as possible. Generative methods can be built very robustly.

But this method is categorized as a non-discriminative training process; as the model-building process does not take into account the negative examples of the speaker.

Gaussian mixtures have three sets of parameters to be adapted: mean vectors (centroids), covariance matrices, and weights. However, experiments have indicated that best results are obtained by adapting the mean vectors only (Reynolds, 2002). Thus improves the importance of mean vectors in GMM model.

Discriminative SVM method require fully labelled training data, can be applied very quickly and often show better recognition accuracy than their generative GMM counterparts.

The SVM is a discriminative model classification technique that mainly relies on two assumptions. First, transforming data into a high-dimensional space may convert complex classification problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions. Second, SVMs are based on using only those training patterns that are near the decision surface assuming they provide the most useful information for classification. The problem with SVM is that the computational burden is excessive compared to other competing methods such as the Gaussian mixture models (GMM).

Clearly, both approaches have their advantages and disadvantage. Several authors have tried to combine the approaches to benefit from the both.

In this paper we describe a new method of integrating discriminative classifiers like the Support Vector Machine (SVM) into speaker recognition environments and show that it is possible to use the SVM directly on the frame-level for datasets with a small amount of speech data. In this approach, speaker models are trained by EM (Expectation-Maximization) algorithms using mixture models and the means parameters of mixture models are used as training set of SVM.

The block diagram of the text-independent SI system proposed in this paper is shown in Fig. 4.

It comprises two phases: enrollment and identification. In the off-line enrollment phase, the first step is to train the GMM, utterances from the reference speaker are pre-processed and the features are extracted, from which the speaker model is trained and stored. The estimation of model parameters is performed by optimizing the likelihood of the training vectors corresponding to each speaker. Typically, the optimization is performed using the algorithm (Beigi, 2011). In the second step the mean vectors from all speaker models are collected in signal matrix F. These vectors are then treated as feature vectors when training an SVM. In the last step the SVM classifier is trained by a "one against all" algorithm for the S class problem. A reduced training set is formed by means vectors of GMM in order to reduce complexity time training, testing and benefice the GMM property. After the SVMs have been trained, the parameters of SVM classifier are stored.

In the on-line identification phase, the utterance of an unknown speaker, represented by a sequence of feature frame vectors $X = \{x_1, x_2, ..., x_L\}$ is evaluated with the SVM classifier. For each frame, the feature vector is classified by SVM Classifier. The output of SVM classier for each vector x_i is the index test score of the identified speaker r_{x_i} where $1 \leq r_{x_i} \leq$S (number of speakers). The outputs of each vector are collected in index test scores vector. Finally the index test scores vector of all frames are combined using an averaging step to give an overall utterance score from which the authenticity of the speaker can be determined.

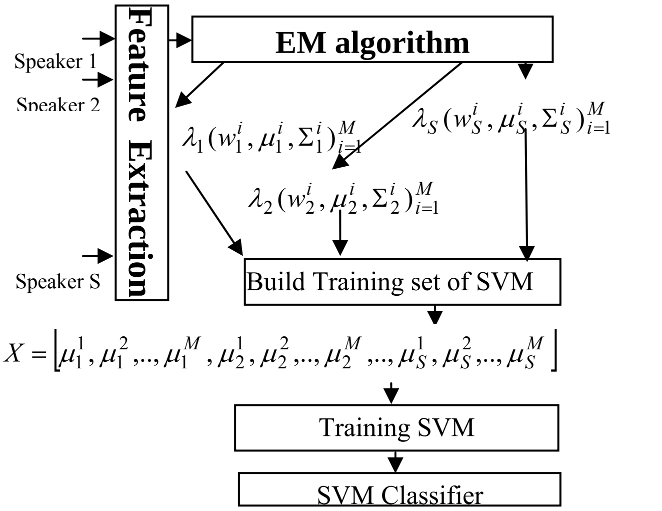

Speaker 1
Speaker 2
Speaker S
Feature Extraction
EM algorithm
$\lambda_1(w_1^i, \mu_1^i, \Sigma_1^i)_{i=1}^M$
$\lambda_2(w_2^i, \mu_2^i, \Sigma_2^i)_{i=1}^M$
$\lambda_S(w_S^i, \mu_S^i, \Sigma_S^i)_{i=1}^M$
Build Training set of SVM
$X = \lfloor \mu_1^1, \mu_1^2, .., \mu_1^M, \mu_2^1, \mu_2^2, .., \mu_2^M, .., \mu_S^1, \mu_S^2, .., \mu_S^M \rfloor$
Training SVM
SVM Classifier

Fig. 4. The block diagram of the proposed text-independent speaker identification system.

A speaker is recognized if, for the entire index test scores, it is selected more frequently than the other speakers.

6. EXPERIMENTAL RESULTS

In this section the database used for the experiments and the experiments performed are discussed. First, we briefly describe the IViE corpus, and then, present simulations results for a GMM-based speaker identification system, and preliminary results for a GMM/SVM-based speaker identification system. Finally, the experiments performed are discussed.

6.1 Speech Corpus

Performance of the proposed approach was evaluated through a number of text-independent speaker recognition experiments. All the experiments were conducted using a subset of the IVIE corpus named "*readpassage*". We focus on comparing and analyzing the GMM-SVM performance, as compared to that of the currently used GMM, and on exploring the effectiveness of GMM-SVM for speaker recognition systems.

The IViE (Intonational Variation in English) corpus contains recordings of nine urban dialects of English spoken in the British Isles. Recordings of male and female speakers were made in London, Cambridge, Cardiff, Liverpool, Bradford, Leeds, Newcastle, and Belfast in Northern Ireland and Dublin in the Republic of Ireland. Three of the speaker groups are from ethnic minorities: we have recorded bilingual Punjabi/English speakers, bilingual Welsh/English speakers and speakers of Carribean descent.

The speech corpus for the experiments reported in this paper is a subset of the IVIE corpus. The subset database "*readpassage"* is a collection of conversational speech from 12 speakers (6 male and 6 female). Each speaker has 5 conversations of approximately one minute each recorded during separate sessions: one for training and the other for testing.

6.2 Front-end Processing

Speech signals were sampled at 16000 Hz. Silence from the speech utterances is removed using an energy-based voice activity detector. A pre-emphasis filter $H(z) = 1 - 0.95z^{-1}$ is used before framing. Mel scale Frequency Cepstral Coefficients (MFCC) were employed as feature analysis (Shaughnessy, 2003; Shaughnessy, 2000). 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 20 triangular filters spaced on the Mel-scale. The log-compressed filter outputs are converted into cepstral coefficients by DCT giving 12 coefficients. The zeroth cepstral coefficient C_0 is not used in the cepstral feature vector and replaced with log of energy of the frame calculated in the time domain and appended to the feature vectors so that the resulting vector length is 13.

6.3 Baseline system

Speaker identification based GMM is a two stage procedure consisting of training and testing. The speaker identification system is operated in a text independent mode. Each enrolled speaker is trained using the speaker training speech. Each speaker is trained using EM algorithm to construct an M-mixtures GMM model.

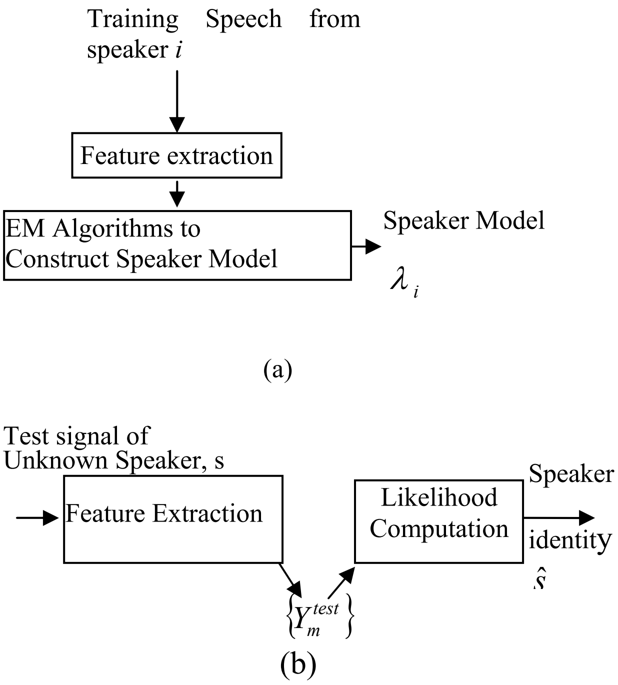


Fig. 5. (a) Training and (b) testing stages in speaker identification system based GMM modeling.

The GMM training procedure must be initialized with some starting model λ_0. The EM algorithm is guaranteed to find a local maximum likelihood model regardless of the starting point, but the likelihood equation for a GMM has several local maxima and different starting models can lead to different local maxima (Fine *et al.,* 2001). In this paper for the training, *k-means* clustering (Richard *et al.,* 2001) method are utilized to initialize the speaker models. To investigate the speaker identification performance of the GMM with respect to the number of component densities per model, the following experiment was conducted on a 32, 64, and 128, 256 component Gaussian densities were trained using 6000 13-dimensional Mel-cepstral vectors corresponding to one minute of speech.

In the SI testing stage M feature vectors x_m^{test} are extracted from a test signal (speaker unknown), scored against all S speaker models using a log-likelihood calculation, and the most likely speaker identity $\hat{s}$ decided according to

$$\hat{s} = \arg\max_{1 \le s \le S} \sum_{m=1}^{M} \log p(x_m^{test} / \lambda_s) \qquad (17)$$

Session one is used for model training and sessions 2, 3, 4 and 5 were used for testing. There is no particular reason why certain utterances were chosen for testing and training. This was done randomly.

All the GMMs were trained with 60 seconds of speech. For testing, we used segments of varied lengths: 1, 0.5 seconds.

For the testing of the systems carried out in this section (GMM 32,.., GMM 256), durations of 1s, 0.5s, 0.25s were utilized since the emphasis in speaker identification tasks is to capture the identity of a speaker with the minimum material in hand. Vectors obtained are 125, 62, and 31 corresponding respectively to durations above.

Table 1 show the percent correct identification performance versus the number of Gaussian components (N_{gmm}) for 0.5 and 1 second test utterance lengths (T_L) respectively.

Table 1. Identification rate (%) for 1 and 0.5 second test utterance length (TL).

T_L N_{gmm}	32	64	128	256
1 s	78.96	85.55	89.80	94.30
0.5 s	74.98	81.50	84.80	87.97

6.4 Hybrid System

For comparison, a GMM part of our GMM-SVM system has also been constructed following the same procedure of the GMM system. In the SVM part of the GMM-SVM system, the GMM means vectors are taken as input feature parameters in order to construct a small data set for training.

Also, to evaluate the effectiveness of the GMM/SVM solution, have compared the recognition accuracy, model size, and execution time of the GMM/SVM solution with GMM speaker identification solution.

6.5 Recognition accuracy

Results are summarized in table 2, 3 and 4 giving recognition rates (%) of a speaker recognition system varies with the number of Gaussian mixtures(M) and second test utterance length (T_L)

Table 2. Identification rate for 1 second test utterance length (TL=1 s).

N_{gmm}	32	64	128	256
GMM	78.965	85.55	89.80	94.30
GMM/SVM	90.71	93.55	94.98	95.27

Table 3. Identification rate for 0.5 second test utterance length (TL=0.5 s)

N_{gmm}	32	64	128	256
GMM	74.98	81.50	84.80	87.97
GMM/SVM	82.92	87.54	88.81	90.30

Table 4. Identification rate for 0.25 second test utterance length (TL=0.25 s).

N_{gmm}	32	64	128	256
GMM	64.58	72.06	76.83	80.76
GMM/SVM	75.21	80.24	83.10	85.75

6.6 Model size and Execution time

In many applications the execution time and storage requirements of a system plays a very important part in determining its practicality. In this section we will look at the execution time of the GMM/SVM proposed system in terms of training and recognition times.

Training time, testing time and storage requirements of the GMM/SVM proposed system depend on the number of total vectors in the training set, support vectors respectively.

Table 5 below shows the execution time of each of the two systems during training and testing step, using 1.8 GHz Core2Duo Intel Desktop Computer machine.

Table 5. Execution times of GMM and GMM/SVM.

Algorithm		Execution time	
		Training	Test
GMM		~ 10mn	0.283345 s
GMM/SVM	GMM part	~10mn	0.020460 s
	SVM part	~0.117321mn	

Tables 6-3 to 6-5 below show the memory requirements of each sub-system

Table 6. Model sizes of GMM and GMM/SVM.

Algorithm	Model size(kB)
GMM	699
GMM/SVM	483

6.7 Discussion and Conclusion

There are several observations to be made from these results. We tested our system's accuracy varying the number of Gaussian mixtures. Table 1 and 2 shows how the accuracy of a speaker recognition system varies with the number of Gaussian mixtures. The GMM and GMM/SVM trained using RBF kernel (σ = 0.1, C=30) maintains high identification performance with increasing number of Gaussian mixtures. It can be easily observed that GMM/SVM has provided a better performance than GMM for speaker identification task for all sessions.

Table 2, 3 and 4 show the experimental speaker identification rate as a function of the test input utterance length and illustrates the effect of the test input utterance length. The identification rates clearly decreases with respect to the test input utterance size for both GMM and GMM/SVM systems It is interesting to observe that the new method outperforms the conventional GMM method for all the utterance lengths considered. The ranges differences in identification rate between the GMM (baseline) and our new method it from 0.93% to 11.74% absolute with various lengths of test input utterance.

In addition, is remarkable that for the identification rate case, the difference between the GMM256 and our new method (GMM/SVM256) are respectively 0.93, 2.33 and 4.99 with the following corresponding lengths of test input utterance 1, 0.5, 0.25. These results experimentally confirm that the new method can yield robust speaker identification with short data segments.

We can see also, the rate of our method SVM/GMM256 with 0.25 s utterances (85.75%) is almost the same as the identification rate of the conventional method GMM64 with 1 s utterances (85.55%). It means that we can achieve about 85% accuracy with one fourth the length of utterances using the new method.

As previous mentioned in section(5) , the main disadvantage of applying discriminative learning SVM directly using full training set is that the number of training examples can be too large and the problem with SVM is that the computational burden is excessive compared to other competing methods such as the Gaussian mixture models (GMM).

For example, SVM training time scales with the square of the number of training examples, while GMM training scales linearly.

The other limitations is extra unknown parameters that are to be specified such as the standard deviation (scale) parameter and the margin of error parameter C in the case of the Gaussian kernel as used for this experiments . The experiments took a very long time to complete as such only region (train set) of the database that was used. To obtain a single value of performance took over 3 hours on a 1.8 GHz Core2Duo Intel machine and it is dependent on the amount of enrolment data.

In the our new GMM/SVM system the above problems are efficiently reduced because the training set's size in SVM part of our GMM-SVM system depend only the number of Gaussian mixtures(M) and the number of speaker(S). For example standard SVM and GMM/SVM speaker identification systems are trained using (S×6000) 13-dimensional mel-cepstral vectors and (M×S) 13-dimensional vectors respectively corresponding to one minute of speech i.e. 120000 vectors for standard SVM and 5120 vectors for GMM/SVM with 256 mixtures.

In GMM/SVM system, the computational cost is proportional to the number of kernel evaluations i.e. to the number of SVs

However, in the GMM method, the major computation loads are the likelihood computation for all mixtures for the speaker model. Such a system uses the majority of the processing power for scoring the Gaussian densities. For large population speaker identification (SI) systems, likelihood computations between an unknown speaker's test feature vectors and speaker models can be very time-consuming and detrimental to applications where fast SI is required.

Table 5 shows that the average training time of the GMM/SVM is longer than that of GMM. The average testing time of the GMM/SVM, however, is much shorter than that of GMM. The GMM/SVM has a great advantage in either real time applications.

Table 6 shows the average model sizes of the two systems. While the model sizes of GMM and GMM/SVM were 699 Kbytes and 483 Kbytes, respectively, the GMM/SVM used only 483 Kbytes to store the model. Therefore, the GMM/SVM has a great advantage in either large applications where millions of people are enrolled or smart card applications where only a few Mbytes of RAM is available. By using this memory-efficient GMM/SVM algorithm, we have successfully realized a real-time application system for speaker identification.

In speaker identification applications, the accuracy and computational load are two major criteria for the selection of a proper system.

GMM/SVM has lower computational complexity and less memory requirement compared to the GMM of the same model order.

Another important issue of proposed system is the choosing optimal parameters for support vector machines (the penalty parameter C and the kernel parameter of the RBF function) use the cross validation technique can be quickly determined because the training time used in very smaller. Choosing optimal parameters for support vector machines is an important step in SVM design

7. CONCLUSION

A simple and efficient statistical speaker identification method has been introduced in this paper. The focus of this work is mainly on applications which require high identification rates and less computation using hybrid approach to take advantage of GMM and SVM approaches. In this paper we introduce a solution to combine GMM means vectors for constructing the SVM training data set to prevail over an important weakness of SVM in large scale databases. Therefore we use SVM for classifying test segment. The hybrid system is very promising in both recognition rate and computational complexity aspects.

REFERENCES

Beigi, H. (2011). *Fundamentals of Speaker Recognition. Springer.*

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*, Springer

Chaudhari, al. (2001). Very large population text-independent speaker identification using transformation enhanced multi-grained models. *In Proc.ICASSP*, volume 1, p. 461-464.

Boujelbene, S.Z., Mezghani, D., Ellouze, N. (2009). Support vector machines approaches and its application to speaker identification. *Digital Ecosystems and Technologies(DEST)*, pp. 662-667.

Fenglei, H. (2000). An integrated system for text-independent speaker recognition using binary neural network classifiers. *In Signal Processing Proceedings, WCCC-ICSP 2000.* Volume 2, pp. 710 - 713

Friedman, J. (1996). Another approach to polychotomous classification. *Technical report*, Department of Statistics, Stanford University.

Guo, G., and Li, S. Z. (2003). Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, volume 14 (1), p 308–315.

Jiang, H., Bai, J., Zhang, S., and Xu, B. (2005). SVM-based audio scene classification. *Proc of the IEEE*, p. 131–136.

Jaakkola, T., Hausser, D. (1998). Exploiting generative models in disciminative classifiers. *In Advances in Neural Information Processing Systems*, pp. 487-493

Knerr, S., Personnaz, L., and Dreyfus, G.(1990). Single-layer learning revisited: a step-wise procedure for building and training a neural network. *In Neuro-computing: Algorithms, Architectures and Applications*, J. Fogelman (eds.), Springer-Verlag, 1990, Vol.F68, p. 41-50.

Picone, J. (1993). Signal Modeling Techniques in Speech Recognition. *In Proc.IEEE*, volume 81 (9), p. 1215-1247.

Reynolds, D. (2002). An overview of automatic speaker recognition technology. In *Proc. ICASSP*,.Orlando.FL., p. 4072-4075

Reynolds, D.A.(1995a). Automatic speaker recognition using gaussian mixture speaker models. *Lincoln Laboratory Journal*, volume 8 (2), p. 173-192.

Reynolds, D.A., Rose, R.C.(1995b). Robust text independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, volume 3 (1), p. 72-83.

Reynolds, D.A(1995c). A gaussian mixture modeling approach to text independent speaker identification, *Ph.D. Thesis*, Georgia Institute of Technology.

Reynolds, D.A., Dunn, R.B., McLaughlin, J.J.(2000). The Lincoln Speaker Recognition System: NIST Evaluation 2000. In *Proc.ICSLP*, Beijing, china.

Richard, O., Peter, E., David G. (2001) *Pattern classification*, Wiley-interscience

Ulrich, H., Kreßel, G.(1999) Pairwise classification and support vector machines. In *Advances in Kernel Methods Support Vector Learning*, Schölkopf B, Burges C J C, Smola A. J (eds.), Cambridge, MA, MIT Press, p. 255–268

Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley and Sons.

Weber, F., Beskin, B.(2000). Speaker recognition in two-speaker data: recent results from dragon systems. In *Proc.ICASSP*, Istanbul, volume 2, pp. 1205 -1208.

Zhong, Yuan X., Yuan, Bo-Ling X., Chong, Zhi Y. (1999). Binary Quantization of Feature Vectors for Robust Text Independent Speaker Identification. *IEEE Transactions on Speech and Audio Processing*, volume 7 (1), pp. 70 - 78