*Comparative Study of hybrid models for robust speaker recognition task**

Kawthar Yasmine ZERGAT, Abderrahmane AMROUCHE

Labo de Communication Parlée & Traitment du Signal. (LCPTS) Faculté d'informatique et d'électronique, USTHB, Bab Ezzouar.

Abstract: *This paper deals with text-independent speaker verification system based on spoken Arabic digits in real environment. In this work, we adopted Mel-Frequency Cepstral Coefficients (MFCC) as the speaker speech feature parameters, Gaussian Mixture Model (GMM) are used for modeling the extracted speech feature and training the support vector machines (SVMs). The experiments were conducted on the ARADIGIT database at different Signal-to-Noise Ratio (SNR) levels and under two noisy conditions issued from NOISEX-92 database. The obtained results show that the GMM-SVM model outperforms the GMM-UBM, especially in noisy environments.*

Résumé : *Cet article traite du cas du système de vérification des textes de locuteurs indépendants sur la base des chiffres énoncés en langues arabes dans un environnement réel. De ce fait, nous avons adopté les coefficients cepstraux de fréquence Mel (MFCC) comme paramètres caractéristiques du discours, le modèle de mélange gaussien (GMM) pour modeler les caractéristiques du discours extrait et avons testés les machines vecteurs de support (SVM). Les tests ont été menés sur la base de données ARADIGIT à différents niveaux du rapport signal sur bruit (SNR) et sous deux conditions bruyantes émises par la base de données NOISEX-92. Les résultats obtenus démontrent que le modèle GMM-SVM surpasse le modèle GMM-UBM plus particulièrement dans un environnement bruyant..*

Keywords: *Speaker verification, MFCC, GMM-UBM, GMM-SVM, Noisy environment.*
Mots clés *: Vérification du locuteur, MFCC, GMM-UBM, GMM-SVM, environnement bruyant.*

**Etude comparative des modèles hybrides pour une reconnaissance vocale robuste*

1 Introduction

Speech signal provides several levels of information. It conveys the words and messages being spoken and also provides the identity of the speaker [1].Speaker recognition is the process of automatically recognizing a user's claimed identity using characteristics extracted from their voices. This is in contrast with speaker verification, where the goal is to verify the person's claimed identity based on his or her utterance [2]. Speaker recognition systems can be classified into text-dependent systems and text-independent systems. Text-dependent speaker recognition systems require that the speaker utter a specific phrase or a given password where text independent speaker recognition systems identify the speaker regardless of his utterance [2]. This paper deals with text-independent speaker verification system.

The GMM-UBM system is the current state-of-the-art for text-independent speaker verification. The advantage of this approach is that both target speaker model and impostor model (UBM) have generalization ability to handle “unseen” acoustic patterns.

In this work we describe the GMM [3] and SVM [4] models and focus on the hybrid ones, which are GMM-SVM [5-6] and GMM-UBM [7-8] used for the verification task.
The main goal of this paper is to evaluate the robustness of these two hybrid systems and to investigate on the performance degradation in adverse conditions conducted on the ARADIGIT database.
The remainder of the paper is structured as follows. In section II, we briefly describe the GMM and GMM-UBM classification methods. In section III we discuss the principles of SVM and GMM-SVM. The experimental protocols used in this work are described in section IV where, experimental results of the speaker verification task in quiet and noisy environment using GMM-UBM and GMM-SVM systems based on the Arabic ARADIGITS database are presented in section V. Finally, a conclusion is given in Section VI.

2 GMM and GMM-UBM

In the Gaussian Mixture Model *(GMM)* [3], the distribution of the parameterized speech vector of a speaker is modeled by a weighted sum of Gaussian densities:

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

Where x is a D-dimensional cepstral vector, λ is the speaker model, $b_i(x)$, $i=1,...,M$ are the component densities characterized by the mean μ_i and the covariance matrix Σ_i and $p_i, i=1,...,M$ are the mixture weights. Each component density is a D-variate Gaussian Mixture function of the form:

$$b_i(x)=\frac{1}{2\Pi^{D/2}\left|\Sigma_i\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_i)'(\Sigma_i)^{-1}(x-\mu_i)\right] (2)$$

The model parameters $\lambda\{\mu_i,\Sigma_i,\Pi_i\}\ i=1,...,M$ are estimated by an Expectation-Maximization (EM) algorithm [9]. It is also used to find the UBM model parameters (mean, variance and weight) by pooling the data from all the speakers' utterances. The hypothesized speaker specific model is derived by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation [10].

The specifics of the adaptation are as follows [10]. Given a UBM and statistically independent T observations feature training vectors from the hypothesized speaker, $X = \{x_1, x_2, ..., x_t\}$, we first determine the probabilistic alignment of the training vectors into the UBM mixture components. That is, for mixture i in the UBM, we compute

$$\Pr(i / x_t) = \frac{\lambda_i p_i(x_t)}{\sum_{j=1}^{M} \lambda_j p_j(x_t)} \qquad (3)$$

$$n_i(X) = \sum_{t=1}^{T} \Pr(i / x_t) \qquad (4)$$

$$E_i(X) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i / x_t) x_t \qquad (5)$$

This is the same as the expectation step in the EM algorithm. Finally, these new sufficient statistics from the training data are used to update the old UBM sufficient statistics for the mixture i to create the adapted mean parameter for the mixture i with the equations:

$$\bar{\mu}_i = \alpha_i E_i(X) + (1 + \alpha_i)\mu_i, i = 1, ..., M \qquad (6)$$

$$\alpha_i = \frac{n_i(X)}{n_i(X) + r} \qquad (7)$$

Where r is a fixed relevance factor r=16[9].

In order to identify a speaker from a group of speaker $S = \{1, 2, ..., s\}$, each speaker is represented by its corresponding model derived from the UBM model by MAP adaptation [10]. For Speakers, the corresponding MAP adapted models can be represented as $\lambda_1, \lambda_2 ... \lambda_s$.

Now the speaker model that maximizes the a posteriori probability for a given sequence of speech utterancescan be written as

$$\hat{S} = \arg\max_{1 \le k \le S} \frac{p(x / \lambda_k) p(\lambda_{UBM})}{p(x / \lambda_{UBM})} \qquad (8)$$

Where Equation 1 is in the form of Bayes rule, for equally likely speaker classes, $p(\lambda_{UBM}) = 1 / S$ and the denominator $p(x / \lambda_{UBM})$ is the same for all speaker models, this reduces the previous equation to a simple form of Log Likelihood detector [3] as follows:

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p(x_t / \lambda_k) \qquad (9)$$

3 SVM AND GMM-SVM

The Support Vector Machines (SVM) are a powerful technique of the statistical learning theory proposed by Vapnick in 1995 and developed from the Structural Risk Minimization (SRM) theory. They can address diverse problems as classification, regression, fusion, etc.

The basic idea of SVM is to project data from the input space, belonging to two different classes, non-linearly separable into a space with larger size called feature space so that data becomes linearly separable. In this space, the construction technique of the optimal hyperplane is used to calculate the function of classification between the two classes. The discriminant function of SVM is given by [11]

$$f(x) = class(x) = sign\left[\sum_{t=i}^{N} \alpha_i y_i K(x, x_i) + b\right] \qquad (10)$$

Here $y_i \in \{-1,+1\}$ are the ideal output values, and $\sum_{t=i}^{N} \alpha_i y_i = 0$ with $\alpha_i \geq 0$. The support vectors x_i, their corresponding weights α_i and the bias term b, are determined from a training set using an optimization process. The kernel function $\mathrm{K}(\cdot,\cdot)$ is designed so that it can be expressed as $K(x, y) = \phi(x)^T \phi(y)$, where $\phi(x)$ is a mapping from the input space to kernel feature space of high dimensionality.

A GMM supervector is constructed by stacking the means of the adapted mixture components [**6**] from the UBM model as follow.

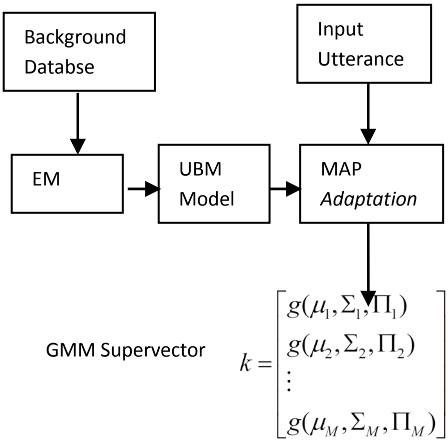
Background Databse
Input Utterance
EM
UBM Model
MAP *Adaptation*
GMM Supervector
$k = \begin{bmatrix} g(\mu_1, \Sigma_1, \Pi_1) \\ g(\mu_2, \Sigma_2, \Pi_2) \\ \vdots \\ g(\mu_M, \Sigma_M, \Pi_M) \end{bmatrix}$

Figure 1: Process of generating the GMM-supervector

4 Experimental protocol

a. Description of data base

While spelling the ten digits in Arabic language we already produce an interesting number of Arabic phonemes. It can be considered as a relative representative elements of this language, which has several specificities, such as germination, emphasis, and sound duration etc.

The database used in this work is a part of ARADIGIT database [12].This database comprises the recording of Algerian Arabic speakers aged between 18 and 50 years from different regions of Algeria. It consists of a set of 10 digits of the Arabic language (zero to nine) spoken by 62 speakers, 31 male and 31female, where each speaker repeats the same list 3 times. All recordings were made in acoustically prepared environment with ambient noise level below 35 dB, using the same microphone. Files were acquired at a sampling rate of 22,050 kHz, and then were downsampled to 16 KHz, in WAV format.

Two independent databases were created: training and testing. We have concatenated the sequences of eights numbers (from zero to seven) for training and used a sequence of two numbers (eight and nine) for testing phase, with three repetitions for each sequence.

b. Parameterization phase

In parameterization phase, we specified the feature space used. This space is defined by vectors of fixed size. Indeed, as the speech signal is dynamic and variable, we presented the observation sequences of various sizes by vectors of fixed size. Each vector is given by the concatenation of the Mel Frequency Cepstrum Coefficients MFCC (12 coefficients), their first and second derivatives are computed and appended to the feature vectors so that the resulting vector length is 36.

To reduce the effect of noisy environments, we applied to the feature vectors a Cepstral Mean Subtraction (CMS). In this method, Cepstral coefficients are averaged over the duration of an entire utterance, and the averaged values are subtracted from the Cepstral coefficients of each frame. This method can compensate fairly well for additive variation in the log spectral domain.

c. Modeling phase

As described in section III, the GMM-SVM method works by using a GMM supervector consisting of the stacked means vectors of MAP-adapted of 16 Gaussians mixture model GMM, that captures the acoustics characteristic of a speaker, the supervector is then presented to a speaker-dependent SVM for scoring.

In order to simulate the impostors, a gender balanced UBMs consisted from 200 speakers (100 male and 100 female), specifics to the corpora: TIMIT database were trained. The models used 2048 mixture components and were trained using EM algorithm. The full background training dataset was made with five sequences spoken in English by each speaker. For better performance, only the mean vectors are adapted.

In the second part of our work, two types of additive noise produced by Speech Babble and Factory production vehicle, reaching high levels of SPL and derived from the NOISEX-92 database (NATO: AC 243/RSG 10) are added to the test speech signal issued from the Arabic ARADIGIT database.

d. Classification phase

In this work, we used the Equal Error Rate (EER) as the evaluations metric performance of the hybrid systems. The error probabilities miss (rejection of a genuine speaker) and false alarm (acceptance of an impostor) are then plotted as Detection Error Trade-off (DET) curves to show the system's performance.

5 Experiment results

5.1 Speaker verification in quite environment using GMM-UBM and GMM-SVM

The following figure presents the performances of these two hybrid methods in term of equal-error rate (EER) shown by DET curve.

In clean environment, a slight superiority of the GMM-UBM model with an EER equal to 1.61% against the GMM-SVM, EER=3.64% is noticed.

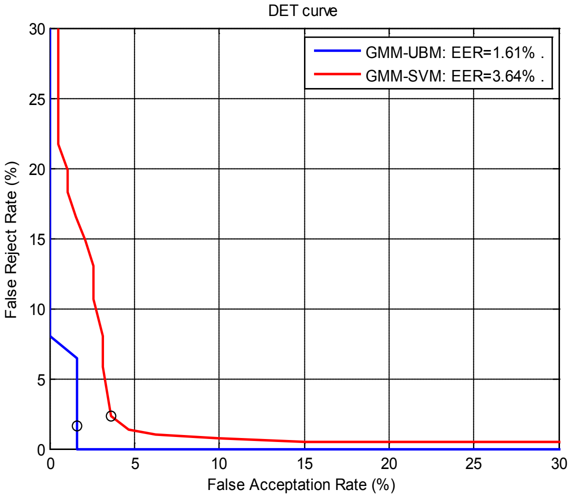

Figure 2: Performances evaluation for GMM-UBM and GMM-SVM based speaker verification systems in clean environment.

5.2 Speaker verification in realistic environment using GMM-UBM and GMM-SVM

The main goal of the experiments done in this section is to evaluate the speaker verification performances of GMM-UBM and GMM-SVM when the speech data is corrupted.

Table1 and Table2 present the experimental results obtained in adverse conditions, such as Babble speech and factory, noises. Clearly, it is seen that the GMM-SVM gives better performance in hard conditions. For example, the EER=17,66% for Babble speech noise which is more than 43% obtained for the GMM-UBM system.

Table1 EERs in speaker verification for GMM–UBM under mismatched data conditions using real world noise.

GMM-UBM				
SNR(dB) Noise	Tested Data			
	SNR=0	SNR= 5	SNR=10	SNR=15
Babble- speech	43.54	22.58	9.76	4.83
Factory	14.51	8	6.45	6.45

Tableau 2: EERs in speaker verification for GMM–SVM under mismatched data conditions using real world noise.

GMM-SVM				
SNR(dB) Noise	Tested Data			
	SNR= 0	SNR=5	SNR=10	SNR=15
Babble-speech	17.66	17.23	13.63	12.5
Factory	12.09	10.34	9.67	9.37

In the Gaussian mixture model, the classes are represented by Gaussians distributions which create such a confusion area between these classes. This explains why the maximum likelihood scoring doesn't converge to optimal classification making the optimal decision boundary difficult to find. Especially in noisy environment, this area increase, which explain the degradation of the GMM-UBM system in hard conditions (SNR =0).

It is not the case for the GMM-SVM model, where the supervectors issued from GMM-UBM are weighted by Lagrange multipliers α_i which eliminates the zone of confusion by finding a good hyperplane to separates classes (see section III, Fig.1.). This explains the robustness of the GMM-SVM model in such noisy conditions.

6 Conclusion

The main goal purpose of our work is to establish two speaker verification systems, GMM-SVM and GMM-UBM systems suitable for Arabic language. The results obtained for text independent speaker verification task are very satisfying in the case of the GMM-SVM in noisy environments. The results are very encouraging and deserve to be applied to a larger Arabic database.

7 Références

M.S. Sinith, S.Anoop, K..GowriSankar, K.V. Sandeep Narayanan, S. Vishnu, “A Novel Method for Text-Independent Speaker Identification Using MFCC and GMM,” Kerala, India, 2010.

Y.Cheang Soon, A.Abdul Manan, “Malay Language Text-Independent Speaker verification using NN-MLP classifier With MFCC,”International Conference on Electronic Design, Penang, Malaysia, 2008.

D.Reynolds, T.Quatieri, R.Dunn, “Speaker verification using adapted gaussian mixture models,” Digital Signal Process. 2000,10 (1), 19–41.

W.Campbell, J.Campbell, D.Reynolds, E.Singer, “a Support vector machines for speaker and languagerecgnition,”Comput. Speech Lang. Torres-Carrasquillo,P.,20 (2–3), 2006, 210–229.

Campbell W.M, Sturim D. E, Reynolds D. A, Solomonoff A,. “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” IEEE, 2006.

R. Dehak, N. Dehak,P. Kenny, P.Dumouchel, “Linear and non linear kernel GMM supervector machines for speaker verification,”Proc. Interspeech, Antwerp, Belgium, 2007, pp. 302–305.

Campbell, W.Sturim, D.Reynolds, “Support vector machines using GMM supervectors for speaker verification,”IEEE Signal Process. Lett. 13 (5), 2006, 308–311.

Minghuil., X. Yanlu, Y. Zhigianng, D. Beigian, “A new hybrid GMM/SVM for speaker verification,” Proc. 18th Int. Conf. Pattern Recognition, 4, 2006, pp. 314–317.

C.Bishop,”Pattern Recognition and Machine Learning,”Springer Science+Business Media, LLC, New York, 2000.

D. Reynolds, T.Quatieri, R.Dunn, “Speaker verification using adapted gaussian mixture models,”Digital Signal Process, 2000,10 (1), 19–41.

B.Yegnanarayana, S.Kishore, ”AANN: an alternative to GMM for pattern recognition,” Neural Networks 15,2002, 459–469.

A. Amrouche, “Automatic speech recognition using connectionist models (Reconnaissance automatique de la parole par les méthodes connexionnistes),” Doctoral thesis, Faculty of Electronics and Computer Science, USTHB, 2007.