2nd International Conference on Advanced Technologies
for Signal and Image Processing - ATSIP'2016
March 21-24, 2016, Monastir, Tunisia

Improving Text-independent Speaker Recognition with GMM

Rania Chakroun[1,4], Leila Beltaïfa Zouari[1,2], Mondher Frikha[1,3], and Ahmed Ben Hamida[1,4]
[1]Advanced Technologies for Medicine and Signals (ATMS) Research Unit
[2]National School of Engineering of Sousse, Sousse, Tunisia
[3]National School of Electronics and Telecommunications of Sfax
[4]National School of Engineering of Sfax, Sfax, Tunisia

Abstract— **The Gaussian mixture models (GMM) represent an efficient model that was broadly used in most of speaker recognition applications. This study introduces a novel method for speaker verification task. We propose a reduced feature vector employing new information detected from the speaker's voice for performing text-independent speaker verification applications using GMM. We use the power spectrum density of the speech signal to improve the system's performance. Speaker verification experiments were evaluated with the TIMIT dataset. The suggested system performance is evaluated against the baseline systems. The decrease in the error rate is well observed and the results have demonstrated the effectiveness of the new approach which avoids the use of more complex algorithms or the combination of different approaches.**

Keywords—GMM, speaker verification, speaker recognition, speaker identification

I. INTRODUCTION

Nowadays, speaker recognition applications are broadly used in several fields. Speaker recognition focus on the process of recognition of a person speaking, taking into account people's speech recordings, which provide specific information about each speaker. This method permits for a speaker to use his voice like identity verification for many purposes such as voice operators, shopping, telephone transactions, and also information or database access, voice mail, remote access computers and security check for some confidential information areas.

The main goal of speaker recognition is to facilitate the individual's everyday life and ameliorate some applications, such as the field of telephone bankind or information services. Speaker recognition is now considered as a strong security component for confidential areas access [3]. For example, a person's unique voice can't be obtained in any way including computer hacking skills, such as passwords. The possible harm would only happen with stealing a person's sample. Considering that security areas use speaker verification systems, so an intrusion requires to be recorded in a noise free environment of an individual's voice. Therefore it is very unlikely to happen. Nowadays, some speaker imitation systems exist and permit to record a person's voice for application with any speech in order to make them pronounce anything, but these systems are still being developed.

In order to have powerful application, current speaker receognition system require a quality recording environment with as large as possible of a set of training and testing data. A more extensive speech database increases the chance of matching during the test phase. There are also some other technical parameters that can be taken into account, which alter the system's effectiveness. The main factors are related to the approach used and the features employed. The system used in this article has been developed using the well-known state-of-the-art GMM [7], [10] approach. Most of the works in this area focus on the use of cepstral coefficients[11]. This work focuses on determining whether power spectral information is useful for improving current automatic speaker verification systems.

For this purpose, a new text-independent speaker verification systems is suggested, which uses both spectral features and power spectral information. The system's performance is compared with the baseline approaches.

This study is organized as follows: First the approach utilized for our speaker verification system is explained. Then, we give a description about the dataset used, the experimental protocol and the results obtained. We evaluate our system's performance against the baseline speaker verification systems. Finally, we will provide a conclusion illustrating the main matter of the new system for speaker verification task.

II. THE GMM APPROACH

A mixture of Gaussian probability densities corresponds to a weighted sum of *M* densities, as illustrated in figure 1 and is represented as follows

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i \, b_i \, (\vec{x}) \qquad (1)$$

978-1-4673-8526-8/16/$31.00 ©2016 IEEE

where $\vec{x}$ corresponds to a random vector having dimension D, $b_i(\vec{x})$, i=1,…,M, represent the density components, and the mixtures weights are p_i, i=1,…,M. Each component density can be calculated as a D variate Gaussian function having $\vec{\mu}_i$ as a mean vector and K_i as a covariance matrix:

$$b_i(\vec{x}) = \frac{e^{(-\frac{1}{2}(\vec{x}-\vec{\mu})^T (K_i)^{-1}(\vec{x}-\vec{\mu}))}}{(2\pi)^{D/2}\sqrt{|K_i|}} \quad (2)$$

The weighting of the mixtures should satisfy that $\sum_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is then parameterized by a covariance matrix, vector of means, and a weighted mixture of all component densities denoted as λ model. To represent these parameters, the following notation is used:

$$\lambda = \{p_i, \vec{\mu}_i, K_i\}, i=1,\ldots, M. \quad (3)$$

The GMM may have various forms based on the of the covariance matrix chosen. In fact, the model can have a covariance matrix per Gaussian component like it is indicated in equation 3. In this case, we use the nodal covariance. The model can have also a covariance matrix for all Gaussian components for a given model, and then we obtain a grand covariance. It is also possible that only one covariance matrix is shared by all models, called global covariance. A covariance matrix can also be diagonal or complete [9].

Gaussian components act in conjunction to model the probability density function. For that, the complete covariance matrix is almost not essential.

The effect of the use of a set of complete covariance matrices corresponds to the same as the use of a larger set of diagonal covariance matrices [6].

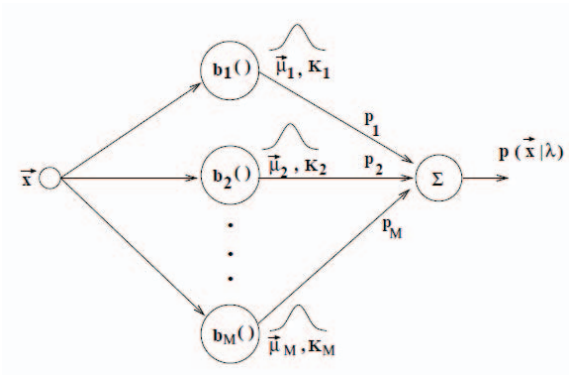

Fig. 1. M component Gaussian mixture density forming a GMM.

To search the maximum likelihood, we use the technique introduced in [4]. Given a sequence of T independent training vectors $X = \{\vec{x}_1, \ldots \vec{x}_T\}$, the GMM likelihood is calculated as

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \quad (4)$$

The likelihood for modeling the claimed speaker (model λ) can be directly calculated with

$$\log p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \quad (5)$$

The scale $\frac{1}{T}$ is employed for the normalization of the likelihood for utterance duration.

The speaker verification application need to make a binary decision, even it accepts or rejects the pretence speaker. A speaker verification system is presented by the following figure.

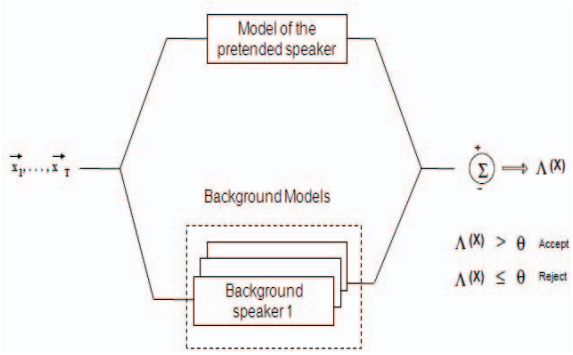

Fig. 2. Speaker verification system with GMM.

The verification system uses a likelihood ratio test to an input speech sequence in order to detect whether the claimed speaker is true or false. Indeed, for an input vector X = $\{\vec{x}_1, \ldots \vec{x}_T\}$., and a claimed speaker having a model λ_c, the likelihood ratio is as follows

$$\frac{p(X \text{ is from the claimed speaker})}{p(X \text{ is not from the claimed speaker})} = \frac{p(\lambda_c|X)}{p(\lambda_{\bar{c}}|X)} \quad (5)$$

With the application of Bayes' rule, the likelihood ratio is transformed into

$$\Lambda(X) = \log p(X|\lambda_c) - p(X|\lambda_{\bar{c}}) \quad (6)$$

The likelihood ratio between the pretence speaker model and other models (back ground models) is evaluated against a given threshold θ [9]. The claimed speaker is accepted only if $\Lambda(X) > \theta$. The threshold is then determined basing on experimental results.

III. Experimental Evaluation

A. TIMIT Database

In this work we consider speaker verification task for TIMIT database [12]. TIMIT contains in totality 6300 sentences with 10 utterances spoken by each one of the 630 speakers. The speakers are from the 8 major dialect regions of the United States. The speech signal was sampled at 16 kHz sampling frequency.

B. Experimental setup

All evaluations are dealt with 64 speakers selected from all the regions of TIMIT database. The speakers are selected as 8 speakers from each region with 4 male and 4 female speakers from each region. To Follow the protocol presented in [9], we divide the sentences recorded from each speaker into 8 utterances for training task (two SA, three SX and three SI sentences) and the remaining 2 utterances (two SX sentences) for the test task.

Mel Frequency Cepstral Coefficients (MFCC) are employed to extract features from the speech signal. These features proved their success in speaker recognition domain since MFCC carry the frequency distribution identifying sounds in addition to the vocal tract shape and length, which are speaker specific features.

In this work, MFCC feature are used, since they are the most popular choice for any speaker recognition system. Our experiments operate on cepstral features extracted from the speech signal with a 25-ms Hamming window. Every 10 ms, 12 MFCC together with log energy were calculated. Then Delta and delta-delta coefficients are calculated to produce 39-dimensional feature vectors. Indeed, this MFCC feature vector constitutes one of the most broadlly used vectors to this day [3], [5]. After that, another set of experiments are dealt with another popular MFCC feature vector [1], [2]. That's why experiments are also conducted with 19-dimensional MFCC feature vector together with log energy, delta and delta-delta coefficients to construct 60-dimensional feature vectors. The features were extracted due to Hidden Markov Model ToolKit (HTK) [8].

Since realistic applications suffer from some constraints like computational resource limitation or reduced memory space, we decide to look for improved approach using more reduced feature vectors and ameliorating the system's performance. We find that the inclusion of new information extracted from the speech signal which is power spectrum density of the signal with reduced MFCC feature vector dimensions can improve the system's performance and give significant resuls. For that, MFCC vector are combined with power spectrum density of the signal. The new structure of the vectors is evaluated and compared with traditional MFCC vectors. The number of mixture components is varied from 1 to 256 mixtures and the Equal Error Rates (EER) given with different feature vectors are plotted in Figure 3.

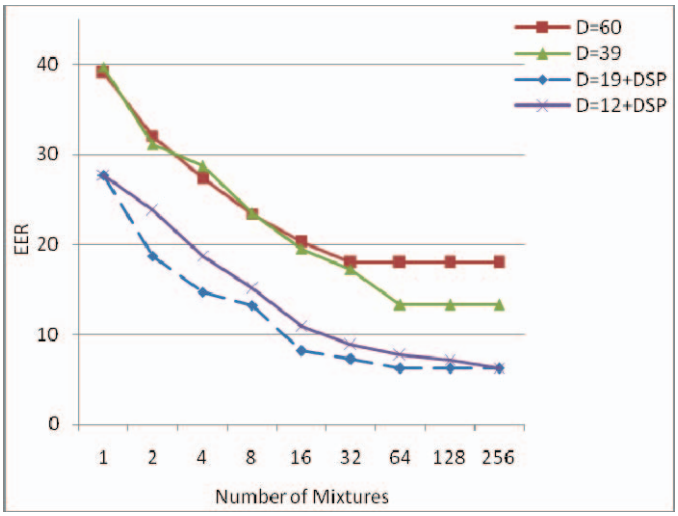

Fig. 3. Speaker EER using different feature vectors for different number of mixtures of GMM.

Throughout this study, verification experiments highlight the feasibility of using power spectrum density of the signal to improve the system's performance. The results obtained with different feature vectors show that the use of MFCC coefficients together with PSD yielding to more significant results. In fact, we succeed to reduce the EER of the system and we obtain 6.2 % of EER with only 19 MFCC together with PSD with 64 mixtures and 6.26 % of EER with 12 MFCC together with PSD with 256 mixtures. We achieve a reduction of nearly 11.7 % with regard to the EER obtained with existing systems which use 60-dimensional feature vectors and a reduction of 7 % with regard to the baseline systems which use 39-dimensional feature vectors.

The DET curves [13] given in the Figure 4 show the results obtained by using the different MFCC feature vectors and MFCC feature vectors combined with PSD coefficients for the verification system. We give comparative results between best verification results obtained with different systems evaluated with 60-MFCC feature vectors, 39-MFCC feature vectors, 19-MFCC coefficients together with PSD and 12-MFCC coefficients together with PSD on speakers from TIMIT database.

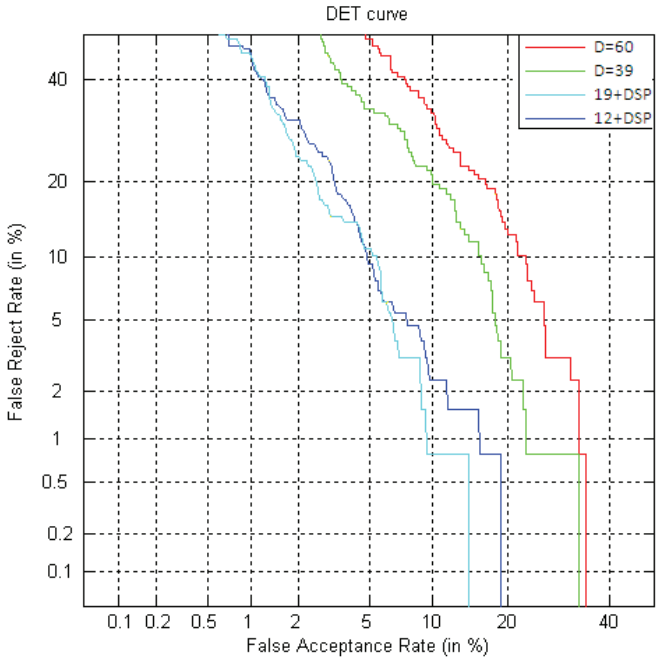

Fig. 4. Detcurves comparison between the different feature vectors.

Our approach provides better results than the results obtained by the baseline speaker verification system. Using the new approach seems to be quite favorable to realistic speaker verification system since it avoids the use of high dimensional feature vectors or the combination of complex algorithms requiring more computational and memory costs.

IV. Conclusions and Perspectives

This paper identified the importance of using power spectrum density of the signal to improve speaker verification. We present a new approach based on reduced MFCC feature vector together with power spectrum density of the signal. Our new approach gives better results than those obtained by the baseline systems with Gaussian mixture models. We substiantialy decrease the error rate and we avoid the use of additional, lengthy and complicated calculations.

References

[1] N. Dehak, Z. Karam, D. Reynolds, R. Dehak, W. Campbell, and J. Glass, “A Channel-Blind System for Speaker Verification”, Proc. ICASSP, pp. 4536-4539, Prague, Czech Republic, May 2011.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, pp. 788-798, May 2011.

[3] R. Togneri and D. Pullella, “An Overview of Speaker Identification: Accuracy and Robustness Issues”, In: IEEE Circuits And Systems Magazine, Vol. 11, No. 2 , pp. 23-61, ISSN : 1531-636X, 2011.

[4] D. A. Reynolds, “A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification”, PhD Thesis. Georgia Institute of Technology, August 1992.

[5] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors”, Speech Communication 52(1): 12-40, 2010.

[6] D. A. Reynolds, “Robust Text-Independent Speaker Identification Using Gaussian Mixture SpeakerModel”, IEEE Transactions on Speech and Audio Processing. vol. 3, n. 1, pp. 72-83, January, 1995.

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

[8] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “Hidden Markov model toolkit (htk) version 3.4 user’s guide”, 2002.

[9] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Commun., vol. 17, no. 1–2, pp.91–108, 1995.

[10] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Process. Lett., vol. 13, no. 5, pp. 308–311, 2006.

[11] D. Reynolds, “Experimental evaluation of features for robust speaker identifi cation,” IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 639–643, 1994.

[12] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J.G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM, "*NIST*, 1993.

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET Curve in Assessment of Detection Task Performance”, in EUROSPEECH, vol. 4, pp. 1895–1898, 1997.