An Improved Approach for Text-Independent Speaker Recognition

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

***Abstract*—This paper presents new Speaker Identification and Speaker Verification systems based on the use of new feature vectors extracted from the speech signal. The proposed structure combine between the most successful Mel Frequency Cepstral Coefficients and new features which are the Short Time Zero Crossing Rate of the signal. A comparison between speaker recognition systems based on Gaussian mixture models using the well known Mel Frequency Cepstral Coefficients and the novel systems based on the use of a combination between both reduced Mel Frequency Cepstral Coefficients features vectors and Short Time Zero Crossing Rate features is given. This comparison proves that the use of the new reduced feature vectors help to improve the system's performance and also help to reduce the time and memory complexity of the system which is required for realistic applications that suffer from computational resource limitation. The experiments were performed on speakers from TIMIT database for different training durations. The suggested systems performances are evaluated against the baseline systems. The increase of the proposed systems performances are well observed for identification experiments and the decrease of Equal Error Rates are also remarkable for verification experiments. Experimental results demonstrate the effectiveness of the new approach which avoids the use of more complex algorithms or the combination of different approaches requiring lengthy calculation.**

Keywords*—GMM; speaker verification; speaker recognition; speaker identification

I. Introduction

The speech signal is an information rich signal that conveys various data to the listener. In addition to the message and words being spoken, the speech conveys much other information such as the language used, the emotion of the speaker, the gender and also the identity of the speaker.

Based on the speech signal, the main goal of automatic speaker recognition is to extract and characterize the information in the speech signal conveying the identity of the speaker.

The broad area of speaker recognition comprehends two fundamental tasks which are speaker identification and speaker verification.

For speaker identification, the system aims at determining who is talking from a set of known voices. The system looks then for the voice which best matches the unknown speaker which makes no identity claim. Speaker identification can be also divided into closed set or open set speaker identification. For the closed set problem, the unknown voice must be among a fixed set of known speakers. However, for the open set speaker identification, the unknown speaker may not exist among the set of known speakers. So, unknown voices are referred to as unknown speakers.

For the task of speaker verification, the system tries to determine whether the unknown person is who he/she claims to be or not. The system makes then a binary decision. Even it accepts the pretended speaker or it reject this unknown speaker.

Depending on the message being spoken by the speaker, the speech used for speaker recognition applications can be either text dependent or text-independent. For text-dependent speaker recognition application, the recognition system has prior knowledge of the text that must be spoken and the system requires that the speaker says exactly the given text. However, for text-independent speaker recognition application, there is no prior knowledge about the text to be spoken, and the speaker is free to say any message he want. Text independent speaker recognition applications are then more difficult but also more flexible.

In this context, this work aims to propose a new approach for Text independent speaker recognition applications based on the use of new information extracted from the speech signal. The proposed system use the Short Time Zero Crossing Rate (STZCR) [14] information with reduced cepstral features to ensure higher performance for the system and also guarantee a reduction of the time and memory complexity of the system. The results are compared to state-of-the-art systems.

This paper is organized as follows: First, related works are summarized. Then, the approach used for the speaker recognition systems is explained. After that, a description of the database used during the experiments is given, followed by a description of the experimental protocol and their results. These results are compared with the state-of-the-art speaker recognition systems. Finally, a conclusion illustrating the main matter of the proposed system for speaker recognition is provided.

II. Related Works

Research and development on speaker recognition field have lead to powerful methods and techniques permitting high performing applications. The most successful approaches are modern statistical approaches [20] where the Gaussian Mixture

Models (GMM) [3] is considered as the most popular approach for current speaker recognition systems [21].

It is also interesting to note that, the process of extracting features from the speech signal is a fundamental process on which the system depends to capture the speaker specific characteristics. For that, many features have been investigated in the literature [2], [11] where the cepstral features [6] are the most appropriate ones for speaker recognition tasks. Up today, the most popular and successful cepstral features are Mel frequency cepstral coefficients (MFCC) [3], [10], [22].

Since speaker recognition application succeed to achieve good performances with appropriate conditions [19], current speaker recognition applications looks for more realistic and challenging conditions. In fact, current speaker recognition systems 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 take into account, which alter the system's effectiveness. The main factors are related to the approach used and the features to be extracted from the speech signal. The systems used in this article have been developed using the well-known state-of-the art approach which is the GMM. Most of the works in this area focus on the use of cepstral coefficients. However, this work focuses on determining whether the Short Time Zero Crossing Rate [14] information is useful for improving current state-of-the-art automatic speaker recognition systems.

III. Proposed Approach

The proposed system for speaker recognition task is shown in Figure 1. The learning phase serves to acquire the characteristics of every speaker from the extracted parameters. A test utterance is input to the system and the recognition task is realized with Gaussian Mixture Models (GMM).

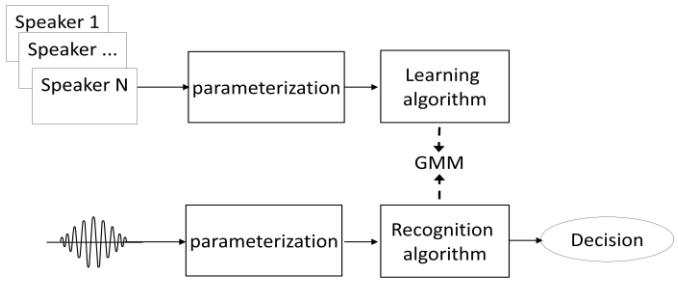

Fig. 1. The architecture of the Automatic Speaker Recognition system

A. GMM approach

The GMM approach can be considered as a model reference for speaker recognition systems [3], [4], [15].

For an utterance of length T frames belonging to a speaker j and *D*-dimensional feature vector extracted for each frame, so for each utterance: $\{\vec{x}_t \in \Re^D : 1 \le t \le T\}$, A Gaussian model for a speaker *j* for any utterance from that same speaker assumes that feature vectors follow a Gaussian distribution, characterized by a mean and a deviation about the mean. Indeed, The Gaussian mixture model for speaker *j*, λ_j is a weighted sum of *M* component densities calculated as follows [3]:

$$p(\vec{x}_t|\lambda_j) = \sum_{i=1}^{M} g_i N(\vec{x}_t; \vec{\mu}_i, \sum_i) \quad (1)$$

Where g_i are mixtures weights having $\sum_{i=1}^{M} g_i = 1$. The individual component densities $N(\vec{x}_t; \vec{\mu}_i, \sum_i)$ represent:

$$N(\vec{x}_t; \vec{\mu}_i, \sum_i) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}}$$

$$\times \exp\left[-\frac{1}{2}(\vec{x}_t - \vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}_t - \vec{\mu}_i)\right] \quad (2)$$

Where $\vec{\mu}_i$ is the mean vector and Σ_i is the covariance matrix.

The GMM model for the speaker *j*, presented by λ_j, is parameterized by the mean vectors, covariance matrices and mixture weights from all *M* component densities:

$$\lambda_j = \{\vec{\mu}_i, \sum_i, g_i\}_j \quad i=1, 2,\ldots, M. \quad (3)$$

a) GMM for speaker identification

In the test phase, an utterance having *T* feature vectors $X = \left\{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\right\}$ is presented to the system. The main objective of an identification system is to find from *N* GMM speaker models, the model of the speaker which has the maximum a posteriori probability [9] for that input feature vector sequence:

$$\hat{j} = \arg\max_{1\le j\le N} P(\lambda_j|X) \quad (4)$$

Where $\hat{j}$ is the identified speaker.

With the use of logarithms and the assumed independence between the observations, the decision can be shown with Maximum-Likelihood (ML) scoring of the log likelihoods:

$$\hat{j} = \arg\max_{1\le j\le N} \sum_{t=1}^{T} \log P(\vec{x}_t|\lambda_j) \quad (5)$$

Where $P(\vec{x}_t|\lambda_j)$ is given above in equation 1.

b) GMM for speaker verification

The speaker verification system need to make a binary decision, even it accepts or rejects the pretence speaker. The verification system uses a likelihood ratio test to an input speech sequence in order to detect if the claimed speaker is true or false. Indeed, for an input vector $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\}$, and a claimed speaker having a model λ_c ,the likelihood ratio is as follows[7]:

$$\frac{p(X \text{ is from the claimed speaker})}{p(X \text{ is not from the claimed speaker})} = \frac{p(\lambda_c|X)}{p(\lambda_{\bar{c}}|X)} \quad (6)$$

With the application of Bayes' rule, the likelihood ratio becomes

$$\Lambda(X) = \log p(X|\lambda_c) - p(X|\lambda_{\bar{c}}) \quad (7)$$

The likelihood ratio between the pretence speaker model and other models (back ground models) is compared to a given threshold θ [9]. The claimed speaker is accepted only if $\Lambda(X) > \theta$.

B. Short Time Zero Crossing Rate (STZCR)

Speech is a signal produced from a time varying vocal tract system with a time varying excitation. That's why, the speech signal is considered as non-stationary in nature. This signal is stationary when it is viewed in blocks of 10-30 msec [16]. Short Time processing divides the input speech signal into short analysis segments that have realatively fixed (non-time varying) properties. These short analysis segments called as analysis frames almost overlap one another.

Zero Crossing Rate is defined as the number of times the zero axes is crossed by the signal per frame. If the number of zero crossings is more in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information which is termed as unvoiced speech. On the other hand, if the number of zero crossing is less, then the signal is changing slowly and accordingly the signal may contain low frequency information which is termed as voiced speech [17]. That's why the Zero Crossing Rate can gives information about the frequency content of the signal, which can be considered as a good indicator about the speaker itself. Short Time Zero Crossing Rate is defined as the weighted average of number of times the speech signal changes sign within the time window [18]. The STZCR for a signal having the window $\omega(n)$ with length *n* is defined as [18]:

$$Z_n = \tfrac{1}{2}\sum_m \left|\text{sgn}[x(m)] - \text{sgn}[x(m-1)]\right| . \omega(n-m) \quad (8)$$

with:

$$\text{sgn}[x(n)] = 1 \text{ if } x(n) \geq 0$$
$$= -1 \text{ if } x(n) < 0$$

IV. Experimental Results

A. TIMIT corpus

In this paper, speaker verification and identification tasks are evaluated with TIMIT (Texas Instruments Massachusetts Institute of Technology) database. The TIMIT corpus has been designed to provide speaker data for the acquisition of acoustic-phonetic knowledge and also for the development and evaluation of automatic speaker recognition systems [12]. TIMIT contains in totality 6300 sentences with 10 sentences 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. Experiments

All evaluations are dealt with 64 speakers selected from all the regions of TIMIT database. Following the protocol suggested in [3], the sentences recorded from each speaker are divided 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) features have been used for extracting features from the speech signal. Since many years, these features proved their success in speaker recognition domain [1], [3], [5]. In this work, MFCC features are used, since they are the most popular choice for any speaker recognition system [3]. The 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]. The features were extracted using the Hidden Markov Model ToolKit (HTK) [8].

Since realistic applications suffer from some constraints like computational resource limitation or reduced memory space, this work looks for an improved approach using more reduced feature vectors and ameliorating the system's performance. In this context, the inclusion of new information extracted from the speech signal which is STZCR of the signal [14] with reduced MFCC feature vectors can improve the system's performance and give significant results. For that, MFCC vector are combined with STZCR features. The new structure of the vectors is evaluated and compared with traditional MFCC vectors.

a) Speaker identification systems

For speaker identification experiments, the number of mixture components is varied from 1 to 256 mixtures and the correct Identification Rates (IR) given with the different feature vectors are plotted in Figure 2.

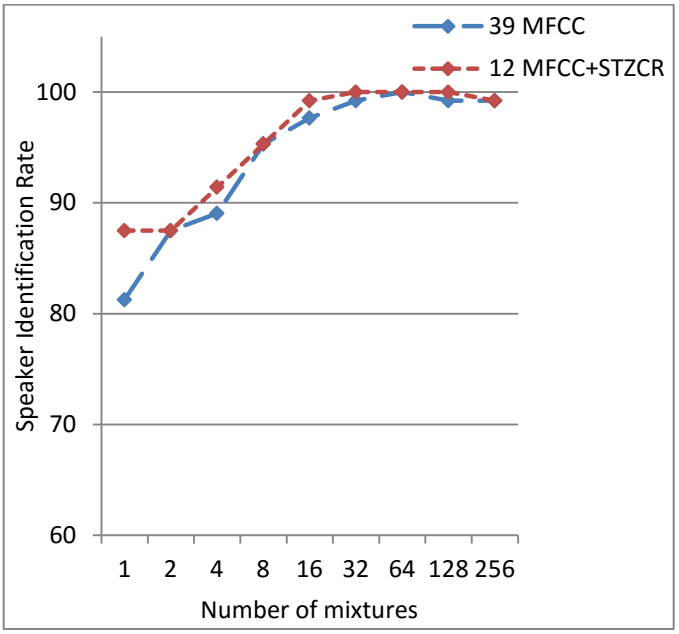

Fig. 2. Speaker IRwith different feature vectors for 8 utterances for the training task and 2 utterances for the test task for various number of mixtures of GMM

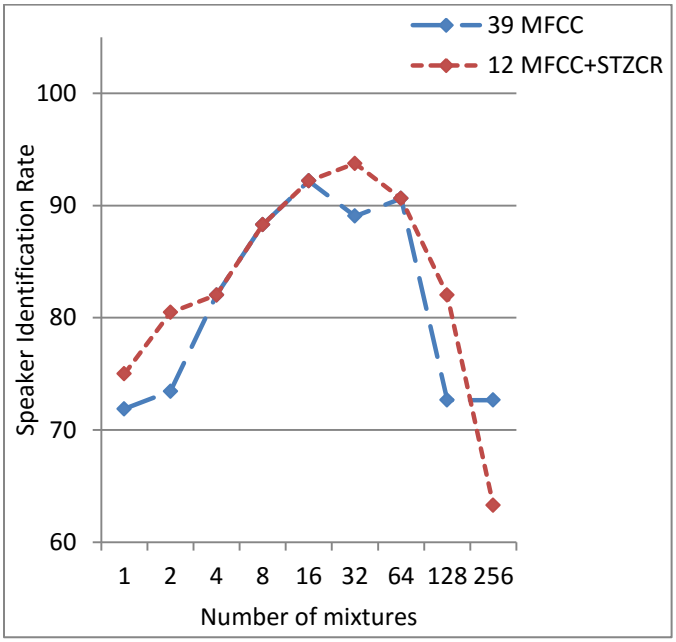

Fig. 3. Speaker IRwith different feature vectors for 3 utterances for the training task and 2 utterances for the test task for various number of mixtures of GMM

In this study, identification experiments are done to demonstrate the feasibility of using STZCR 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 STZCR can give more significant results. In fact, the proposed system succeed to improve the performance of the system and achieved the best result of 100% of correct Identification Rate with the number of 32 mixtures components with only 12 MFCC together with STZCR for 8 utterances for training and 2 utterances for the test task. However, with 39 MFCC feature vectors, the system achieved 100% of correct Identification Rate only with 64 mixtures components with deterioration of the system performance when more components are added.

To further examine the effectiveness of the proposed parameterization, the system's performance is evaluated for more reduced training data. Experimental results were then evaluated for 3 utterances for training and 2 utterances for the test task. The curves given in figure 3 present the results obtained by using 39-dimensional MFCC feature vectors and 12-MFCC coefficients together with STZCR on speakers from TIMIT database.

From the curves presented above, it can be seen that the use of the proposed feature vector composed of MFCC coefficients together with STZCR gives more significant results than the standard MFCC coefficients. In fact, the performance of the system is improved and achieved the best result of 93.75 % of correct Identification Rate with 32 mixtures components with only 12 MFCC together with STZCR for 3 utterances for training and 2 utterances for the test task. However, the system achieved only 92.19 % of correct Identification Rate with 16 mixtures components with 39 MFCC feature vectors. These results explained the superiority of the proposed approach towards the state of the art applications.

b) Speaker verification systems

Speaker verification experiments for the different speakers are dealt for the different feature vectors with 8 utterances for the training task and 2 utterances for the test task. The Equal Error Rates (EER) given with each of the feature vectors are plotted in Figure 4.

Figure 4 shows the results obtained by using the two kinds of feature vectors for 8 utterances for training and 2 utterances for the test task. As can be seen from the graph, the new feature vector succeed to reduce the EER of the system and gives the best result of 2.26% of EER for 128 mixtures components. However, for the state-of-the-art system based on 39-dimensional feature vectors, the best result realized was only 13.28 % of EER for 64 mixtures components.

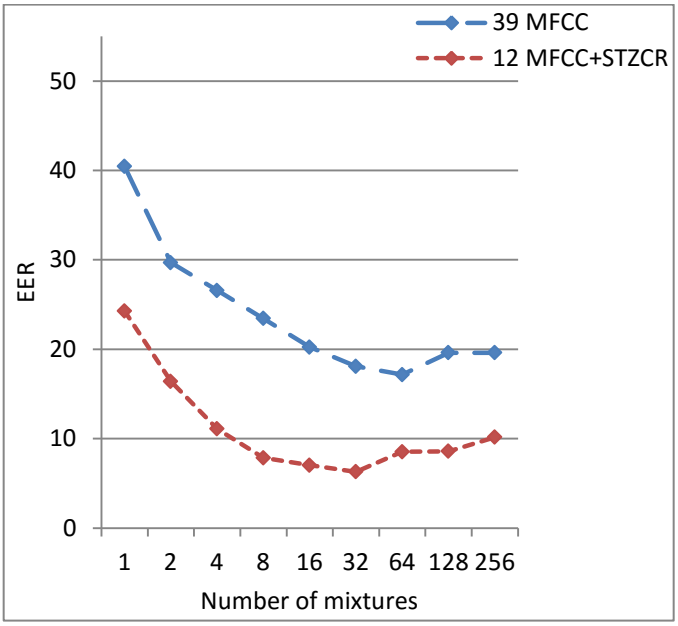

Fig. 4. EERwith different feature vectors for 8 utterances for the training task and 2 utterances for the test task for various number of mixtures of GMM

Comparative results between best verification results obtained with the different verification systems evaluated with 39-MFCC feature vectors and 12-MFCC coefficients together with STZCR are given with The DET curves [13] plotted in Figure 5.

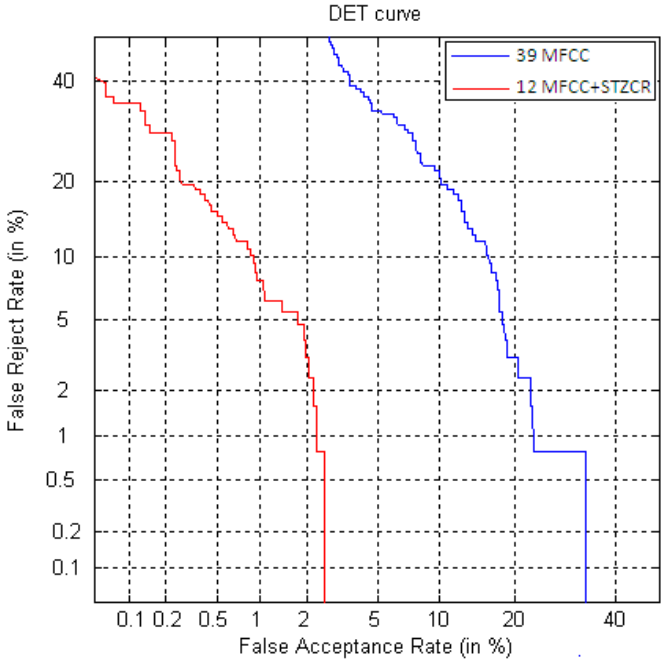

Fig. 5. Detcurves comparison between the different feature vectors with 8 utterances for the training task and 2 utterances for the test task

By examining the results given above with DET curves, it is clear to conclude that the proposed methods yield to more significant results. The use of the combination between reduced MFCC features and STZCR is more appropriate because it avoids the use of high dimensional feature vectors and it gives more efficient system.

In addition to the reduction of the dimension of feature vectors to resolve the problems related to computational resource limitation for realistic applications, the need of limited speakers data is also essential to diminish the system's complexity. That's why, the performance of the system with more reduced training data. The Equal Error Rates (EER) given with each of the feature vectors for 3 utterances for the training task and 2 utterances for the test task are plotted in Figure 6.

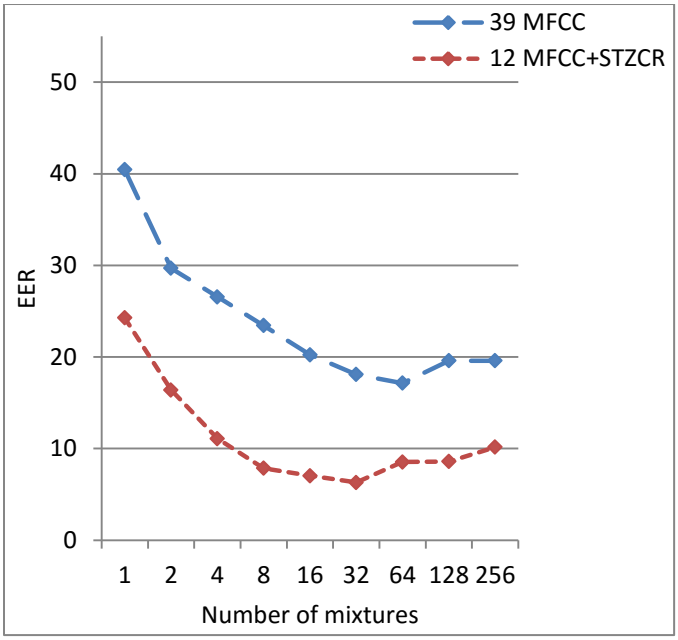

Fig. 6. EERwith different feature vectors for 3 utterances for the training task and 2 utterances for the test task for various number of mixtures of GMM

The results presented above highlight the influence of the amount of training data on the system performance. Indeed, experimental results show that the performance of the GMM-based speaker recognition system decreases when the speech utterance duration becomes shorter.

The results obtained clearly demonstrate that the inclusion of the new information extracted from the speech signal which is the STZCR of the signal with reduced MFCC feature vectors dimension can improve the system's performance and give more significant results. In fact, the proposed system achieved a reduction of nearly 11 % of EER with regard to EER obtained with the state-of-the-art systems based on 39-dimensional feature vectors with a reduced training time. Indeed, the best result achieved with the proposed system achieved 6.30 % of EER with 32 mixtures components. However, the state-of-the-art system achieved the best result of 17.15% of EER for 64 mixtures components.

The DET curves given in Figure 7 present comparative results between best verification EER obtained with verification systems evaluated with the different feature vectors.

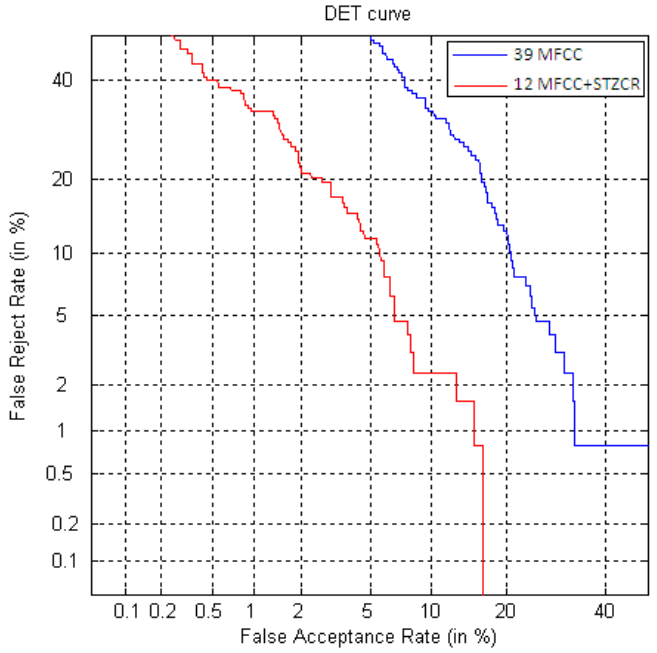

Fig. 7. Detcurves comparison between the different feature vectors with 3 utterances for the training task and 2 utterances for the test task

Throughout this study, it can be seen that the proposed approach gives better results than the results obtained by the state-of-the-art speaker identification and verification systems. The use of the new approach seems to be quite favorable to realistic speaker recognition systems since it avoids the use of high dimensional feature vectors or the combination of complex algorithms requiring more computational and memory costs.

V. Conclusions and Perspectives

This paper identifies the importance of using Short Time Zero Crossing Rate of the signal to improve speaker recognition. It presents a new approach based on low dimensional feature vectors composed by reduced MFCC feature vector together with STZCR of the signal. This new approach gives better results than those obtained by the baseline systems with Gaussian mixture models. The proposed method substantially improves the system performance and avoids the use of additional, lengthy and complicated calculations.

Future work will also investigate the performance of the proposed system with other features or applications.

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 R. Mammone, X. Zhang, and R. Ramachandran, “Robust speaker recognition: A feature-based approach,” IEEE Signal Processing Mag., vol. 13, no. 5, pp. 58–71, 1996.

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

[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] Ramaligeswararao, N. M., V. Sailaja, and K. Srinivasa Rao. "Text Independent Speaker Identification using Integrated Independent Component Analysis with Generalized Gaussian Mixture Model." THE SCIENCE AND INFORMATION ORGANIZATION (2011): 85.

[11] Kekre, H. B., and Vaishali Kulkarni. "Speaker Identification using Frequency Dsitribution in the Transform Domain." www. thesai. org| info@ thesai. org (2012).

[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.

[14] L.R. Rabiner, and R.W. Schafer, Digital speech processing. The Froehlich/Kent Encyclopedia of Telecommunications, 6, pp.237-258, 2011.

[15] C. Barras & J. Gauvain, “Feature and score normalisation for speaker verification of cellular data”, International Conference on Acoustics, Speech, and Signal Processing(ICASSP), in Hong Kong SAR, China, April 6-10, 2003.

[16] Ronald W.Schafer and Lawrence R.Rabiner, “Digital Representations of Speech Signals”, Proceedings of the IEEE, Vol.63, No.4, April 1975.

[17] R.G.Bachu, S.Kopparthi, B.Adapa and B.D.Barkana, “Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy”, Advanced Techniques in Computing Sciences and Software Engineering, pp.279-282, 2010.

[18] Lawrence R.Rabiner and Ronald W.Schafer, “Introduction to Digital Speech Processing”, Foundations and Trends in Signal Processing, Vol.1, No.33-53, 2007.

[19] Reynolds, D. An overview of automatic speaker recognition Technology. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075). (2002).

[20] N. Fatima and T. F. Zheng, “Short Utterance Speaker Recognition A research Agenda”, In International Conference on Systems and Informatics (ICSAI), 2012.

[21] P. Motlicek, S. Dey, S. Madikeri, and L. Burget, "Employment of Subspace Gaussian Mixture Models in speaker recognition." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[22] Almaadeed, Noor, Amar Aggoun, and Abbes Amira. "Text-Independent Speaker Identification Using Vowel Formants." *Journal of Signal Processing Systems* 82.3: 345-356, 2016.