International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1528-1535

A Novel Approach for Efficient Speaker Identification System in Mismatch Conditions

El bachir Tazi [1, 2] and Noureddine El makhfi [2]

[1] Assistant Professor at Moulay Ismail University Meknes
[2] Member of Research Team: Physics, Computer Science and Process Modeling
Higher School of Technology Khenifra, Morocco
e.tazi@estk.umi.ac.ma [1] n.elmakhfi@gmail.com [2]
([1] corresponding author)

Abstract

Although the field of automatic speaker recognition has been the subject of extensive research over the past decades, the lack of robustness against background noise has remained a major challenge. In this paper we study the use of a novel approach combining the conventional MFCC method with the robust GFCC one for features extraction. We adopted this approach because the MFCC coefficients are the most widely used in speaker identification task, but it deteriorates in the presence of noise. On the other hand the GFCC features have shown very good robustness against noise and acoustic changes. Our main objective consists to enhance the accuracy and efficiency of a speaker identification system in presence of varying noise. Given the fact that the performances of speaker identification systems are strongly influenced by the quality and quantity of the used speech signal, we first deployed a Voice Activity Detection VAD, which removes the silence zones and reduces background noise in the speech signal. Given that most speech information is represented by the static parameters, we have also proposed the use of a reduced number of dynamic parameters. The mismatch between the training conditions and the testing ones has a deep impact on the accuracy of these systems and represents a barrier for their operation in real conditions generally affected by noises disturbances. To simulate these conditions, we have mixed the white Gaussian noise with various SNR levels to the test utterances. For classification and identification, we have used the well-known Gaussian Mixture Models GMM method. The performances evaluation of our approach is carried out on our proper corpora database containing 50 Arabic speakers. The experiments show significant improvements of the accuracy and time response performance. The results provide an analytical comparison between MFCC, GFCC and MFCC-GFCC combined features.

Keywords: Combination features strategy, Gammatone frequency coefficients, Mel frequency cepstral coefficients, Speaker identification system, Voice activity detection.

\

INTRODUCTION

Automatic Speaker Identification is the task performed by the machine to identify a person from his/her voice. The structure of the system achieving this task contains typically three stages: features extractor, pattern classifier and decision logic [1,2]. The features extractors are short-term cepstral coefficients such as Mel Frequency Cepstral Coefficients, Gammatone Frequency Cepstral Coefficients and perceptual linear Predictive coefficients, or long-term features such as prosody [3,4]. For pattern recognition, GMM are widely used to model the feature distributions and it considered actually as the state of the art, in text independent speaker identification task [5]. Such systems usually do not perform well under noisy conditions [6,7,8] because the extracted features are distorted by noise, causing mismatched likelihood calculation. A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as voice activity detection, pre-emphasis, framing, windowing, spectral, cepstral analysis and identification of the spoken utterance. The following figure 1 describes the general structure of our system based on the combination of MFCC and GFCC front-ends combined to the VAD enhancement signal method.

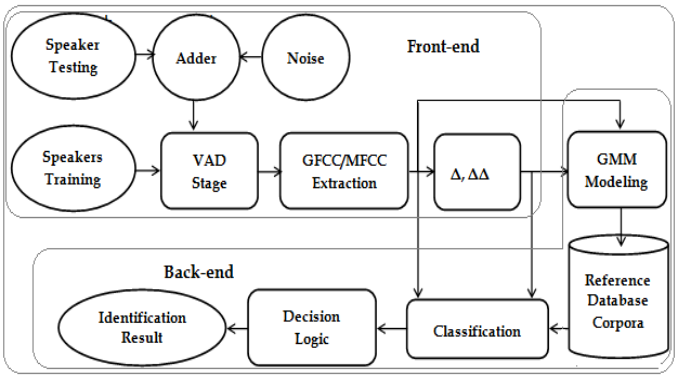

Figure 1: General structure of the proposed system

This type of system operates on two distinguished phases:

- The first phase is training sessions. This step consist to build a reference database corpora that will serves as reference for comparing and identifying the speaker in the next step
- The second phase is a testing phase that consists of searching the identity of the speaker under test.

In previous studies, we have shown the accuracy of MFCC [9] under low noise conditions and the robustness of GFCC [10] in noisy environments. It would be beneficial to integrate the advantages of these two approaches, to reduce or eliminate their individual drawbacks.

VAD SPEECH ENHANCEMENT

The VAD technique is frequently used in a number of applications including speech coding, speech enhancement, speech identification and speaker identification. This technique consists of extracting only the parts containing the useful speech signal by removing the parts corresponding to a silence and background noise. This will reduce the duration of recordings to their useful parts only. Hence there improved speed and performance of the SIS systems. Several implementations are reported in the literature to design a VAD module [11,12,13,14]. In this study we have choose the solution using the Zero Crossing Rate (ZCR) combined to the energy of the speech signal. Indeed a low rate of zero crossing and high energy are a good indicator of the presence of a speech signal, while a high rate of zero crossing rate and a low energy characterize a silence zone containing only background noise [15]. Given the fact that the noise is characterized by its random nature, and then usually it has a zero-crossing rate higher than the parts corresponding to a speech signal. In this implementation we have used the equation (1) to compute the zero crossing rate.

$$ZCR = 0.5 * \sum_{n=0}^{N-1} |sign(s_n) - sign(s_{n-1})| \qquad (1)$$

Where sign(s_n) is the sign of the instantaneous sample value of signal s(n) acquired at time n and N is the total length of the processing speech signal. In practice to discriminate between the presence and absence of the speech signal we have fixed two thresholds one for the energy and one other for the ZCR. Bellow here are the main steps of the proposed algorithm:

- Step 0: initialize all parameters like thresholds of energy and ZCR (thr_zcr, thr_energy), length of frame (lengthf) etc.
- Step 1: for i=1 to length of noisy speech signal to process
- Step 2: framing the speech signal using the initialized lengthf
- Step 3: for j=1 to length of frames do
- Step 4: calculate the energy and the ZCR of the j^{th} frame
- Step 5: if ZCR>thr_zcr and energy <t hr_energy
- Step 6: suppression the j^{th} frame from original speech signal next j
- Step 7: Improved VAD speech signal ← speech resulting in step 6 next i

The figure 2 shows an example of the resulting signal after VAD processing applied to an utterance of speech signal corrupted by 20dB SNR white Gaussian noise. We can show that the length of the resulting signal is short than the original one. This indicates that some parts of the original signal were suppressed by the VAD algorithm. These parts correspond to silence/non speech segments of the original signal and background noise. This action reducing the signal will subsequently contribute to accelerate the speaker identification process, requires less storage and finally increase the efficiency of the system.

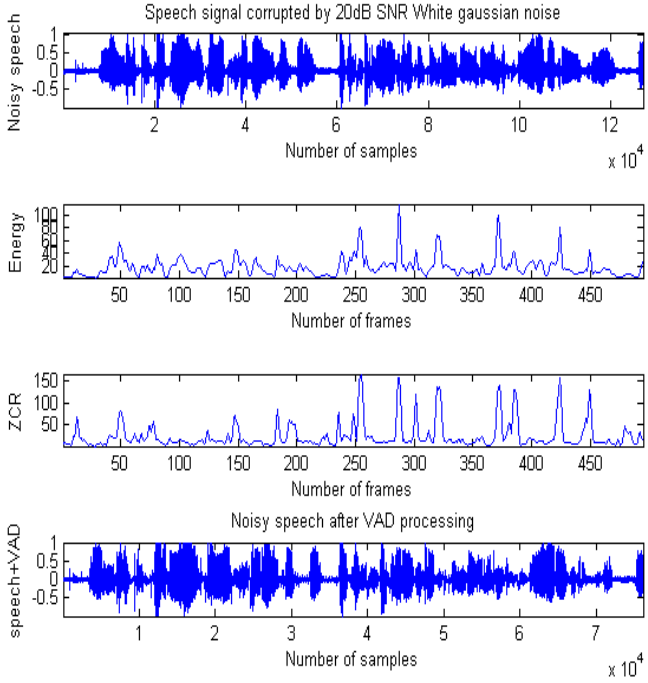

Figure 2: Example of the output signal obtained with the VAD method.

FEATURES EXTRACTION TECHNIQUES

The quality and quantity of the features influence highly the performance of any speaker identification system. For this, it is recommended to give a lot of importance to the parametric extraction phase. This by choosing the best parameters but also the best approaches of combination between these parameters depending possibly also on the type of application envisaged. The main purpose of feature extraction is to compute a set of acoustic vectors to provide a compact representation of the speech signal. In addition, it removes unwanted and redundant information and retains only useful information. Unfortunately, in practice, this process is complicated, especially, in the presence of different types of noise that mix with the useful information of the speech signal and finally generate confusion during the last speaker recognition and decision phase. Based on the fact that static parameters are more representative of information carried on a voice signal than their derivatives. Similarly, the first derivative parameters are more representative than the second derivatives, we selected the top 24 ranked coefficients in proportion of static=12, delta=8, and delta-delta=4. In this

study, we noticed that this reduced selection of parameters gives better performances, especially in terms of time response of the system, than the 36 original coefficients (12x3) that we studied previously in [16].

Mel Frequency Cepstral Coefficients (MFCC)

MFCC is based on a perceptually scaled frequency axis. The Mel-scale provides higher frequency resolution on the lower frequencies and lower frequency resolutions on higher frequencies according to equation (2) giving the relation between frequency in Hz and Mel scales

$$mel = 2595.log_{10}\left(1 + \frac{f(Hz)}{700}\right) \quad (2)$$

The figure 3 bellow shows the shape of the used 20 Mel filter-bank with 16 kHz sampling frequency.

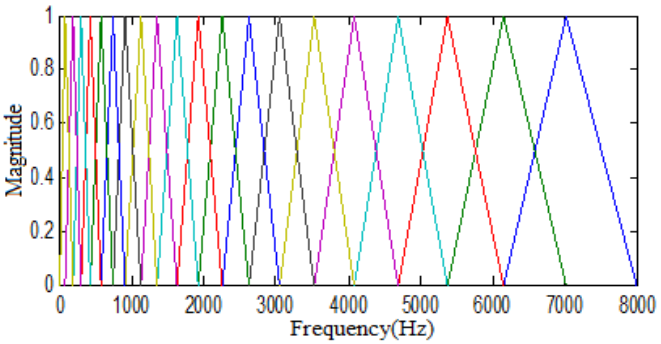

Figure 3: Shape of 20 Mel filter-bank

This scaling is based on hearing system of human ear. The principle of the MFCC method is illustrated by the following block diagram at figure 4.

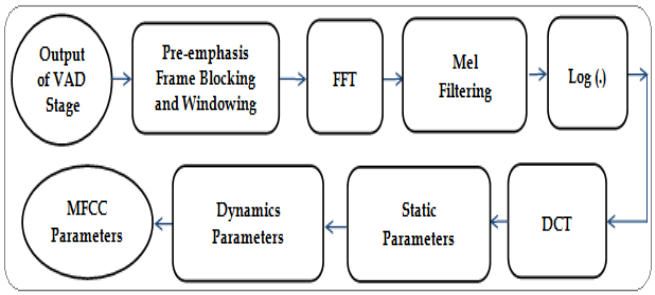

Figure 4: Block diagram of MFCC method

Nowadays, the conventional MFCC algorithm is widely used for speech signal parameterization and is accepted as the baseline. It gives better performance in a fairly quiet environment but it is not robust enough in noisy environments.

The figure 5 and 6 show respectively an example of the evolution of the logarithmic energy at the outputs of the used 20 Mel filter-bank and the magnitude evolution of the 12 first static MFCC parameters.

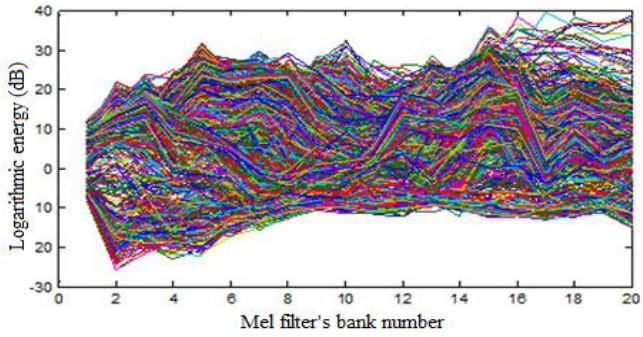

Figure 5: Logarithmic energy at the output of the Mel filter-bank

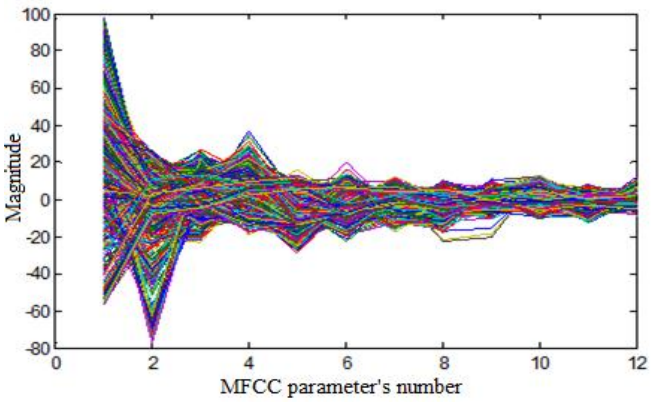

Figure 6: Magnitude evolution of 12 static MFCC parameters

According to this last figure, we can be said that the amplitudes of the first MFCC parameters are higher compared to the last ones. This means that the first MFCC parameters can be sufficient to calculate most of the speech signal information that is supposed to discriminate between different speakers.

Gammatone Frequency Coefficients (GFCC)

The overall process of the GFCC algorithm is shown in the block diagram at the following figure 7.

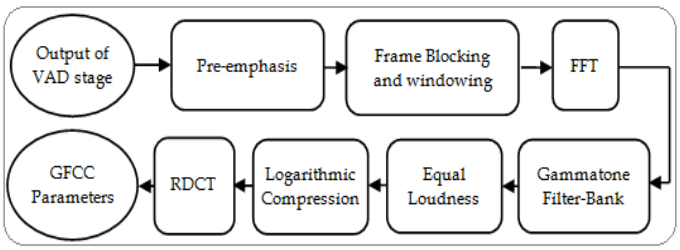

Figure 7: Block Diagram of the GFCC algorithm

This novel and robust algorithm based on the GammaTone Filter Bank (GTFB), has been successfully applied to speech recognition [17]. The filters in the bank are designed to simulate the auditory process of human ear [18,19,20] that are formulated as follows:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \varphi) \quad (3)$$

a is a constant which is generally equal to 1.

n is the filter order which is set less or equal 4

φ is the phase shift between filters

Fc and b are respectively the center frequency and the bandwidth of the filter in Hz which is related by:

$$b = 1.019 * ERB = 1.019 * 24.7\left(4.37 * \frac{f_c}{1000} + 1\right) \quad (4)$$

The following figure 8 shows the shape of the used gammatone filter-bank with 16 KHz sampling frequency.

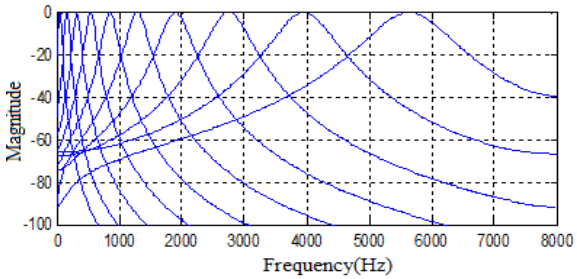

Figure 8: Impulse response of a set of 10 Gammatone filters

As in the conventional MFCC the robust GFCC technique are calculated from the spectrum of a series of windowed speech frames of 32ms and overlapping by 16ms. First, the spectrum of a speech frame is obtained by applying the Fast Fourier Transformation (FFT), 512 point. Then the speech spectrum is passed through 20 gammatone filter-bank GTFB. Equal-loudness is applied to each of the filter output, according to the centre frequency of the filter. After that, the logarithm is taken to each of the filter outputs. Finally we applied the Reverse Discrete Cosine Transform (RDCT) to the gammatone filter-Bank outputs in order to transit from spectral domain to cepstral domain. The figure 9 bellow shows an example of the gammatonegram of 10 GFCC filter-bank.

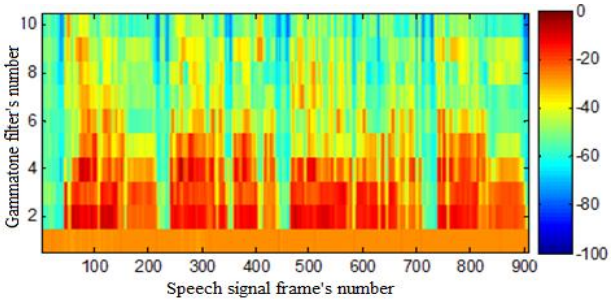

Figure 9: Gammatonegram of 10 GFCC filer-bank

According to this gammatonegram, we can be said that most of the energy of the speech signal is concentrated only at the outputs of the first gammatone filters. The next figure 10 shows the magnitude evolution of 12 static GFCC parameters. The following equations (5) and (6) give the expressions used to compute these parameters.

$$GFCC_k = \sqrt{\frac{2}{M}}\sum_{m=1}^{M}\left\{X_{m(ln+e)}\cos\left(\frac{\pi k(m-0.5)}{M}\right)\right\} \quad (5)$$

Where $1 \leq k \leq 11$

$$GFCC_0 = \sqrt{\frac{1}{M}}\sum_{m=1}^{M} X_{m(ln+e)} \quad (6)$$

m is the number of the filter in the bank

k is the number of the GFCC parameter

Xm is the output of the gammatone filter-bank after equal loudness and logarithmic compression processing.

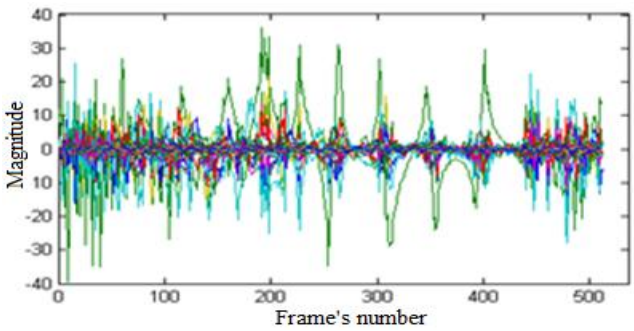

Figure 10: Magnitude evolution of the static GFCC parameters

MFCC and GFCC Combination Strategy

To benefit from the above mentioned advantages of each of the MFCC and GFCC methods, we grouped them in the same features extractor according to the following figure 11. We have therefore concatenated the acoustic vectors resulting respectively from these two methods in order to verify if the approach can makes an improvement performance in term of accuracy, assuming that both types of parameters may contain additional and complementary features. In details, we used 24 static parameters (12 MFCC + 12GFCC) and only 24 derived parameters (8ΔMFCC + 4ΔΔMFCC + 8ΔGFCC + 4ΔΔGFCC). We mention that this reduced selection of parameters contributes to increase the speed of the system.

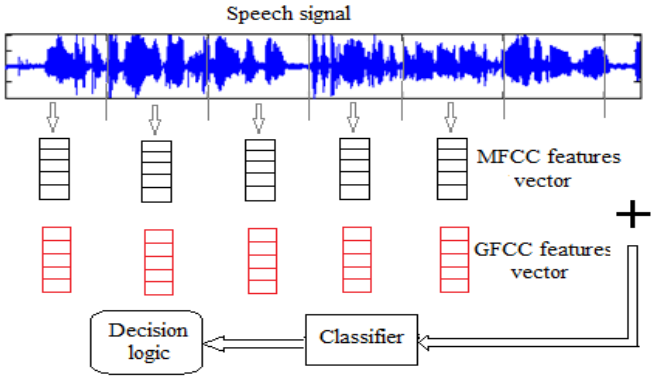

Figure 11: Principle of the proposed combination strategy

SPEAKER RECOGNITION TECHNIQUE

The classifier used in this study, as a back-end of our system, is the well-known Gaussian Mixture Model (GMM) that is very suitable for text-independent speaker identification task [5,21]. Let i be the number corresponding to one speaker in the database, xi represents a signal belonging to the speaker i and Xxi represents the model of speaker i resulting of the signal xi. We also note £(xi/ Xxj) the likelihood of xi knowing the model Xxj.

For a yt vector of d dimension, the multi-dimensional Gaussian distribution denoted N(μ,Σ) has a probability density function $\mathcal{F}$μ,Σ(yt) given by (7).

$$f_i(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\Sigma i}} \exp\left[-\frac{1}{2}(x-\mu i)'\sum_i^{-1}(x-\mu i)\right] \quad (7)$$

Where μ and Σ are respectively the average vector of d dimension and the covariance matrix of dxd dimension of the distribution. The function £(yt/μ,Σ)=$\mathcal{F}$μ,Σ(yt) is called the likelihood function of the distribution. The Xxi models used are the GMM. Each GMM X is a weighted sum of multivariate Gaussians (4) defined by the vector of parameters Θx=(c1, ... ck, μ1, ... μk, Σ1, ..., Σk).

Where k is the number of Gaussian components and ck the weight of the mixture associated with the k^{th} component given that:

$$ck>=0 \text{ and } \sum_{l=1}^{K} ci = 1 \quad (8)$$

The likelihood for a test vector yt is produced by the mixture of Gaussian GMM X is expressed by (9)

$$£(yt/X) = £(yt/\Theta x) = \sum_{i=1}^{K} ci\, £(yt/\mu i, \Sigma i) \quad (9)$$

For a speech signal y containing n samples y=(y1, y2, y3, ..yn), the likelihood of this signal knowing the GMM X model is given by (10)

$$£(y/X) = \prod_{i=1}^{N} £(y_i/X) \quad (10)$$

Where y_i is the i^{th} sample of y signal.

The Learning phase aims to estimate the parameters of Gaussian distributions that make up the models corresponding to all acoustics vectors in the database. These parameters are obtained by the K-means algorithm, and then the optimization of the values of these parameters is provided by the Expectation Maximization algorithm (EM) described in [23].

EXPERIMENTATION AND PERFORMANCE MEASURE

Experimental conditions

In this study, we have interested to evaluate the benefit of the VAD method. For this we have used it in order to improve the GFCC front-end extraction in a text-independent monaural speaker identification context. First we have built our proper corpus database which corresponding to a population of 50 Arabic-speakers (27 male and 23 female). Each speaker had participated by 2 different recordings: one for learning the database for about 20s and one other for the test step for about 10s. All the productions sound from the speakers, were directly digitized to wav format with a sampling frequency of 16 kHz and 16-bit monophonic quantification using the well-known software Wavesurfer®8 [24]. A white Gaussian noise, with 0 mean and unit variance, of variable level was added to the recorded signals to examine the robustness of described techniques in noisy environments that are inevitable in most real applications. The features extractors that will be considered in this set of experiments are MFCC, GFCC, and MFCC+GFCC without and with the VAD stage. The entire system is implemented under MATLAB®10 programming environment. The following table 1 describes in details the experimental conditions.

Table 1: Experimental Conditions of the Speaker Identification Systems

Task system	Text-independent automatic speaker identification
Language	Arabic
Front-ends	MFCC, GFCC, MFCC+GFCC
Back-end	Gaussian mixture models (GMM) with 4 mixture
Silence suppression and noise reduction	Voice activity detection (VAD)
Number of coefficients in a feature vector	24 (12 static + 8 delta + 4 delta-delta) for GFCC and MFCC 48 (24x2) for GFCC+ MFCC
Window size	32 ms
Step size	16 ms
Sampling rate	16kHz
Training set	50 speakers (one utterance per speaker for about 20s)
Test set	50 speakers (one utterance per speaker for about 10s)
Noise Type	White Gaussian Noise (WGN) with 0 mean and unit variance
SNR range	40 to 0 dB with 5dB Step
Platform	HP Elite book core i7 - 2.7 Ghz
Programming Language	MATLAB®10

EXPERIMENTAL RESULTS

The performance evaluation of speaker identification systems is measurable in terms of their accuracy and speed. Accuracy can be measured in terms of Identification Rate (IR in percent) that is computed by equation (11)

$$IR = \frac{Positive\ tests\ number}{Total\ tests\ number}\% \qquad (11)$$

The following table II shows the obtained identification accuracy respectively with MFCC, GFCC and the combination of these last features extraction methods without and with VAD technique for different levels of signal per noise ratio SNR. From the results obtained, it can be said that the GFCC method combined with the MFCC method gives the best score in terms of accuracy. However, the high number of parameters used (48 parameters) makes the identification phase slower and generates more memory usage. This can be an obstacle for embedded systems and in real time applications. The experimental tests show also that:

- The VAD stage improuves the accuracy for both the conventional MFCC and the robust GFCC front-ends. The dynamic variants parameters of MFCC and GFCC give minors improvement accuracy than the only static variant but they occur a long time to estimate the parameters of the GMM models and leads to high memory consumption.
- The high values of the Gaussian mixture models give substantially better accuracy performance with the MFCC parameters, but they take a long time to estimate the GMM parameters and obviously require a lot of memory.
- According to a previous study [x], the optimal value of NG=4 is sufficient for obtaining the good results in terms of accuracy and efficiency, especially with the GFCC parameters.
- The GFCC method gives better accuracy and efficiency than conventional MFCC method, especially in noisy conditions. But in quiet environments MFCC parameters give the best score of accuracy performance.
- The use of reduced number of dynamic parameters is an optimal solution that gives a good accuracy with fast time response of the system.

Figure 12 below shows the performance results obtained with the different methods implemented according to the data on table 2.

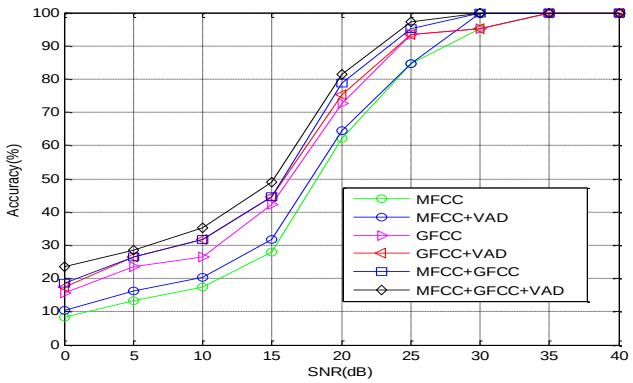

Figure 12: Accuracy performance of studied methods

Table 2: The accuracy of different studied Methods versus various SNR Level

Metod / Noise level	Conventional MFCC (Ncoef=24, NG=4)		Robust GFCC (Ncoef=24, NG=4)		MFCC + GFCC (Ncoef=48, NG=4)	
SNR (dB)	**12MFCC+8Δ+4ΔΔ without VAD**	**12MFCC+8Δ+4ΔΔ with VAD**	**12GFCC+8Δ+4ΔΔ without VAD**	**12GFCC+8Δ+4ΔΔ with VAD**	**12MFCC+8Δ+4ΔΔ+ 12GFCC+8Δ+4ΔΔ without VAD**	**12MFCC+8Δ+4ΔΔ+12GFCC+8Δ+4ΔΔ with VAD**
0	08.45%	10.38%	15.63%	17.52%	18.43%	23.45%
5	13.28%	16.14%	23.45%	26.37%	26.37%	28.52%
10	17.52%	20.32%	26.37%	31.63%	31.63%	35.19%
15	27.86%	31.63%	42.15%	44.67%	44.67%	48.85%
20	62.28%	64.36%	72.57%	75.42%	78.72%	81.54%
25	84.67%	84.67%	93.56%	93.56%	95.23%	97.12%
30	95.23%	100%	95.23%	95.23%	100%	100%
35	100%	100%	100%	100%	100%	100%
40	100%	100%	100%	100%	100%	100%
Average Accuracy	**56.58%**	**58.61%**	**63.21%**	**64.93%**	**66.11%**	**68.29%**

CONCLUSION

The performance of speaker identification systems has improved due to recent advances in speech processing techniques but there is still need of improvement. In addition there are, to our knowledge, no single features capable to give acceptable performance in adverse conditions. In this paper a combined approach strategy for feature extraction has been studied and compared with MFCC and GFCC features extractors. The objective of this study was to demonstrate to what extent speaker identification systems can benefit from the combination of these kinds of features. A voice activity detection technique was also implemented in order to enhance the performance of our system in terms of accuracy and efficiency against noise. In order to find out the performance of the system based on the proposed features combination, we have used our proper corpora database containing 50 Arabic speakers corrupted by additive white Gaussian noise. The average improvement of 11,71% relative to the baseline MFCC front-end is achieved with about 2% improvement is due only to the VAD module when SNR changes from 40dB to 0db. In addition, the experiments show that the GFCC front-end combined to the VAD method gives considerable speed and accuracy improvement. Despite the improvement due to the reduced number of parameters used, we found that the system based on the proposed combination strategy is still slow to be operational in practical applications. In the future, we will focus on improving efficiency by finding better ways to combine features. Most likely, we will explore a hybrid approach based on the projection in a reduced parameter space to improve the speed of the system.

REFERENCES

[1] J.P. Campbell, “Speaker identification: A tutorial” in Proc. IEEE, vol. 85, pp. 1437-1462, 1997.

[2] S. Furui, “Digital speech processing, synthesis, and identification” in New York: Marcel Dekker, 2001.

[3] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, “Using Prosodic and Lexical Information for Speaker Identification,” ICASSP 2002

[4] Markel, J.D., et al, "Long term feature averaging for speaker recognition". In IEEE trans. Acoust. Speech signal processing, Vol. ASSP-25, n°4, pp 330-337, 1977

[5] D.A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models” in Speech Comm., vol. 17, pp. 91108, 1995.

[6] Y. Shao and D.L. Wang, “Robust speaker identification using binary time-frequency masks” in Proc. ICASSP, vol. I, pp. 645-648, 2006.

[7] M. Zamalloa, L. Rodríguez, M. Penagarikano, G. Bordel and J. Uribe, "Improving robustness in open set speaker identification," in Odyssey 2008: The Speaker and Language Recognition Workshop, Stellenbosch, South Africa, 2008.

[8] D. A. Reynolds, "An overview of automatic speaker recognition technology," in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002

[9] El Bachir Tazi A. Benabbou and M. Harti, "Design of an Automatic Speaker Recognition System Based on Adapted MFCC and GMM Methods for Arabic Speech", International Journal of Computer Science and Network Security (IJCSNS), pp. 45-50, Vol.10, No.1, January 2010

[10] El Bachir Tazi, Abderrahim Benabbou and Mostafa Harti, "Robust Features For Noisy Text-independent Speaker Identification Using GFCC Algorithm Combined to VAD and CMN Techniques", Journal of Theoretical and Applied Information Technology (JATIT), pp. 206-216, Vol.36, No.2, February 2012

[11] Sohn, J., Sung, W., 1998. “A voice activity detector employing soft decision based noise spectrum adaptation” in: Internat.Conf. on Acoust. Speech Signal Process., Vol. 1, pp. 365–368

[12] J. A. Haigh and J. S. Mason, “Robust voice activity detection using cepstral features,” in IEEE TEN-CON, 1993, pp. 321–324

[13] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, “The voice activity detector for the pan European digital cellular mobile telephone service,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing, May 1989, pp. 369–372.

[14] M. H. Moattar and M. M. Homayounpour “A simple but efficient real-time voice activity detection algorithm” 17th European Signal Processing Conference (EUSIPCO 2009) Glasgow, Scotland, August 24-28, 2009

[15] W. Abdulla, “Auditory based feature vectors for speech recognition systems advances” in Communications and Software Technologies, N. E. Mastorakis & V. V. Kluev, Editor. WSEAS Press. pp 231-236, 2002.

[16] El Bachir Tazi "A Robust Speaker Identification System Based On the Combination of GFCC and MFCC Methods” in the 5th International Conference on Multimedia Computing and Systems (ICMCS'16), IEEE Conference, organized from 29 September to 1 October 2016, Marrakech, Morocco.

[17] M. Kleinschmidt, J. Tchorz and B. Kollmeier, “Combining speech enhancement and auditory feature extraction for robust speech recognition” in Speech Communication, Vol. 34, Issues 1-2, pp. 75-91, 2001.

[18] Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M., "Complex sounds and auditory images, Auditory Physiology and Perception", (Eds.) Y Cazals, L. Demany, K.Horner, Pergamon, Oxford, pp. 429-446, 1992.

[19] M. Slaney, "An Efficient Implementation of the

Patterson-Holdsworth Auditory Filter Bank", Apple Technical Report No. 35, Advanced Technology Group, Apple Computer, Inc., Cupertino, CA, 1993

[20] Glasberg, B. and Moore, B., "Derivation of auditory filter shapes from notched-noise data", Hearing Research, Vol. 47, pp. 103-108, 1990.

[21] Douglas A. Reynolds et Richard C. Rose; " Robust text-independent speaker identification using gaussian mixture speaker models" in IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 3, N° 1 pp: 72-83, january 1995.

[22] Reynolds, Douglas A. Thomas F. Quatieri, and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models" in Digital Signal Processing. vol. 10, pp. 19-41, 2000.

[23] Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum Likelihood from Incomplete Data via the EM Algorithm" in Journal of the Royal Statistical Society, B, 39, 1–38. December 1976.

[24] http://www.speech.kth.se/wavesurfer/