Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ROBUST FEATURES FOR NOISY TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING GFCC ALGORITHM COMBINED TO VAD AND CMNTECHNIQUES

[1]E. B. TAZI, [2]A.BENABBOU and [1]M.HARTI

[1]UFR INTIC Département d'InformatiqueFaculté des sciences DharMehraz Fès Maroc

[2] Département d'Informatique Faculté des sciences et techniques SaissFès Maroc

E-mail: elbachirtazi@yahoo.fr, abenabbou@yahoo.fr, mharti@rocketmail.com

ABSTRACT

A major problem of most speaker identification systems is their unsatisfactory robustness in noisy environments. The performance of automatic speaker identification systems degrade drastically in the presence of noise and other distortions, especially when there is a noise level mismatch between the training and testing environments. In this experimental research we have studied a recently robust front-end algorithm based on Gammatone Frequency Cepstral Coefficients GFCC associated to Voice Activity DetectorVAD and Cepstral Mean Normalization CMN techniques. Our system using a Gaussian Mixture Models GMM classifier are implemented under MATLAB®7 programming environment. An Expectation Maximization EM algorithm was used to maximize the sum of Gaussian densitiesuntil convergence was reached.Evaluation is carried out on our owndatabase containing 51 mixed Arabic speakers. All test utterances are corrupted by a multilevel White Gaussian Noise WGN.Our aim is to study the performances of the suggested architecture and make a comparison with the conventional Mel Frequency Cepstral Coefficients MFCC method which we have successfully implemented and tested in the previous work. The obtained experimental results confirm the superior performance of the proposed method over MFCC and outperform it in different noisy environments. Theevaluationresultsbased on the recognition rate accuracy show that both MFCC and the proposed features extractor have perfects performances in low-noise environments when Signal per Noise Ratio SNR is greater than 35 dB (practically 100% in all cases). But when the SNR of test signal changed from 0 to 40 dB, the average accuracy of the MFCCs methods is only 52.14%, while the proposed GFCCsfeatures extractors associated to VAD and CMN techniques still achieves an average accuracy of 57.22%.

Keywords:*Cepstral Mean Normalisation(CMN); Gammatone Frequency Cepstral Coefficients (GFCC); Gaussian Mixture Models (GMMs); Mel Frequency Cepstral Coefficients (MFCC); Robust speaker identification;Voice Activity Detector (VAD); White Gaussian Noise (WGN).*

1. INTRODUCTION

Speaker identification assigns an identity to a test utterance from a set of known speakers. Normally the utterance is assumed to be from a known set of speakers [1,2] and is therefore called closed-set speaker identification but open-set speaker identification can also reject a speaker if the best speaker score does not exceed a certain threshold. Speaker identification is normally carried out by training a speaker model for every speaker in the set and the test is implemented by pattern matching, as shown in figure 1. The preprocessing and feature extraction process treats the speech signal waveform
and represents it as feature vectors. These feature vectors are modeled in the training part of the system by using Gaussian mixturemodels GMMs.

The identification process performs pattern matching of the derived feature vectors with each speaker model to give Maximum Log Likelihood MLL and the decision module selects the most likely speaker. In general, a typical speaker identification systemcan be divided into two major parts:front-end and back-end.The front-

Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

end of the system is principally a features extractor device, while the back-end consists of a template or statistical classifier and a referencedatabase. The following figure 1 illustrates the architecture of ourrobust speaker identification system.

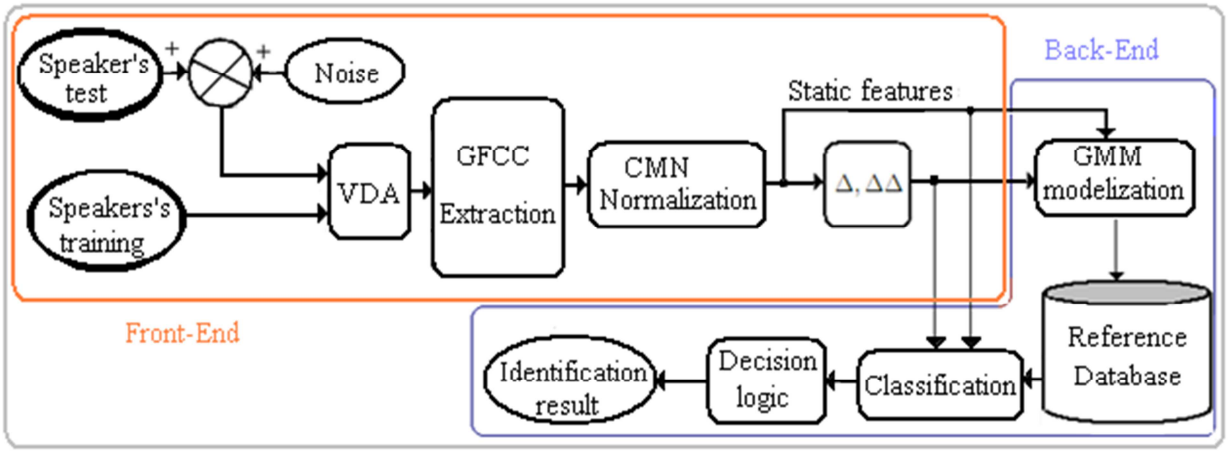

Figure 1: Architecture of the suggestedrobust speaker identification system

The main task of the front-end is to extract features from a speech signal. The aim is to sufficiently represent the Characteristics of the speech signal with reduced redundancy[1,2]. Features are extracted based on frames (windows). In other words, a short speech segment, typically 32ms, is first obtained from the entire speech signal. Then speech extraction algorithm is applied to the segment. As a result, a few numbers of coefficients are calculated. These coefficients are grouped together to form a feature vector, which represents some useful characteristics of that particular frame. After that, the above procedure repeats with subsequent utterance frames. Each new frame is some time posterior to its previous frame, typically 16ms. One feature vector is calculated for every subsequent frame. Hence, a sequence of feature vectors is generated as a result. After feature extraction, the sequence of feature vectors is passed to the back-end of the speaker identification system, which is primarily a classifier. Based on the feature vectors, the back-end of the system selects the most likely utterance out of all the possibilities from the reference database. In this work we have used a GMM statistical classifier which we have described in the previous study [3]. After training, the statistical models are stored in the database. When an unknown utterance is presented, feature vectors are obtained. The classifier calculates the maximum log likelihood based on the models and decides the most likely utterance. In order to have a good identification performance, the front-end of the system should provide feature vectors that can capture the important characteristics of an utterance. Besides, the front-end should also demonstrate reasonable performance in adverse environment.

In this research we have studied a recently robust front-end algorithm Gammatone FrequencyCepstral Coefficients (GFCC) [4,5]. Our goal is to study the performances of this front-end and make a comparison with the MFCC baseline method previously studied in [3,6]. The first part of the paper, in section 2, describes the used VDA, GTCC and CMN techniques in detail. It will explain how feature vectors are extracted. Then the second part of the paper, in section 3, will explain and depict the experimental conditions and identification results. The last part, in section 4, concluded the paper.

2. DESCRIPTION OF VAD, GFCCAND CMN USED METHODS

2.1. Description of the VoiceActivity Detector VAD

A voice activitydetector VAD permit to extract only the parts containingthe speech signal by removing the parts corresponding to a silenceand background noise. This will reduce the duration of recordings to theiruseful parts only. Hence there improved speed andperformances of the

Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

system. Several implementations are reported in the literature to design a VAD. In this study we havechoosed the solutionusing theZero crossing Rate ZCR combined to the energy of the speech signal. Indeed alow rate of zero crossingand high energy are a good indicator of the presence of aspeech signal, while a high rate of zero crossing rate and a lowenergy characterize a silencezone conta ining only backgroundnoise [7].Given the fact that the noise is characterized by its random nature, 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

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

With sign(sn) is the sign of the instantaneous value of signal s(n) acquired at time n.
N is the total length of the processed speech signal. In practice to discriminate between the presence and absence of the speech signal we have fixedtwo thresholds one for energy and one other for ZCR.Figure 2 shows an example of the evolution curve of the zero crossingrate corresponding to an utterance of speech for about 8 seconds.

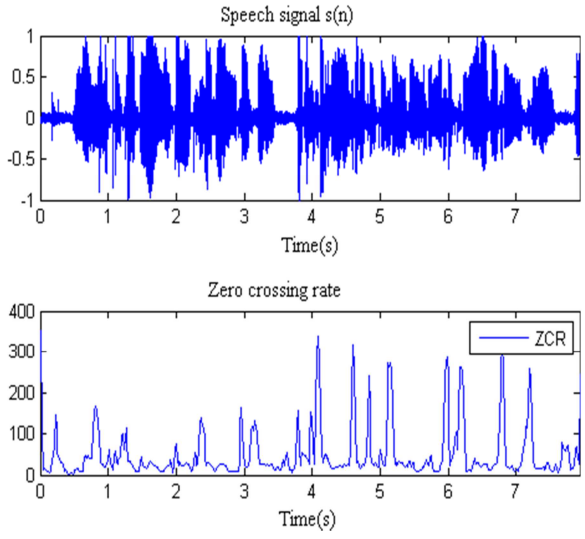

Figure 2: Example of Zero crossing rate of speech signal

2.2. Description of the GFCC algorithm

GammatoneCepstral Coefficients (GFCC) is another FFT-based feature extraction technique in speaker identification systems. The technique is based on the Gammatone filter bank (GTFB), which attempts to model the human auditory system as a series of overlapping band-pass filters [8,9,10,11]. Like MFCC, feature vectors in this technique are calculated from the spectra of a series of windowed speech frames. The following block diagram at figure 3 explains the feature extraction process.

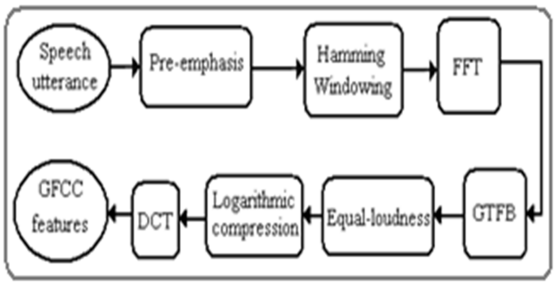

Figure3: Block diagram of the used GFCC algorithm

First, the spectrum of a speech frame is obtained. Then the speech spectrum is passed through the Gammatone filter bank. Equal-loudness is applied to each of the filter output, according to the centre frequency of the filter. After that logarithm is taken to each of the filter outputs and Discrete Cosine Transform (DCT) is applied to the filter outputs. The following sub-sections describe each step of the algorithm in detail.

A. Pre-emphasis stage

The first step of the algorithm is pre-emphasis. The idea of pre-emphasis is to spectrally flatten the speech signal and equalize the inherent spectral tilt in speech [1,2]. Pre-emphasis is implemented by a first order FIR digital filter. The transfer function of the pre-emphasis digital filter is given by the following equation (2)

$$Hp(z) = 1 - az^{-1} \quad (2)$$

where a is a constant, which has a typical

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

value of 0.97.

B. Hamming windowing and FFT stage

This stage consists first to subdivide a speech sequence into frames. The windowing function used is the Hamming window given by equation (3), which aims to reduce the spectral distortion introduced by windowing [2].

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{2N-1}\right), 0 \le n \le N-1 \\ 0 \quad otherwise \end{cases} \quad (3)$$

After windowing, Fast Fourier Transform (FFT) is applied to the windowed speech frame. The 512point FFT spectrum, S[k] for $0 \le k \le N-1$, of the speech frame is obtained as a result.

C. Gammatone filter banks stage

The Gammatonefilterbank consists of a series of bandpass filters, which models the frequency selectivity property of the human cochlea. The impulse response of each filter was introduced by Patterson [10], as shown in the following equation (4)

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \varphi) \quad (4)$$

where a is a constant, usually equals to 1. n is the order of the filter. φ is the phase shift. fc and b are respectively the centre frequency and the bandwidth of the filter in Hz.

According to Patterson [10], the centre frequency and the bandwidth of each Gammatone filter can be derived from the filter's Equivalent Rectangular Bandwidth (ERB). Under the concept of auditory modeling, the human cochlea can be modeled by a series of rectangular auditory filters, whose bandwidths are called the Equivalent Rectangular Bandwidth (ERB). The ERB of an auditory filter is related to the filter's centre frequency. Glasberg and Moore [12] suggested the following equation (5), which expresses the mathematical relationship between the centre frequency and the ERB of an auditory filter.

$$ERB(f_c) = 24.7\left(4.37\frac{f_c}{1000} + 1\right) \quad (5)$$

Patterson [10] adopted the idea of ERB to Gammatone filters and suggested that the bandwidth of a Gammatone filter should be approximately 1.019 times the ERB at its centrefrequency according to the following equation (6).

$$b = 1.019ERB = 1.019(24.7\left(4.37\frac{f_c}{1000} + 1\right) \quad (6)$$

He also suggested that a 4th order Gammatone filter (n=4) would be a good model of the auditory filter.

After defining the order and the bandwidth of the Gammatone filters, here comes a question on how to determine the centre frequencies of the filters. Slaney [4] suggested that each Gammatone filter should be spaced a given fraction (a step factor) of an ERB away from the previous filter. By integrating the reciprocal of equation 4 with a proper step factor [4], he showed that the centre frequency of a Gammatone filter can be determined by the equation (7) bellow.

Where fcm is the centre frequency of the mthGammatone filter (1 ≤ m ≤ M), fL and fH are respectively the lower and upper frequency boundaries of the filter bank in Hz. There are a total M Gammatone filters distributed between fL and fH in the filter bank.

Slaney [4] proposed an efficient implementation of the Gammatone filter bank. In his design, each 4thorder Gammatone filter in the filter bank comprises four cascaded filter stages. Each stage is essentially a 2nd order digital filter. The equations (8.1) to (8.4) bellows describe the transfer functions of these digital filters.Where T is the sampling interval.

Journal of Theoretical and Applied Information Technology
29th February 2012. Vol. 36 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$f_{cm} = \frac{-1000}{4.37} + \left(f_H + \frac{1000}{4.37}\right) exp\left(\frac{m}{M}\left(-ln\left(f_H + \frac{1000}{4.37}\right) + ln\left(f_L + \frac{1000}{4.37}\right)\right)\right) \quad (7)$$

$$H^{(1)}(z) = \frac{-2T + \left(2Te^{-2\pi bT}\cos(2\pi f_c T) + 2\sqrt{3 + 2^{3/2}}Te^{-2\pi bT}\sin(2\pi f_c T)\right)z^{-1}}{-2 + 4e^{-2\pi bT}\cos(2\pi f_c T)\,z^{-1} - 2e^{-4\pi bT}z^{-2}} \quad (8.1)$$

$$H^{(2)}(z) = \frac{-2T + \left(2Te^{-2\pi bT}\cos(2\pi f_c T) - 2\sqrt{3 + 2^{3/2}}Te^{-2\pi bT}\sin(2\pi f_c T)\right)z^{-1}}{-2 + 4e^{-2\pi bT}\cos(2\pi f_c T)\,z^{-1} - 2e^{-4\pi bT}z^{-2}} \quad (8.2)$$

$$H^{(3)}(z) = \frac{-2T + \left(2Te^{-2\pi bT}\cos(2\pi f_c T) + 2\sqrt{3 - 2^{3/2}}Te^{-2\pi bT}\sin(2\pi f_c T)\right)z^{-1}}{-2 + 4e^{-2\pi bT}\cos(2\pi f_c T)\,z^{-1} - 2e^{-4\pi bT}z^{-2}} \quad (8.3)$$

$$H^{(4)}(z) = \frac{-2T + \left(2Te^{-2\pi bT}\cos(2\pi f_c T) - 2\sqrt{3 - 2^{3/2}}Te^{-2\pi bT}\sin(2\pi f_c T)\right)z^{-1}}{-2 + 4e^{-2\pi bT}\cos(2\pi f_c T)\,z^{-1} - 2e^{-4\pi bT}z^{-2}} \quad (8.4)$$

It can be shown that the transfer function of the Gammatone filter, H(z), is the product of the transfer functions of these cascaded filters given by the following equation (9)

$$H(z) = H^{(1)}(z).H^{(2)}(z).H^{(3)}(z).H^{(4)}(z) \quad (9)$$

In GFCC algorithm, we are interested to the magnitude response, ||H(ω)|| (or in digital domain, ||H[k]||), of the filter transfer function. The magnitude response can be obtained by substituting $z=e^{j\omega}$ into H(z). The following figure 4 illustrates the magnitude responses of a series of Gammatone filters in a filter banks.

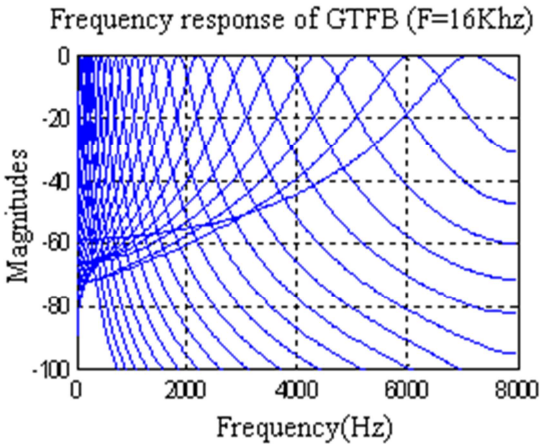

Figure 4: Frequency response of Gammatone filter banks

In addition the frequency responses of the filters are normalized. Hence the magnitude response at the centre frequency is equal to unit. As previously described in sub-section B) we applying FFT to a windowed speech frame for obtaining its frequency spectrum, S[k]. The power spectrum, $|S[k]|^2$, is then calculated on the first half of the frequency spectrum (the second half of the spectrum is just the mirror image of the first half and is discarded). The next step is to find out the filter output, X_m, which can be expressed by the following

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

equation (10),

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |[S[k]|^2 |H_m[k]| \quad 1 \le m \le M \quad (10)$$

Where N is the number of FFT points in each windowed frame and $H_m[k]$ is the frequency response of the m[th]Gammatone filter.

D. Equal-loudness stage

As described in a previous paper [6], equal loudness technique which we have used in PLP front-end, is also applied to the Gammatone filter outputs according to their centre frequencies. The equal loudness weight of the m[th] filter, E_m, can be found by evaluating either equation (11) for applications, whose Nyquist frequency is up to 5 kHz or equation (12) for applications, whose Nyquist frequency is above 5 kHz.

$$E = \frac{(w^2 + 56.8 \times 10^6)w^4}{(w^2 + 6.3 \times 10^6)^2(w^2 + 0.38 \times 10^9)} \quad (11)$$

$$E = \frac{(w^2 + 56.8 \times 10^6)w^4(w^6 + 9.58 \times 10^{26})^{-1}}{(w^2 + 6.3 \times 10^6)^2(w^2 + 0.38 \times 10^9)} \quad (12)$$

The filter output after equal loudness, $X_m(e)$, is simply the product of the filter output and the equal loudness weight is given by the following equation (13)

$$X_m(e) = E_m . X_m \quad 1 \le m \le M \quad (13)$$

Alternatively, the weight can be applied to the magnitude response of each Gammatone filter according to equation (14) and after that the filter outputs can be directly obtained from the weighted Gammatone filters by applying equation (15).

$$H_m(e)[k] = E_m . H_m[k] \quad 1 \le m \le M \quad (14)$$

$$X_m(e) = \sum_{k=0}^{\frac{N}{2}-1} |[S[k]|^2 . |H_m(e)[k]| \quad 1 \le m \le M \quad (15)$$

E. Logarithmic compression stage

The next step of the algorithm is to apply logarithm to each filter output according to equation (16). The aim of this procedure is to simulate the human perceived loudness given certain signal intensity. Let $X_m(\ln + e)$be the logarithmically-compressed filter output of the m[th]Gammatone filter.

$$X_m(\ln + e) = \ln(X_m(e)) \quad 1 \le m \le M \quad (16)$$

F. Discrete cosines Transformation DCT stage

The last stage of the algorithm consistsof correlating the filter outputs. For this the Discrete Cosine Transform (DCT) is applied to the filter outputs. Suppose p is the order of GFCC. The feature vector of one speech frame, which has p number of GFCC coefficients, contains the first p DCT coefficients of the filter outputs. The k[th] GFCC coefficient of the feature vector is defined by the following equation (17) where: $1 \le k \le p$

$$GTCC_k = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} \left\{ X_m(\ln + e) . \cos\left(\frac{\pi k(m - 0.5)}{M}\right) \right\} \quad (17)$$

Often an additional component is appended to the p[th] order feature vector. The additional component can either be a log energy term or the zeroth order GFCC coefficient or both. The log energy term is simply the logarithm of the energy of one speech frame. The zeroth order GFCC coefficient is the zeroth order DCT coefficient of the filter outputs given by (18).

$$GTCC_0 = \sqrt{\frac{1}{M}} \sum_{m=1}^{M} X_m(\ln + e) \quad (18)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Given the fact that these two components do not contain any specific information relevant to the discrimination of speakers [1,2], they are not used in this study and only the 12 first static coefficients (K=1 to 12) plus their first and second derivative are used. In total we have used 36 parameters.

2.3. Cepstral Mean NormalizationCMN

CMN normalization ensures that the values in the feature vectors have zero mean and unit variance. This will help avoid the risk that larger values will have a greater influence on the behavior of different treatments in subsequent identification steps. Given the fact that the cepstral coefficients of speech signals have generally a zero mean then to remove noise, we must simply subtract each cepstral coefficient the average of all cepstral coefficients characterizing the speech signal in question. This operation is known by Cepstral Mean subtraction (CMS). These two treatments will help improve the identification scores of our system. For this we have used the empirical definitions of mean and variance of feature vectors given respectively by equations(19) and (20) below.

$$\mu = \frac{1}{N}\sum_{n=1}^{N} c(n) \quad (19)$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (c(n) - \mu)^2 \quad (20)$$

Wherec(n) is the features vector of the n^{th} frame, N is the total number of frames in the analyzed speech signal.

Finally to obtain the normalized features vector $\hat{c}(n)$ using the CMN technique we have first subtract the average cepstral features vector μ for each c(n) and then dividing the result by the standard deviation σ according to equation (21) follows.

$$\hat{c}(n) = \frac{c(n) - \mu}{\sigma} \quad (21)$$

The CMN is an alternate way to high-pass filter cepstral coefficients which allows to compensate the effects of unknown linear filtering and force the average value of cepstral coefficients to be zero in both the training and testing domains. Nevertheless, CMN can compensate directly the combined effects of additive noise and linear filtering.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental conditions

In this study, we are interested to evaluate the performance of the suggested front-end based on GFCC method in a text-independent monaural speaker identification context. For this we are built our proper database which corresponding to a population of 51 Arabic-speakers (35 male and 16 female). Each speaker had participated by 2 different recordings: one for learning the database for about 20s per utterance and one other for the test step for about 10s per utterance. All the productions sound from the speakers, were directly digitized to .wav format with a sampling frequency of 16 kHz and 16-bit quantification using the well-known software Wavesurfer®8 [13]. A white Gaussian noise of variable level (0db <SNR <40dB) was added to the recorded test utterances only to examine the robustness of described technique in noisy environments that are inevitable in most real applications. The features extractors that will be considered in this set of experiments are MFCC, ΔMFCC, ΔΔMFCC, GFCC, ΔGFCC, and ΔΔGFCC. The classifier of our system is the Gaussian Mixture Models (GMM) which is considered actually as the state of the art in text-independent speaker identification task [14,15,16]. The entire identification system is implemented in a MATLAB®7 programming environment. The following table 1 describes the experiment conditions in detail.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE 1: EXPERIMENT CONDITIONS OF SPEAKER IDENTIFICATION SYSTEMS

Task system	Text-independent automatic speaker identification
language	Arabic
Front-ends	MFCC, ΔMFCC, ΔΔMFCC, GFCC, ΔGFCC, ΔΔGFCC
Back-end	Gaussian mixture models (GMM)with 8 mixture
Number of coefficients in a feature vector	36(12static + 12delta + 12delta-delta) for MFCC & GFCC
Window size	32 ms
Step size	16 ms
Sampling rate	16kHz
Training set	51 speakers (one utterance per speaker for about 20s)
Test set	51 speakers (one utterance per speaker for about 10s)
Noise Type	White Gaussian Noise (WGN)
SNR range	0, dB, 5dB, 10dB, 15dB, 20dB, 25dB , 30dB, 35dB, 40dB
Platform	Laptop PC HP 512 Intel Pentium M 2.13Ghz
Programming Language	MATLAB®7
Acquisition tool	Wavesurfer®8

3.2. Experimental results

The evaluation of the identification performances of our systems was done by applying the empirical equation (22)

$$C = \frac{H}{N} \times 100\% \quad (22)$$

Where C is the percentage of correctly identified speakers called identification/recognition rate RR, H is the number of correctly identified speakers and N is the total number of speakers that have participated to test identification. The following table 2 and figure 5 show the identification rate of MFCC, ΔMFCC, ΔΔMFCC, GFCC, ΔGFCC and ΔΔGFCC front-ends in various SNR conditions. These results indicate clearly that the algorithm GFCC produces interesting results. In a clean speech environment GFCC works perfectly like MFCC. However, in noisy environments, all variants of GFCC exceed all variants of MFCC. Second the dynamic variants of these two algorithms work better than static variants but they occur a long time to estimate the parameters of the GMM models. The average identification rate of GFCC is about 55.31% while the average identification rate of MFCC is still equal 50.53% when SNR changes from 40dB to 0dB. In others words, these results indicate that in noisy environments the GFCC algorithm works better than MFCC and the dynamic variants of these algorithms are better suited to robust conditions.

TABLE 2: PERCENTAGE OF CORRECTLY IDENTIFIED SPEAKERS (C) IN VARIOUS SNR ENVIRONMENTS CORRESPONDING TO BASELINE MFCCS AND PROPOSED GFCC FRONT-ENDS COMBINED TO VAD AND CMN TECHNIQUES

SNR(dB)	0	5	10	15	20	25	30	35	40	**MOY.**
MFCC	07.22	9.45	12.23	23.67	43.15	63.43	95.67	100	100	**50.53**
ΔMFCC	10.47	12.78	15.93	25.15	45.54	66.35	96.65	100	100	**52.54**
ΔΔMFCC	11.34	13.65	16.43	26.83	46.52	68.74	96.89	100	100	**53.37**
GFCC	13.87	17.13	24.32	32.17	48.23	66.83	95.25	100	100	**55.31**
Δ GFCC	17.34	21.87	27.70	35.28	51.35	70.34	96.54	100	100	**57.82**
ΔΔ GFCC	18.68	22.05	28.18	36.38	52.82	72.13	96.75	100	100	**58.55**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

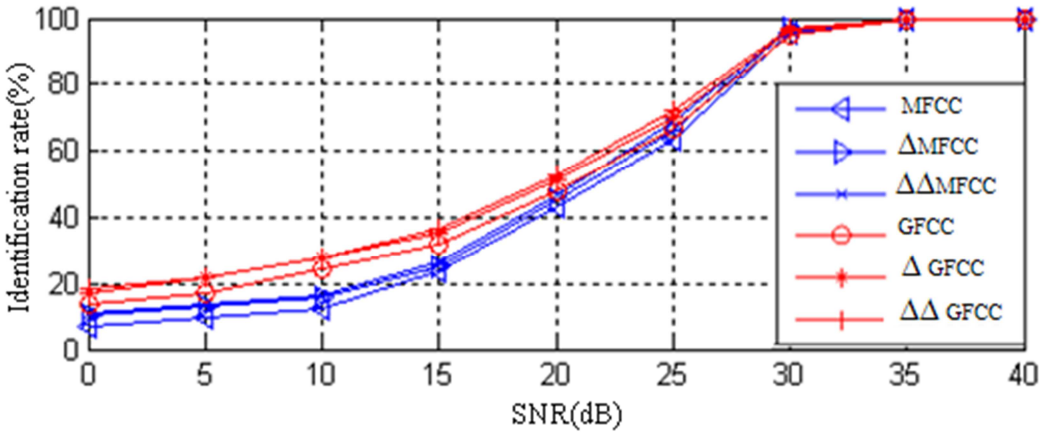

Figure 5: Performances of GFCCs and MFCCs front-ends versus SNR.

4. CONCLUSION AND FUTURE WORK

A typical speaker identification system consists of a front-end and a back-end. The front-end is basically a features extraction module, while the back-end is primarily a classifier. The function of the back-end is to make accurate identification decisions based on the speech features extracted by the front-end. Hence, the quality of a front-end algorithm plays an important role in speaker identification systems. In this experimental study the robust front-end GFCC algorithm combined to VAD and CMN techniques was implemented, evaluated and compared to the baseline and conventional front-end MFCC. The experiment results show that the proposed approach outperforms the baseline methods MFCCs in noisy conditions by approximately **5.08%** when SNR changes from 40dB to 0dB. However in clean environment (SNR≥35dB)our architecture performs equally well with MFCC and all its variants. Finally, compared to the ordinary Mel frequency cepstral coefficients, the speaker identification system based on GFCC combined to VAD and CMN techniques presented gives better identification rate and robustness characteristics.

In perspective to this work, we intend to implement the suggested architecture on a Digital Signal Processor DSP in order to use it in a real application for secured access to high secure areas.

REFRENCES:

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

[2]Yassine Mami, Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence, Thèse de ENST, Paris France octobre 2003.

[3] E. B. Tazi, A. Benabbou, M. Harti, "Design of an automatic speaker recognition system based on adapted MFCC and GMM methods for arabic speech", in IJCSNS journal, Vol.10 No.1, January 2010.

[4] 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

[5] O. Cheng, W. Abdulla, Z. Salcic, performance evaluation of front ends processing for speech recognition systems, faculty of engirneering university of AUCKLAND Report No. 621, 2005

[6]B. Tazi, A. Benabbou, M. Harti, Etude Comparative des Méthodes de Paramétrisation Conventionnelles MFCC, PLP et LPCC pour la Reconnaissance Automatique du locuteur Arabophone, RNIOA'11, FST Errachidia le 24,25 mars 2011.

[7] Atsuzuka Y, Highly sensitive speech detector and high-speed voice-band data discriminator in DSI-ADPCM systems, IEEE Trans. On

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Communications, vol. COM-30, no.4, pages 739-750, USA 1982.

[8] 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.

[9]D. Kim, S. Lee and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments" IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, pp. 55-69, 1999.

[10]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

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

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

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

[14]Reynolds, Douglas A. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker, Models. IEEE Transactions on Speech and Audio Processing. vol. 3, n. 1, pp. 72-83, January, 1995.

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

[16]Bing Xiang and Toby Berger, Efficient text independent speaker verification with structural gaussian mixture models and neural network, IEEE transactions on speech and audio processing, Vol. 11, NO.5, September 2003.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTHOR PROFILES:

Mr El Bachir TAZIgraduated in Electronic Engineering from ENSET Mohammedia in 1992. He obtained his postgraduate diploma DEA and University Doctorate in Automatics and signal processing at the Faculty of Sciences, USMBA University, Fez in 1995 and 1999 respectively. Now he is working on his PhD. His thesis is on the Robust Speaker Identification. Professor of Computer sciences and Member of the Research Unit "INTIC" at USMBA Fez, Morocco.

MrAbderrahim BENABBOUreceived the PhD in Applied Computer Sciences from ENSIAS, Mohammed V-Souissi University, Rabat, in 2002 andUniversity Doctorate in Computer Sciences from Mohammed V University, Rabat, in 1997. Professor in Sidi Mohamed Ben Abdellah University USMBA, FST, Department of Computer Engineering Fez, Morocco. Major Fields: Natural Language Processing, Speech processing, Human-Machine Interfaces, Embedded Systems, Artificial Intelligence and Object Paradigm.

MrMostafa HARTIreceived the PhD in Computer Sciences and Statistics from ULB University, Belgium, in 1996. He received the University Doctorate in Computer Sciences from University, Nancy I, France in 1986. Professor in Sidi Mohamed Ben Abdellah University, Fez, Morocco. Major Fields: G. I. S, Information System Governance, Databases, development, Language Processing Text and Speech, Statistics. Chairman of the Research Unit " INTIC" at the Faculty of Sciences DharMahraz, Sidi Mohamed Ben Abdellah University USMBA Fez,