Robust speaker identification system over AWGN channel using improved features extraction and efficient SAD algorithm with prior SNR estimation

Riadh Ajgou, Salim Sbaa, Said Ghendir, Ali Chemsa, Abdelmalik Taleb-Ahmed

Robust Speaker Identification System Over AWGN Channel Using Improved features Extraction and Efficient SAD Algorithm with Prior SNR Estimation.

Riadh AJGOU, Salim SBAA, Said GHENDIR, Ali CHEMSA and A. TALEB-AHMED

***Abstract*—** This paper motivates the combination of Autoregressive (AR) parameters and Mel Frequency Cepstral Coefficients (MFCC) features for remote robust text-independent speaker identification. All speaker identification techniques start by converting the raw speech signal into a sequence of acoustic feature vectors carrying distinct information about the signal. The most commonly used acoustic vectors are Mel Frequency Cepstral Coefficients (MFCC) which is a very useful feature for speaker identification system but it deteriorates in the presence of noise, thus to have better identification rate, we have developed a robust feature extraction method based on the combination of MFCC and Autoregressive model (AR) parameters modeled with GMM model. To improve the identification rate accuracy, an efficient speech activity detection (SAD) algorithm based on prior SNR estimation are proposed in the pre-processing phase. To validate our work TIMIT database with speech from 630 speakers has been used. The first four utterances for each speaker could be defined as the training set while 1 utterance as the test set. The use of AR-MFCC approach has showed significant improvements in identification rate accuracy when compared with MFCC. However, in terms of runtime, AR-MFCC requires more time to execute than MFCC. Our SAD algorithm has provided a suitable contour of speech activity in noisy conditions.

***Keywords*— S**peaker recognition, feature extraction, AR-MFCC, GMM; SAD; EZR.

I. Introduction

Speaker recognition or voice classification is the task of recognizing people from their voices[1]. Such systems extract features from speech signal, process them and use them to recognize the person from the voice [1].

Riadh AJGOU, Said GHENDIR, Ali CHEMSA are with, LI3CUB Laboratory, Electric engineering department, University of Biskra. B.P 145, 07000 Biskra ALGERIA. (Email: Riadh-ajgou, said-ghendir {@univ-eloued.dz }, chemsadoct@yahoo.fr) and with department of Sciences and Technology, University of Eloued, PO Box 789, 39000 El-Oued, ALGERIA.

Salim SBAA is with LI3CUB Laboratory, Electric engineering department, University of Biskra. B.P 145 R.P, 07000 Biskra ALGERIA. (Email: s.sbaa@univ-biskra.dz).

A. TALEB-AHMED Author is with LAMIH Laboratory University UVHC, 59313 Valenciennes Cedex 9 FRANCE. (Email: abdelmalik.taleb-ahmed@univ-valenciennes.fr)

Speaker recognition combines speaker verification and speaker identification. Speaker verification is the technique to verify a person's claimed identity by making use of the speech cures. On the other hand, in speaker identification process, no identity claims are made and the system has to identify the speaker.

In our work, the system that we described is classified as text independent speaker identification system.

Otherwise, the most difficult tasks in the speaker recognition are feature extraction, speech activity detection (SAD) (Speech/non-speech determination) and speaker modeling. Gaussian Mixture Model (GMM) represents vocal tract configurations that are effective for text-independent speaker identification system [2]. In this paper we have used GMM for speaker modeling. The individual component of GMM represents some vocal tract configurations that are speaker dependent for identifying the speaker [2].

All speaker identification techniques start by converting the raw speech signal into a sequence of acoustic feature vectors carrying distinct information about the signal. This feature extraction is also called “front-end” in the literature [1]. The most commonly used acoustic vectors are Mel Frequency Cepstral Coefficients (MFCC) [3, 4]. However, MFCC’s are very useful in clean conditions but deteriorates in the presence of noise [5]. Furthermore, Autoregressive models, is also important to represent speech signal [6, 7].

In this paper, we have suggested to improve the speaker identification accuracy by combining MFCC and Autoregressive features (AR-MFCC) over AWGN channel. Furthermore, we have proposed efficient speech activity detection (SAD) algorithm based on prior SNR estimation where the proposed SAD algorithm is based on Zero Crossing Rate and Energy Measurements.

The recognition is usually performed over features extracted from the decoded signal, although it is also possible to extract the recognition features directly from the codec parameters. Fig. 1 presents a scheme of this system architecture where the implementation is over an IP (Internet Protocol) network [8].

Our proposal speaker identification system on the remote

communication channel is described in more detail in the next section.

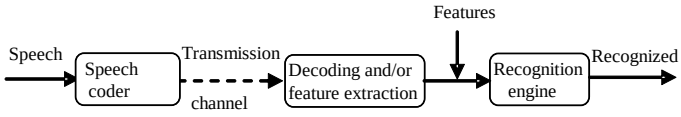

Fig.1. Scheme of a network speaker/speech recognition system.

II. PROPOSED IDENTIFICATION SYSTEM CONFIGURATION

In this section we consider the identification system of speaker. Any identification system consists of three parts training stage, test stage and decision stage. In the proposed system, test and training stages based on two main blocks, pre-processing where our proposed speech activity detection algorithm is used to detect speech/non-speech zone and feature extraction where feature extracting is accomplished by the combination of MFCC and Autoregressive model (AR) parameters. The system we used for experiments was established according to the following diagram in Fig. 2.

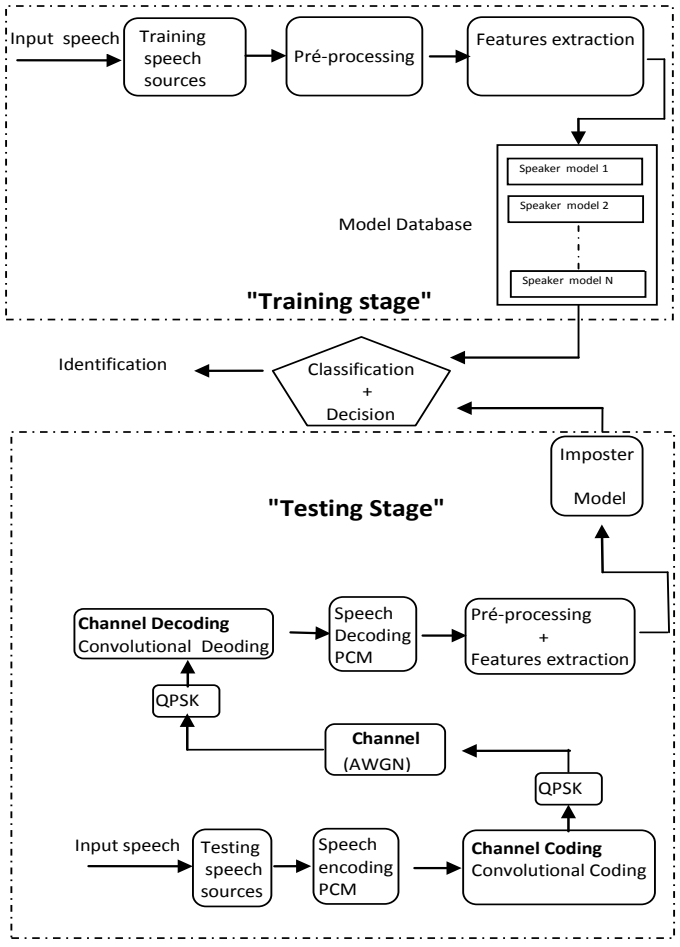

Fig.2. Different stages of the proposed speaker identification system.

A. Pre-processing stage

In pre-processing step, silence portions are removed from the speech signals by using the speech activity detection algorithm as in [9] where we have developed a robust SAD (the detailed procedure is explained in section III). Then, each utterance is pre-emphasized with a pre-emphasis factor of 0.97 which leads to improve identification rate accuracy?

The sampled speech is pre-emphasized to enhance the high frequency components of the spectrum, especially the so-called formants, against the lower frequencies which contain most of the signal's power, but are known to be rather irrelevant for speech intelligibility. Pre-emphasis of the high frequencies is done to obtain similar amplitudes for all the formants [10]. Speech signal was emphasized using a high pass filter. Commonly a digital filter with 6dB/ Octave is used. The constant μ in equation (1), is usually chosen to be 0.97 [11]:

$$y(z) = 1 - \mu z^{-1} \quad (1)$$

B. Proposed feature (AR-MFCC)

The goal is using features suitable for speaker identification. Thus, in this work, we have combined MFCC features with autoregressive model coefficients. The number of coefficients is 64 (32 MFCC and 32 AR). The acoustic signal contains different kinds of information about the speaker. The signal processing involved changes depending on the type of characteristics we are interested in the speaker.

The MFCC feature set is based on the human perception of sound, on the known evidence that the information carried by low-frequency components of the speech signal are phonetically more important for humans than the high-frequency components [12]. The human perceptual frequency is represented in mel scale which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz . The human perception of sound is assumed to be consist of bank of filters. Each filter is of triangular in shape. The triangular filter banks in mel scale are uniformly spaced [13].

In our work, speech is segmented in frames of 20 ms, and the window analysis is shifted by 10 ms. Each frame is converted to 32 MFCCs. The mapping from linear frequency to Mel-Frequency is shown in equation (2), f in Hz [13]:

$$\mathrm{mel}(f) = 2595 \log 10\left(1 + f/700\right) \quad (2)$$

f_{low} and f_{high} are the low and high frequency boundaries of filter bank, they are given as [13]:

$$f_{low} = \frac{fs}{N} \quad (3)$$

N: is the frame size which is done with a frame size of 160 samples (corresponds to 20ms).

$$f_{high} = \frac{fs}{2} \quad (4)$$

AN: f_{low}=100 Hz, f_{high}=8000Hz.

Accordingly to equation (2) we have the Fig.3 that represents the relationship between the actual frequency scale and its Mel-frequency scale, when the actual frequency f is below 1000Hz, the relationship is linear; however, the relationship of the features becomes logarithmic when f is above 1000Hz, the bandwidth of the individual filter (filter bank) increases logarithmically in the normal scale. Each triangular filter is of length 1,000 (arbitrarily chosen) in frequency domain. Also note that the 1,000th sample corresponds to: Fs/2 [14]. Also, the number of filters used is 24. The Fig.4 illustrates the bank of filters in mel-frequency scale. (1 in x-axis corresponds to Fs/2).

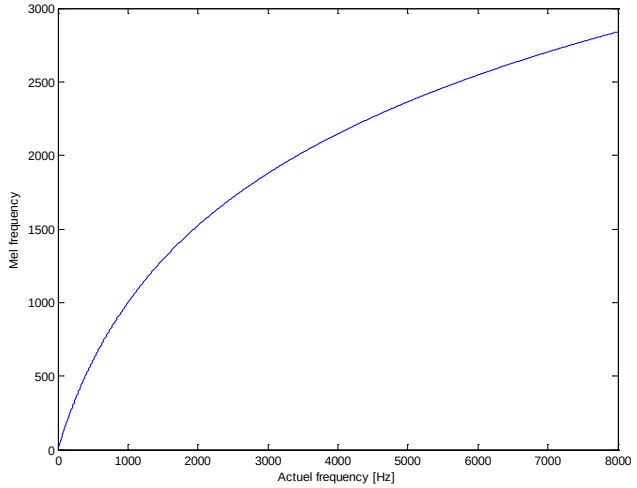

Fig.3. Curve relationship between the actual frequency scale and its Mel-frequency scale.

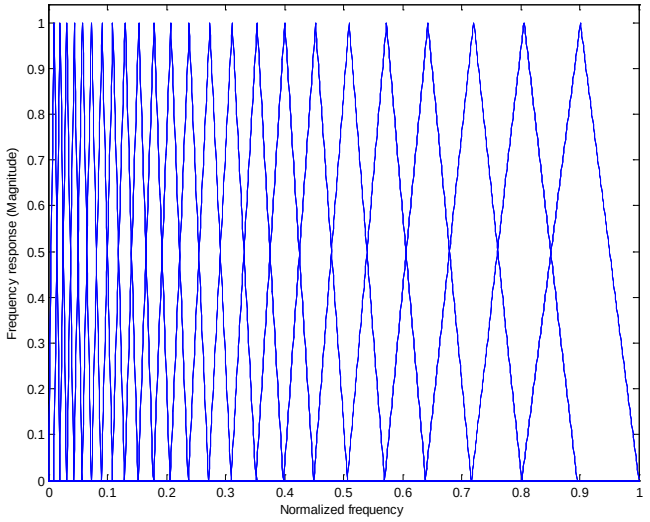

Fig.4. Bank of filters in mel-frequency scale. 1 in x-axis corresponds to Fs/2: (8000 Hz)

Once log mel spectrum has been computed, it has to be converted back to time domain by using Discrete Cosine Transform (DCT). The result is called the mel frequency cepstrum coefficients (MFCCs). Using the same procedure, a set of mel-frequency cepstrum coefficients are computed for each speech frame of about 20 ms with overlapping manner.

Autoregressive models are widely used models, [15]. In general, the number of prediction coefficients $\{a_1, a_2, \ldots a_P\}$ is infinite since the predictor is based on the infinite past, In our case, we limited the number of coefficients to 32 (p=32). The calculation of AR coefficients has been carried out by the Yule-Walker method [13], this method solves the Yule-Walker equations by means of the Levinson Durbin recursion. The Autoregressive Model (signal model B(z)) is an all-pole filter of the type [15]:

$$B(Z) = \frac{1}{A(Z)} = \frac{1}{1 + a_1 Z^{-1} + a_2 Z^{-2} + \ldots\ldots a_p Z^{-p}} \quad (5)$$

Where:

$$A(Z) = 1 + a_1 Z^{-1} + a_2 Z^{-2} + \ldots\ldots a_p Z^{-p} \quad (6)$$

The figure 5presents the AR-MFCC extraction procedure for a speech signal consists of five frames that represents speech, where we extract MFCC and AR features from each frame and reconstruct one matrix of AR-MFCC as shown in this figure.

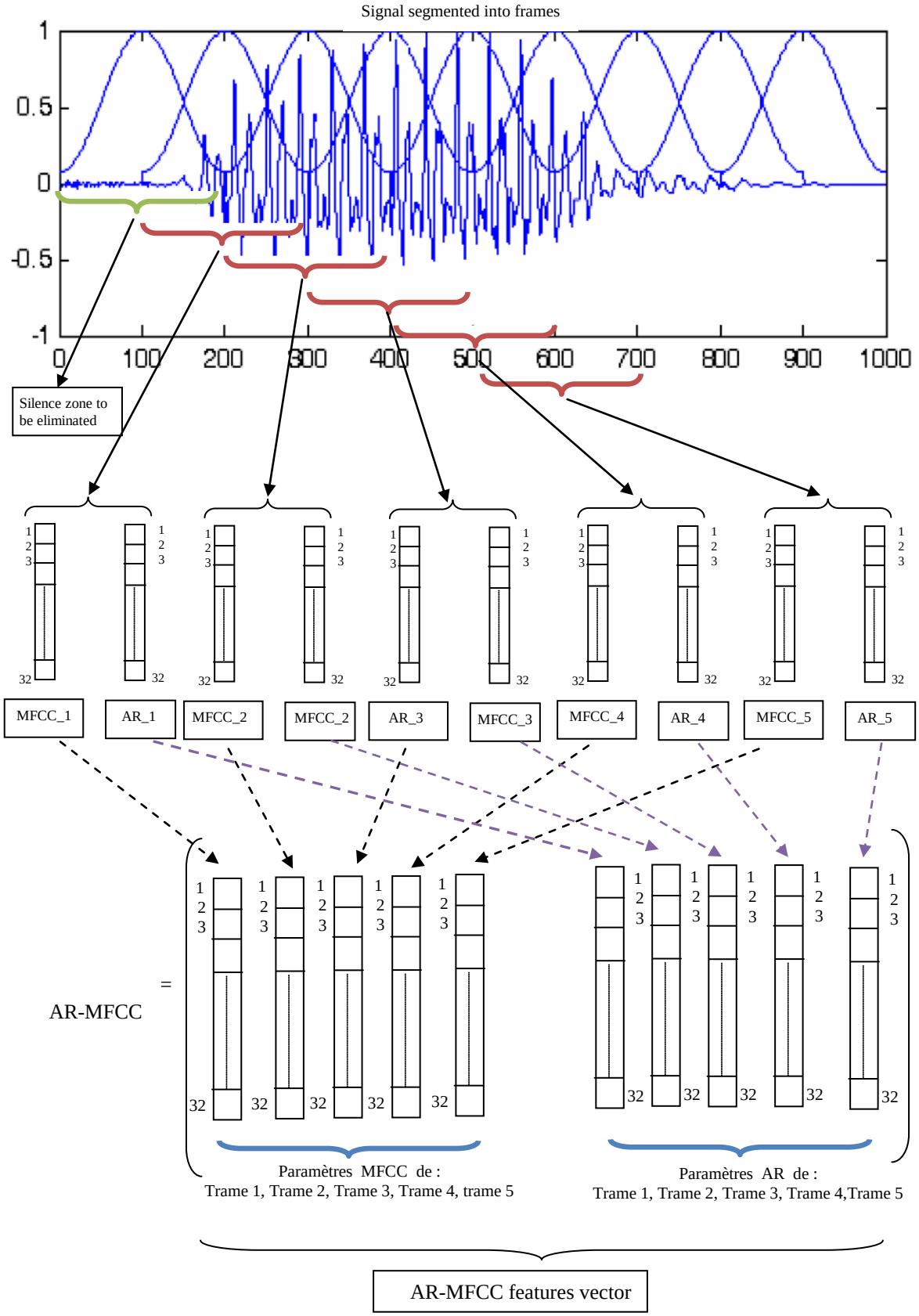

Fig.5 AR-MFCC extraction procedure

C. Speaker modeling using GMM

In the training stage, pattern generation is the process of generating speaker specific models with collected data. In our, work, the generative model used is the Gaussian mixture model (GMM) [2]. In GMM, we model the speaker data (feature vectors) using statistical variations of the features. Hence, it provides a statistical representation of how speaker produces sounds. Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker.

The system was trained using speakers from the TIMIT database [16] where we have chosen 200 speakers from different regions. Moreover, in the training stage, we have used four utterances for each speaker. Speech signal passed through pre-processing phase (emphases + SAD), and sixty-four coefficients are extracted (32 mel-frequency cepstral coefficients and 32 parameters of the autoregressive model) and models characterization using GMM are formed.

D. Testing phase

Speech signal is coded using PCM code. In addition, a convolutional code [8], with a rate of ½ as channel forward error correction, has been introduced in order to make the channel more robust to noise, the coded signal is then transmitted through AWGN channel. After demodulation (QPSK), convolutional decoding, and PCM decoding, the binary data is converted back to a synthesized speech file. as a final point, from file synthesized speech, MFCC coefficients and AR parameters are extracted.

*E. D*ecision phase

Pattern matching is the task of calculating the matching scores between the input feature vectors and the given models. The GMM forms the basis for both the training and classification processes. The principle of GMM is to abstract a random process from the speech, then to establish a probability model for each speaker [2]. A Gaussian Mixture density is a weighted sum of M component densities.

In the GMM model, the features distributions of the speech signal are modeled for each speaker as follows [2, 17]:

$$p(\bar{x}/\lambda) = \sum_{i=1}^{M} P_i \;\; b_i(\bar{x}) \tag{7}$$

Where:

$$\sum_{i=1}^{M} P_i = 1 \tag{8}$$

And x is a random vector of D-dimension, $p(x/\lambda)$ is the speaker model; *pi* is the ith mixture weights; *b*i(x) is the ith pdf component that is formed by the ith mean μ_i and ith covariance matrix, where *i* = 1,2,3,…,*M*. *M* is the number of GMM components [2, 17], each density component is a D-variants Gaussian distribution of the form:

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2}.\left|\sum_m^s\right|^{1/2}} \exp\left[-\frac{1}{2}(\bar{x}-\overline{\mu_i})'(\sum_i^{-1})(\bar{x}-\bar{\mu}_i)\right] \tag{9}$$

A statistical model for each speaker in the set is developed and denoted by λ. For instance, speaker *s* in the set of size *S* can be written as follows [2, 17]:

$$\lambda_s = \left(P_i, \bar{\mu}_i, \sum_i\right), \text{ i=(1.....M), s= \{1.....S\}} \tag{10}$$

1) ML Parameter Estimation Steps (training)

To obtain an optimum model for each speaker we need to obtain a good estimation of the GMM parameters. The Maximum-Likelihood Estimation (ML) approach can be used; where for a given *T* vectors used for training, **X**=(*x*1,*x*2,...,*xT*), the likelihood of GMM can be written as [2, 17]:

$$p(X/\lambda_s) = \prod_{t=1}^{T} p(x_t/\lambda_s) \tag{11}$$

Since it's impossible to directly maximize a nonlinear function with GMM likelihood approach, the ML estimations can be done using the EM algorithm iteratively [17]. The training phase consists of two steps, namely *initialization* and *expectation maximization (EM)*. The initialization step provides initial estimates of the means for each Gaussian component in the GMM model. The EM algorithm recomputed the means, covariances, and weights of each component in the GMM iteratively. The EM algorithm steps and formulas are [17]:

- new estimates of *'i'* th weight :

$$\bar{P}_i = \frac{1}{T}\sum_{t=1}^{T} P(i \setminus x_t, \lambda) \tag{12}$$

- new estimates of mean:

$$\bar{\mu}_i = \frac{\sum_{i=1}^{T} P(i \setminus x_t, \lambda) x_t}{\sum_{n=1}^{N} P(i \setminus x_t, \lambda)} \tag{13}$$

- New estimates of diagonal elements of *'i'* th covariance matrix [2, 17]:

$$\bar{\sigma}_i^2 = \frac{\sum_{i=1}^{T} P(i \setminus \bar{x}_t, \lambda) x_t^2}{\sum_{n=1}^{T} P(i \setminus \bar{x}_t, \lambda)} \bar{\mu}_i^2 \tag{14}$$

- where the likelihood a *posteriori* of the i-th class is given by posterior probability [2, 17]:

$$P(i \setminus \bar{x}_t, \lambda) = \frac{p_i b_i(\bar{x}_t)}{\sum_{k=1}^{M} P_k b_k(\bar{x}_t)} \quad (15)$$

This process is repeated until convergence is achieved.

2). Classification and identification

After estimating the GMM models for each speaker, the problem is to find a model with maximum likelihood posteriori for an observation sequence. The input to the classification system is denoted as [17]:

$$X = \{x1, x2, x3, \ldots, x_T\}. \quad (16)$$

The rule through which we determine whether *X is* coming from speaker *'s'* can be stated as:

$$p(\lambda s \mid X) > p(\lambda r \mid X), \quad r = 1, 2, \ldots, S \ (r / = s) \quad (17)$$

The classification system needs to compute and find the value of 's' that maximizes $p(\lambda s \mid X)$ according to [2, 17] :

$$\hat{S} = \arg\max_{1<s<S} P(\lambda_s \setminus x) = \arg\max_{1<s<S} \frac{P(x \setminus \lambda_s)\Pr(\lambda_s)}{P(x)} \quad (18)$$

The classification is based on a comparison between the probabilities for each speaker. If it can be assumed that the prior probability of each speaker is equal, then the term of $p(\lambda s)$ can be ignored [17]. The term $p(X)$ can also be ignored as this value is the same for each speaker, so $p(\lambda s \mid X) = p(X \mid \lambda s)$, where [2, 17]:

$$p(X / \lambda_s) = \prod_{t=1}^{T} p(x_t / \lambda_s) \quad (19)$$

The speaker of the test data is statistically chosen according to [17]:

$$\hat{S} = \arg\max_{1<s<S} P(x \setminus \lambda_s) \xrightarrow{take\log} \hat{S} = \arg\max_{1<s<S} \sum_{t=1}^{T} \log P(\bar{x}_t \setminus \lambda_s) \quad (20)$$

The speaker identification Rate is given by:

$$Id = \frac{No.of\ .utterance..correctly.identified}{Total.No.of\ .utterance.under.test}.100\% \quad (21)$$

III. PROPOSED SPEECH ACTIVITY DETECTION ALGORITHM

The robust SAD algorithm is based on two original works: [18] and [19]. In [18] the author had used the LPC residual energy and zero crossing rates to detect speech activity using adaptive threshold, where this threshold is calculated for every frame introduced in comparison with previous calculated features of frames. This means probable mistakes for the first frames since the algorithm is initiated and spans up to a few frames (0-15) frames considered as non-speech. The second author in [19] used energy and zero crossing rates ratios to voiced/non voiced classification of speech using a fixed threshold.

Our robust SAD is based on Energy and Zero crossing rate Ratios (EZR) using efficient thresholds based on Prior SNR estimation for detecting voiced segments. The procedure of calculating threshold is as follows:

A. Segmenting the whole speech signal into frames:

At first the speech signal segmented into frames of 8ms with rectangular window and without overlapping.

B. Calculating energy and zero crossing rates:

We calculate short term energy $\bar{E}[m]$ and zero crossing rate ZCR[m] [20]:

$$\bar{E}(m) = \sum_{n=0}^{N-1} x^2(n).w(m-n) \quad (22)$$

Where: w is a rectangular window of length N (length of a frame) and x(n) is the frame signal with N samples. ZCR is defined as [20, 21]:

$$ZCR(m) = \sum_{n=0}^{N-1} \left|\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)\right| w(m-n) \quad (23)$$

Where sgn(.) is the signum function which is defined as [20, 21]:

$$\mathbf{sgn}[x(n)] = \begin{cases} +1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases} \quad (24)$$

C. Calculating the ratio of the average energy and zero crossing rate (EZR)

Usually speech segments have high energy and low zero crossing rate and non-speech segments have low energy and high zero crossing rate. Thus, our method works on the principle of extracting energy and zero crossing rate features from the input speech signal. So that, we calculate the ratio of the energy average and zero crossing rate (EZR) and comparing them to the threshold to classify the frames into speech and nonSpeech classes.

If the EZR of the frame is greater than the threshold, the frame is judged as a speech frame. Otherwise, the frame is considered to be non-speech frame (the recognition system does not extract features from this frame which leads to a good recognition rate):

$$EZR[m] = \frac{\bar{E}[m]}{ZCR[m]} \quad (25)$$

D. Calculating the maximum and minimum of EZR

After segmenting speech signal into frames, we calculate EZR for each frame. So that, we calculate the values of minimum and maximum of EZRs.

E. Signal-to-Noise Measure of speech signal by assuming noise variance

Our SAD algorithm requires knowledge of SNR to estimate threshold. There are two possibilities of SNR calculation. One is to estimate the ratio of the signal power and the noise variance directly where noise variance is a measure of the statistical dispersion of the noise magnitude of a received signal; the other is to obtain the signal power estimate and the noise variance estimate. SNR estimation is as [22]:

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_v^2}\right) \tag{26}$$

Where σ_x^2 is the variance value of the signal and σ_v^2 is the variance value of the noise process.

In this section, we describe the method of estimating noise variance from the given segment of the noisy signal. In the presence of additive white Gaussian noise *v(n)*, the observed signal y(n) can be written as [15]:

$$y(n) = x(n) + v(n) \tag{27}$$

Where x *(n)* is the uncontaminated signal and v *(n)* is the zero-mean white noise of variance σ_v^2. The aim here is to estimate the noise variance σ_v^2, from the observed noisy signal *y (n)*. In order to solve this problem, we assume that the uncontaminated signal *x(n)* follows the *p*-th-order AR model (Equation 5). AR parameters *ai* satisfy the following set of Yule-Walker equations [23]:

$$\sum_{k=1}^{p} a_k R_x(|i-k|) = -R_x(i) \quad , \quad i > 0 \tag{28}$$

Where $R_x(i)$ are the autocorrelation coefficients of the uncontaminated signal *x (n)* . Since the additive noise *v(n)* is white, the autocorrelation coefficients $R_x(i)$ of the uncontaminated signal *x(n)* are related to the autocorrelation coefficients $R_y(i)$ of the noisy signal *y(n)* as follows [23]:

$$R_x(0) = R_y(0) - \sigma_v^2 \tag{29}$$

and

$$R_x(i) = R_y(i) \tag{30}$$

We have three-step procedure for estimating the noise variance σ_v^2. These steps are outlined below [22]:

Step *1:* From the observed (noisy) signal *y(n)* , compute the unbiased estimates of the autocorrelation coefficients $R_y(i)$, i = 0,1.., p + *q*. where *q* > p.

Step 02: Compute the least-squares estimate of AR coefficients by using the Cadzow's method [24] from the q(>p) high-order Yule-Walker equations [i=p+1, p+2,...,p+q].

Step 3: Use the AR coefficients obtained from Step 2 and compute the least-squares estimate of the noise variance from the over-determined set of p low-order Yule-Walker equations [i = 1, 2, . . 1 , p]. This is given by [24]:

$$\hat{\sigma}_v^2 = \left[\sum_{k=1}^{p} a_i \left\{\hat{R}_x \mathrm{x}(i) + \sum_{k=1}^{p} a_k \hat{R}_y(|i-k|)\right\}\right] \Big/ \sum_{i=1}^{p} a_i^2 \tag{31}$$

After noise variance estimation $\hat{\sigma}_v^2$, we estimate the variance value of the signal $\hat{\sigma}_x^2$ and calculate the noised signal variance $\hat{\sigma}_y^2$ (Speaker signal + AWGN):

$$\hat{\sigma}_x^2 = \hat{\sigma}_y^2 - \hat{\sigma}_v^2 \tag{32}$$

From equation (27) we have:

$$\mathrm{SNR}_{dB} = 10\log_{10}(\hat{\sigma}_x^2) - 10\log_{10}(\hat{\sigma}_v^2) \tag{33}$$

F. Threshold Adaptation and Decision

SAD algorithm calculates EZRs of all frames (for speaker's signal) and estimate threshold by:

$$Threshold = \min(EZRs) + \alpha * [DELTA] \tag{34}$$

$$DELTA = \max(EZRs) - \min(EZRs) \tag{35}$$

α: is a real number in the interval of]0,1[. The value of α depends on noise level (SNR). To estimate "α" value, a work is done. The results are reported in Table 1. These results, show that to have high identification accuracy we increase the value of α as SNR level increase. Threshold calculation procedure can be represented by the Fig.6.

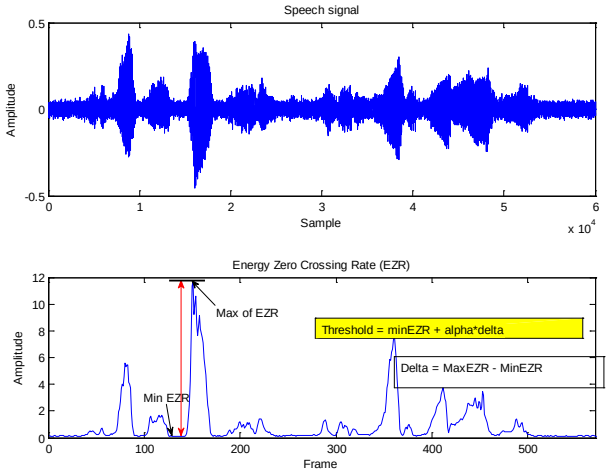

Fig.6. Illustration of threshold calculation (SNR= 10dB, α=0.2)

G. Implementation

The method is described by the diagram shown in fig.7.

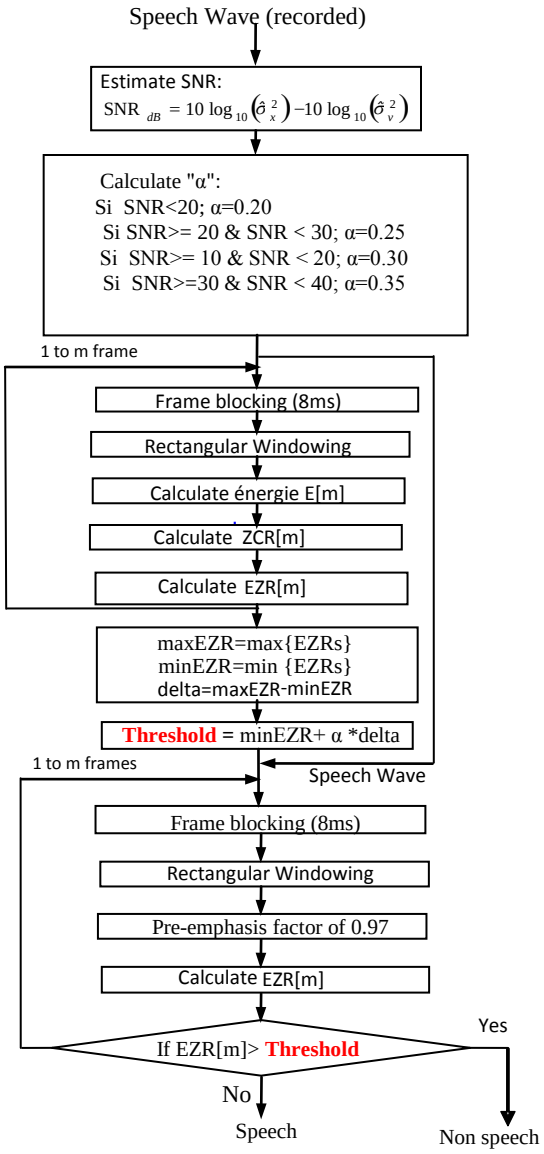

Fig.7. Block diagram for the SAD algorithm method.

IV. RESULTS AND DISCUSSION

The evaluation of the proposed feature extraction method was performed by text-independent closed-set speaker identification experiments on the TIMIT database.

TIMIT database allows identification to be done under almost ideal conditions. The TIMIT database consists of 630 speakers, 70 % male and 30 % female from 10 different dialect regions in America. Each speaker has approximately 30 seconds of speech spread over ten utterances. The speech was recorded using a high quality microphone in a sound proof booth at sampling frequency of 16 kHz, with no session interval between recordings.

We have chosen 200 speakers from different regions and defined the first 4 utterances for each speaker as the training set and 1 utterance as the test set. The TIMIT database files are sampled with a rate of 16000 samples/s, these files were downsampled to a rate of 8000 samples/s. The speech signal is segmented into frames. Processing was performed using Hamming windowed frames of 20 ms; it takes 160 samples overlapping by 50% (10ms) of 80 samples.

In our work, the number of filters used is 24 of triangular shape (as shown a above in fig.4) . Each frame is converted to 32 MFCCs and 32 Autoregressive coefficients. These features (AR-MFCC) are used to train the GMM. The GMM forms the basis for both the training and classification processes. We fixed the number of Gaussian mixture at G=64 and a maximum number of iteration by 100 to model the speaker's voice sample.

A. Demonstrate the performance of SAD algorithm

We pass the speech signal through the SAD algorithm (α = 0.45). The fig.8 represents the speech signal before and after processed through speech activity detection algorithm we can observe the efficiency of the SAD algorithm where silent segments are eliminated. The fig.9 illustrates a speech signal (clean speech) and its speech activity contour. Fig.10, fig.11 and fig.12, represent speech signal and their speech activity contour as function of SNR for 10dB, 5dB and 0dB respectively. These figures indicate that the SAD algorithm is robust down to SNR=5 dB

B. Degradation of speaker identification over Channel

To show channel degradation effect on our proposed speaker recognition system, we use original speech and reconstructed wave files after transmission over AWGN channel. The results are reported in Table 2, where we observe performance degradation of speaker identification accuracy when using reconstructed files after transmission.

C. Performance comparison of Proposed Method AR - MFCC with MFCC and ΔMFCC

We compare the proposed AR-MFCC with MFCC and ΔMFCC features in noisy conditions (Additive White Gaussian Noise). These results are reported in figure 13. From these results, it can be seen that the proposed AR-MFCC features provide good improvements of speaker identification

in comparison with MFCC and ΔMFCC (first derivative of MFCC) over AWGN channel. As a second test we compare AR-MFCC and MFCC in terms of runtime. Table 3 shows simulation results in terms of runtime, where we can observe that AR-MFCC is time consuming more than MFCC, where we have used a Laptop that is: Intel (R) core (TM) i5-3210M CPU @ 2.5GHZ 2.50GHZ.

Otherwise, we have evaluated our system in presence of different kind of noise like Pink, Blue, Red and violet noise but not over communication channel. Pink noise is used for replacing ambient noise in sound-related experiments. It is also used in theaters and studios where the human ears must evaluate the quality of sound [25]. We have added different kind of noise to speech signals of TIMIT database (we used 300 speaker signals from TIMIT database). All results are reported in Table 4. Table 4 shows by comparing different feature extraction techniques (AR-MFCC, MFCC, ΔMFCC) that AR-MFCC has higher performance for speaker identification rate.

D. Threshold Calculation for 20 speech signals

We choose 20 speech signals from TIMIT database, we passed every speech signal over SAD algorithm that calculates the threshold. Figure 14 shows threshold calculation of each speech signal to detect Speech/non-Speech.

TABLE I. "α" VALUE VERSUS SNR .

SNR dB	Alpha (α)	Identification rate %
50	0.25	86.00
	0.35	85.66
	0.45	**87.33**
	0.5	85.33
30	0.20	86.33
	0.25	87
	0.35	**79**
	0.4	77.33
	0.5	76.66
20	0.20	63.66
	0.25	64.33
	0.3	**67.67**
	0.4	57.66
	0.5	42.33
10	**0.25**	**60.33**
	0.3	59.67
	0.35	57.00
	0.4	54.33
	0.5	51.33

TABLE II. IDENTIFICATION RATE ACCURACY USING ORIGINAL SPEECH WAVEFORM AND RECONSTRUCTED SPEECH AFTER TRANSMISSION OVER AWGN CHANNEL

System	Speaker identification system	Speaker identification rate over AWGN Channel
Identification rate %	96	87

TABLE III. RUN TIME OF: MFCC AND AR_MFCC (FOR 100 SPEAKERS).

Features	MFCC	AR_MFCC
Elapsed time [sec]	299.091383	433.50024

TABLE IV. IDENTIFICATION RATE ACCURACY FOR AR-MFCC, MFCC and ΔMFCC IN PRESENCE OF DIFFERENT KIND OF NOISE: PINK, BLUE AND VIOLET NOISE (NOT OVER COMMUNICATION CHANNEL).

Noise	dB	Speaker recognition system %		
		AR MFCC	MFCC	ΔMFCC
Pinknoise	clean	99	85	90
	30	95	85	90
	20	95	75	83
	15	70	55	60
	10	55	30	41
	5	35	5	15
	0	15	7	10
Rednoise	30	99	75	90
	20	99	75	90
	15	99	75	90
	10	99	75	90
	5	99	75	90
	0	95	70	85
Bluenoise	30	88	75	80
	20	45	30	32
	15	10	15	18
	10	6	5	5
	5	5	5	5
	0	5	5	5
violetnoise	30	75	75	70
	20	40	30	45
	15	15	10	15
	10	9	5	5
	5	5	5	5
	0	5	5	5

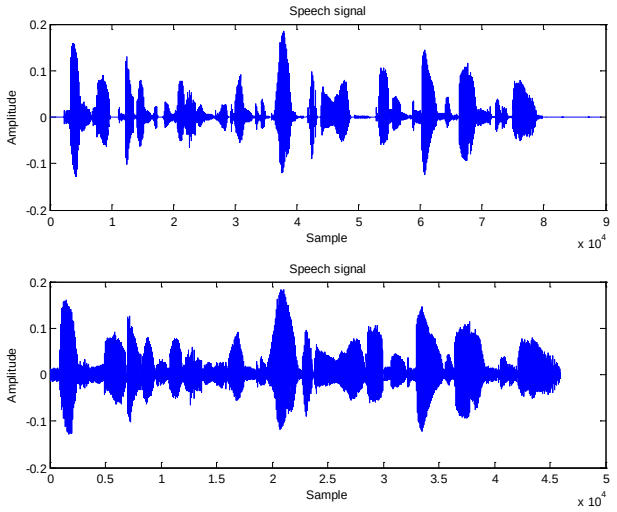

Fig.8 Original speech signal before and after processed through speech activity detection algorithm

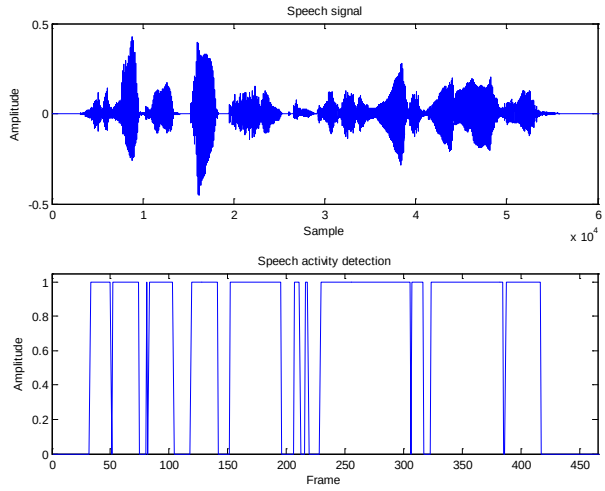

Fig.9 Clean speech signal and its speech activity contour using SAD algorithm

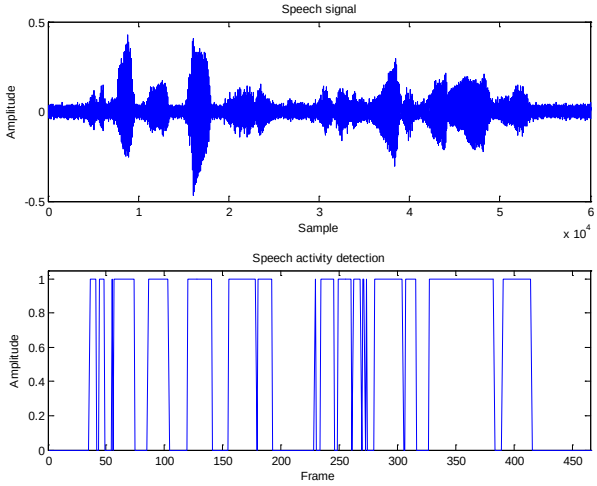

Fig.10. Speech signal and its speech activity contour at SNR= 10dB

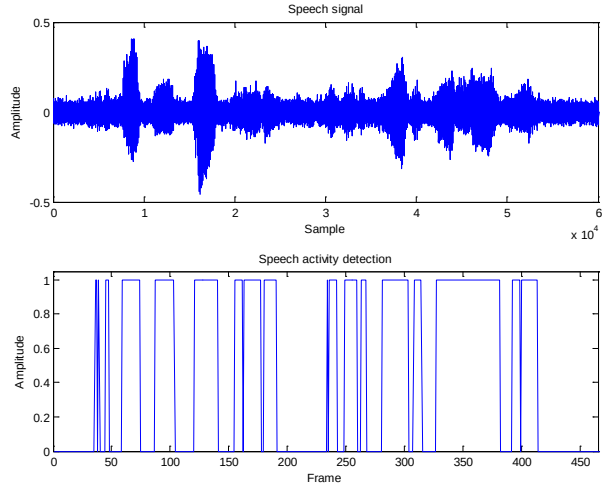

Fig.11 Speech signal and its speech activity contour at SNR= 5dB

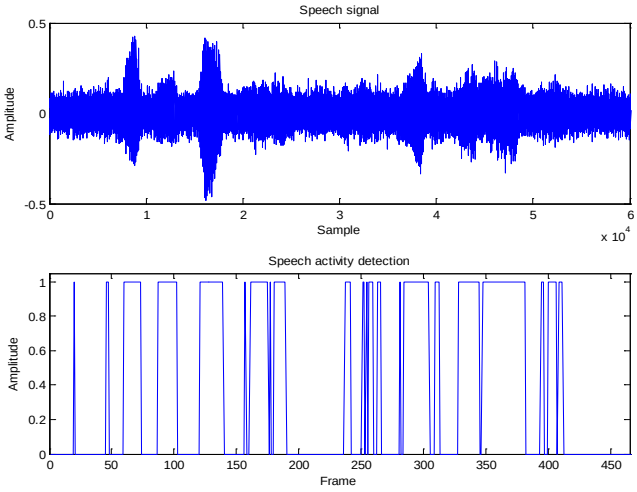

Fig.12. Speech signal and its speech activity contour at SNR= 0dB

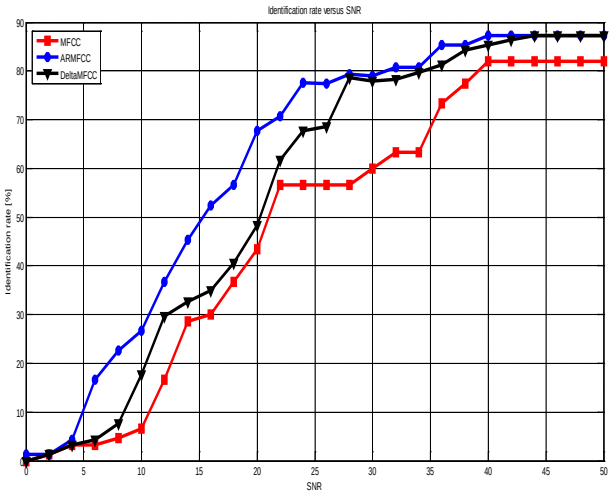

Fig.13 Identification rate accuracy of AR-MFCC, MFCC and ΔMFCC versus SNR over AWGN channel.

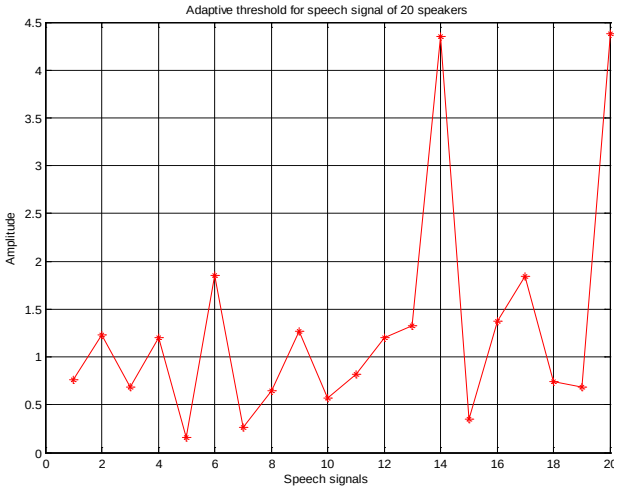

Fig.14. Threshold for speech signal of 20 speakers from TIMIT database under SNR=35dB

V. CONCLUSION

In this paper, we have tried to provide a robust features extraction based on MFCC and AR modeling approach (AR-

MFCC) to enhance the performance of an Automatic Speaker Recognition System over AWGN channel in noisy environment. We have also proposed an efficient SAD algorithm based on knowledge of the noise level as the first step of the algorithm and Zero Crossing Rate and Energy Measurements as second step. Our SAD based on efficient threshold for speech/non-speech determination. SAD algorithm showed a high performance of speech /non-speech discrimination in noisy environments which improves memory capacity and identification rate accuracy. The SAD depends on a factor "α" , where to have a high identification rate accuracy we should increase α as SNR level increase. A comparison study of MFCC (32 coefficients), ΔMFCC (32) and AR-MFCC (64) was made in view of their effects on the identification system performance of remote automatic speaker in noisy environment.

We have evaluated our system in presence of different kind of noise like Pink, Blue, Red and violet noise but without communication channel. Where the maximum identification rate of 99% was found for AR-MFCC in the presence of different kind of noise.

The results of experiments indicate that the performance of the speaker identification system is improved for combined AR and MFCC feature. However, in term of runtime, AR-MFCC requires more time to execute than MFCC. SAD performs suitable contour of speech activity. Furthermore, it works accurately in low SNR environments (down to SNR=5 dB) and leads to a good identification rate accuracy.

Our system may be very effective by decreasing the run time of AR-MFCC. The performance of this system can also be improved by improving the noise removing technique of the speech signal.

REFERENCES

[1] AL-SAWALMEH, Wael, DAQROUQ, Khaled, AL-QAWASMI, Abdel-Rahman, et al. The use of wavelets in speaker feature tracking identification system using neural network. WSEAS Transactions on Signal Processing, 2009, vol. 5, no 5, p. 167-177.

[2] NIDHYANANTHAN, S. SELVA et KUMARI, R. SHANTHA SELVA. Language and Text-Independent Speaker Identification System Using GMM. *Wseas Trans. Signal Process*, 2013, vol. 4, p. 185-194.

[3] NARAYANA, M. Laxmi et KOPPARAPU, Sunil Kumar. Effect of noise-in-speech on MFCC parameters. In : *Proceedings of the 9th* WSEAS international conference on signal, speech and image processing, and 9th WSEAS international conference on Multimedia, internet & video technologies. World Scientific and Engineering Academy and Society (WSEAS), 2009. p. 39-43.

[4] GOH, Yeh Huann, RAVEENDRAN, Paramesran, et JAMUAR, Sudhanshu Shekhar. Robust speech recognition using harmonic features. *Signal Processing, IET*, 2014, vol. 8, no 2, p. 167-175.

[5] Zhao, Xiaojia et Wang, DeLiang. Analyzing noise robustness of MFCC and GFCC features in speaker identification. In : Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013. p. 7204-7208.

[6] El Ayadi, M. (2008). Autoregressive models for text independent speaker identification in noisy environments (Doctoral dissertation, University of Waterloo).

[7] Delima, Charles B., Alcaim, Abraham, et Apolinario JR, J. A. GMM Versus AR-Vector Models for text independent speaker verification. In : Proc. of SBT/IEEE International Telecommunication Symposium (ITS 2002), Brazil. 2002.

[8] Peinado, Antonio et Segura, Jose. Speech recognition over digital channels: robustness and standards. John Wiley & Sons, 2006.

[9] Riadh Ajgou, Salim Sbaa, Said Ghendir,Ali Chamsa and A. Taleb-ahmed. Effects of speech codecs on a remote speaker recognition system using a new SAD. Proceedings of the 2014 International Conference on Systems, Control, Signal Processing and Informatics II (SCSI '14). Prague, Czech Republic April 2-4, 2014.

[10] Tan, W. C., Jaafar, H., Ramli, D. A., Rosdi, B. A., & Shahrudin, S. Intelligent frog species identification on android operating system. *International journal of circuits, systems and signal processing.* Volume 8, p137-148. 2014.

[11] Seng, G. H., & Swee, T. T. Spectral Coefficients System for Osteoarthritis Detection. *International journal of circuits, systems and signal processing.* Issue 3, Volume 7, 2013.

[12] Chavan, Mahesh S. "Speaker Identification in Mismatch Condition using Warped Filter Bank Features.". *International journal of circuits, systems and signal processing*, vol. 9, no 4, p. 88-93. 2015.

[13] Gopi, E. S. (2014). *Digital Speech Processing Using Matlab*. Imprint: Springer.

[14] KUSUMOPUTRO, B., & BUONO, A. (2012). Identification of Noisy Speech Signals using Bispectrum-based 2DMFCC and Its Optimization through Genetic Algorithm as a Feature Extraction Subsystem. *WSEAS Transactions on Computers*, *11*(8).

[15] Tang, Yinggan. "Parameter estimation of Wiener model using differential evolution algorithm." *International Journal of Circuits, Systems and Signal Processing* 5: 315-323. (2012).

[16] 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..

[17] Cherifa, Snani et Messaoud, Ramdani. New technique to use the GMM in speaker recognition system (SRS). In : Computer Applications Technology (ICCAT), 2013 International Conference on. IEEE, 2013. p. 1-5.

[18] HARSHA, B. V. A noise robust speech activity detection algorithm. In : Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on. IEEE, 2004. p. 322-325.

[19] Kotnik, Bojan, Hoge Bojan, Hoge, Harald, et Kacic, Zdravko. Evaluation of pitch detection algorithms in adverse conditions. In : Proc. 3rd international conference on speech prosody. 2006. p. 149-152.

[20] Ajgou, Riadh, Sbaa, Salim, Ghendir, Said, *et al.* Novel Detection Algorithm of Speech Activity and the impact of Speech Codecs on Remote Speaker Recognition System. *WSEAS Transactions on Signal Processing*, vol. 10. 2014

[21] SOON, Ing Yann, KOH, Soo Ngee, YEO, C. K., *et al.* Transformation of narrowband speech into wideband speech with aid of zero crossings rate. *Electronics Letters*, vol. 38, no 24, p. 1607-1608. 2002

[22] DE DIEULEVEULT, François et ROMAIN, Olivier. Électronique appliquée aux hautes fréquences-2ème édition-Principes et applications: Principes et applications. Dunod, 2008.

[23] PALIWAL, K. K. Estimation of noise variance from the noisy AR signal and its application in speech enhancement. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no 2, p. 292-294. 1988,

[24] J. A. Cadzow, "Spectral estimation: An overdetermined rational model equation approach," Proc. *IEEE*, vol. 70, pp. 907-939, Sept. 1982.

[25] Kyon, D. H., Lee, W. H., Kim, M. S., & Bae, M. J. Hi-pass Pink Noise and Standard Volume for Auditory Experiments. 2013.