An Efficient Approach for MFCC Feature Extraction for Text Independant Speaker Identification System

R. Ajgou, S. Sbaa, S. Ghendir, A. Chemsa And A. Taleb-Ahmed

***Abstract*—**This paper presents an efficient noise-robust feature extraction method for remote speaker identification system. Mel frequency cepstral coefficients (MFCCs) are the most widely used front ends in the state of the art speaker identification systems. One of the major problem with MFCCs is that it deteriorates in the presence of noise. To overcome this problem, we have propsed an efficient feature extraction technique based on the combination between the MFCC and parameters of too pole filter parameter (Autoregressive model parameters) that characterize the human vocal tract. The system employs a robust speech feature based on MFCCAR modeled by GMM. An effective speech enhancement methods is essential for speaker recognition, an overview of some recent speech enhancement techniques of the state of the art have been presented where we have investigated its effects on our speaker identification system accuracy based on MFCCAR. TIMIT database with speech signals from 200 speakers has been used in Matlab simulation. The first four utterances for each speaker could be defined as the training set while 1 utterance as the test set. Experimental results show that proposed methods achieve better performance. The use of MFCCAR approach has provided significant improvements in identification rate accuracy when compared with MFCC, deltaMFCC and PLP in noisy environment. However, with regard to runtime, MFCCAR requires more time to execute. In terms of effects of reverberant speech enhancement methods, it is shown a significant improvement for Tracking of noise algorithm method.

***Keywords*—** speaker recognition; MFCC; AR, MFCCAR,

I. INTRODUCTION

Speaker modeling and feature extraction are the main part of a speaker recognition system. The Gaussian mixture model (GMM) is the most common approach for speaker modeling in text-independent speaker recognition [1]. It it is important to extract features from the speech signal which capture the speaker specific characteristics [2]. The most important features extraction is Mel Frequency Cepstral Coefficients (MFCC) [3]. However, performance using MFCC features deteriorates in the presence of noise [4]. A more direct and conceptually simpler way to characterize speaker differences would be to look at the differential parameters that characterize the speech production apparatus [5]. The vocal tract filter is modeled as an all-pole filter system. Auto-Regressive (AR) Vector Models is a signficant subject of interest in the field of Speaker Recognition [6]. Whereas the idea of modeling a speaker by an AR-vector model estimated on sequences of speech frames is common to these works, the way to measure the similarity between two speaker models is addressed very differently. The use of AR-vector model is often motivated by the belief that such an approach is an eficient way to extract dynamic speaker characteristics. In this sense, in this work we have combined MFCC parameters with AR-model vector to have noise robust Text-Independent Speaker Identification.

Speech/non-speech detection is an important task in speech processing. Through the last decade, numerous researchers have developed different strategies for detecting speech on a noisy signal and the influence of the speech activity detection (SAD) effectiveness on the performance of speech processing systems[7]-[8]. In this paper, we have used the SAD algorithm developped by [9] to improve the identification accuracy.

Speech signal can be corrupted by noise in various situations, such as trains, cars, airport, babble, factory, street..etc. As a result, robustness to environmental noise is essential to most practical applications of speaker identification systems. Speech degradations as imposed by various telephone networks and internet have been proven to have large effects on the performance of the automated speaker recognition systems [10]. Speech enhancement improves the quality and intelligibility of voice communication for a range of applications including speaker recognition system. Thus, in this paper we present an overview of single-channel speech enhancement methods. Although many significant techniques have been introduced over the past decade because there are many areas where it is necessary to enhance the quality of speech that has been degraded by background noise. Hence, this paper evaluated the effects of seven famous speech enhancement algorithms of the state of the art on our system of text-independent speaker identification

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)

performance that are based on MFCC and AR-vector.

In our work we suggest to provide a comprehensive assessment of speech extraction based on MFCC and Autoregressive model using the GMM model in remote speaker recognition system over communication channel (Rayleigh). Our proposal speaker identification system on the remote communication channel is described in more detail in the next section.

II. Remote Speaker Identification and Proposed Feature Extraction (MFCCAR)

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 the SAD is used to detect speech/non-speech zone and feature extraction where feature extracting is accomplished by the combination of MFCC and AR Vectors. The system we used include a remote text independent speaker recognition system which was established according to the following diagram in Fig. 1.

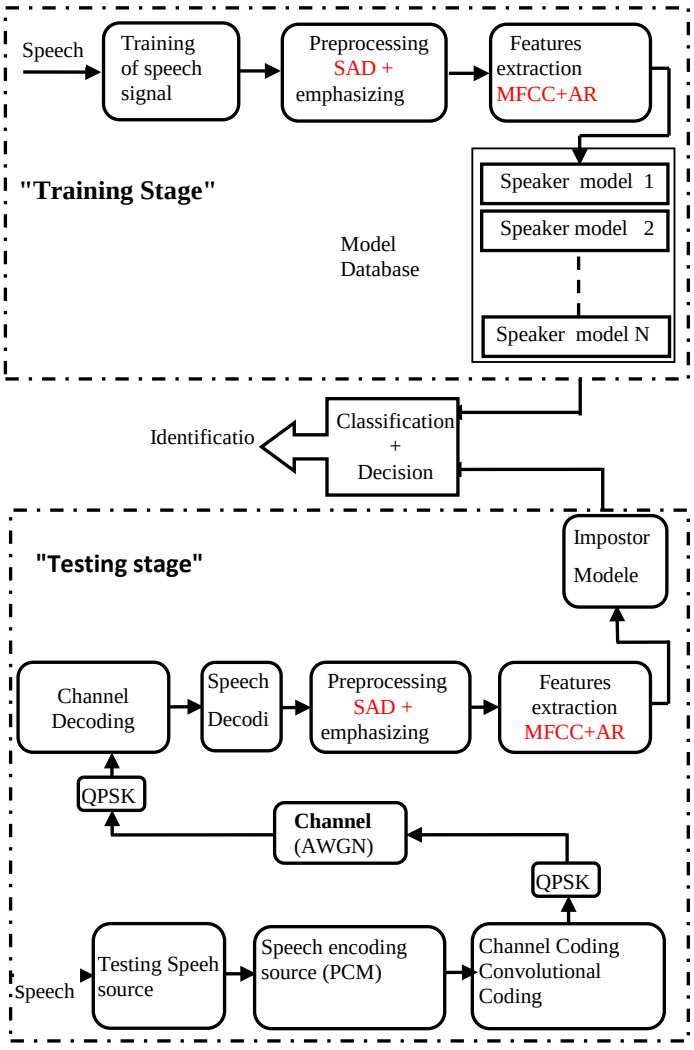

Fig 1. Block diagram of the proposed system.

A. Preprocessing stage

In this step, silence portions are removed from the speech signals by using the speech activity detection algorithm as in [10]. Then, each utterance is pre-emphasized with a pre-emphasis factor of 0.97 which leads to improve identification rate accuracy. 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]:

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

B. Proposed features extraction (MFCCAR)

We have combined MFCC features with autoregressive model coefficients (AR vectors). The number of coefficients is 64 (32 MFCC and 32 AR) extracted from each frame. The acoustic signal contains different kinds of information about the speaker. Although the proposed feature (MFCCAR) is performed using the following steps:

- MFCC feature:

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. 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 consisting of bank of filters. Each filter is of triangular in shape. The triangular filter banks in mel scale are uniformly spaced [12].

In our work, speech signal 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 [12]

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

f_{low} and f_{high} are the low and high frequency boundaries of filter bank, they are given as [12]:

$$f_{low} = \frac{f_s}{N} \tag{3}$$

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

$$f_{high} = \frac{f_s}{2} \tag{4}$$

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

Also, the number of filters used is 24. The Fig.2 illustrates

the bank of filters in mel-frequency scale. (1 in x-axis corresponds to Fs/2).

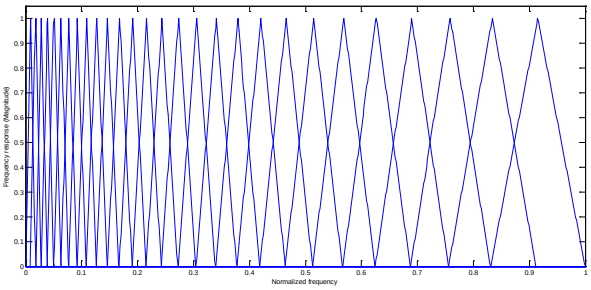

Fig.2. Bank of filters in mel-frequency scale. 1 in x-axis corresponds to Fs/2: (8000 Hz) [12].

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.

For a given input speech sequence $s(n)$:

$$s = \begin{bmatrix} s_0 & s_1 & ..... & s_{N-1} \end{bmatrix}^T \quad (5)$$

where: N is speech signal sample.

The mel-cestral coefficients for the whole speech signal described as:

$$c = \begin{bmatrix} c_0 & c_1 & ..... & c_M \end{bmatrix}^T \quad (6)$$

where: M is frame number, $c_0 \quad c_1 \quad ..... \quad c_M$ represents MFCC for frames: 0 1…….M. the c_0 represents MFCCs for frame number 0 which described as:

$$c_0 = \begin{bmatrix} cc_{00} & cc_{01} & ..... & cc_{0L} \end{bmatrix}^T \quad (7)$$

where: L is mel-frequency cepstral coefficients to be considered (in our work we have considered L=32). From Eq.6 and Eq.7 we will have MFCCs of a speaker signal:

$$c = \begin{bmatrix} cc_{00} & cc_{10} & cc_{20} & .....cc_{M0} \\ cc_{01} & cc_{11} & cc_{21} & .....cc_{M1} \\ & & & \\ cc_{0L} & cc_{1L} & cc_{2L} & .....cc_{ML} \end{bmatrix} \quad (8)$$

- Autoregressive Vectors (AR vectors):

The vocal tract is usually modeled as a concatenation of nonuniform lossless tubes of varying cross-sectional area that begins at the vocal cords and ends at the lips. The vocal tract is excited by a broad-band noise source during the production of unvoiced sounds. Plosive sounds result from building up air pressure in the mouth and abruptly releasing it [13].

A linear model of speech production was developed by Fant in the late 1950s [14], where the glottal pulse, vocal tract, and radiation are individually modeled as linear filters. The source is either a quasi-periodic impulse sequence for the voiced sounds or a random noise sequence for unvoiced sounds with a gain factor G set to control the intensity of the excitation. The transfer function V(z) for the vocal tract relates volume velocity at the source to volume velocity at the lips. It is generally an all-pole model for most speech sounds [13]. Each pole of V(z) corresponds to a formant or resonance of the sound. For nasals and fricatives that require both resonances and anti-resonances (poles and zeros), an all-pole model is still preferred because the effect of a zero in the transfer function can be achieved by including more poles [13]. The radiation model *R(z)* describes the air pressure at the lips, which can be reasonably approximated by a first-order backward difference. Combining the glottal pulse, vocal tract, and radiation yields a single all-pole transfer function [13]-[15] given by :

$$H(z) = G(z)V(z)R(z) = \frac{G}{A(Z)} = \frac{G}{1 + a_1 Z^{-1} + a_2 Z^{-2} + ...a_p Z^{-p}} \quad (9)$$

where *p is* the order of prediction coefficients (prediction coefficients are $\{a_1, a_2, \ldots a_P\}$), with this transfer function, we get a difference equation for synthesizing the speech samples *s(n)* as [13]:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \quad (10)$$

It can be noted that s(n) is predicted as a linear combination of the previous samples. Therefore, the speech production model is often called the linear prediction (LP) model, or the autoregressive model (AR).

In practice, the predictor coeflcients{ai}describing the autoregressive model must be computed from the speech signal. Since speech is time-varying in that the vocal-tract configuration changes over time, an accurate set of predictor coefficients is adaptively determined over short intervals (typically 10 ms to 30 ms) called frames, during which time-invariance is assumed. The gain G is usually ignored (G=1) to allow the parameterization to be independent of the signal intensity. The autocorrelation method and the covariance method are two standard methods of solving for the predictor coefficients [16]. 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, we limited the number of coefficients to 32 (*p*=32*)*. The calculation of {ai} has been carried out by the Yule-Walker method this method solves the Yule-Walker equations by means of the Levinson Durbin

recursion [15].

For a given input speech sequence $s(n)$ (Eq. 5)., AR coefficients for the whole speech signal described as:

$$A = \begin{bmatrix} A_0 & A_1 & ..... & A_M \end{bmatrix}^T \quad (11)$$

where: M is frame number, $A_0 \quad A_1 \quad ..... \quad A_M$ represents AR vectors for frames: 0 1…….M.

the A_0 AR vector for frame number 0 which described as:

$$A_0 = \begin{bmatrix} AA_{00} & AA_{01} & ..... & AA_{0p} \end{bmatrix}^T \quad (12)$$

where: in our work we have considered p=32.

From Eq.11 and Eq.11 we will have AR vectors that characterize speech signal of a speaker:

$$A = \begin{bmatrix} AA_{00} & AA_{10} & AA_{20} & ..... & AA_{M0} \\ AA_{01} & AA_{11} & AA_{21} & ..... & AA_{M1} \\ & & & & \\ AA_{0L} & AA_{1L} & AA_{2L} & ..... & AA_{MP} \end{bmatrix} \quad (13)$$

From Eq.8 and Eq.13 we set MFCCAR feature that describe the combination between MFCCs and AR vectors (L=p=32):

$$MFCCAR = \begin{bmatrix} cc_{00} & cc_{10} & cc_{20} & .....cc_{M0} & AA_{00} & AA_{10} & AA_{20} & .....AA_{M0} \\ cc_{01} & cc_{11} & cc_{21} & .....cc_{M1} & AA_{01} & AA_{11} & AA_{21} & .....AA_{M1} \\ & & & & & & & \\ cc_{0L} & cc_{1L} & cc_{2L} & .....cc_{ML} & AA_{0L} & AA_{1L} & AA_{2L} & .....AA_{MP} \end{bmatrix} \quad (14)$$

The Fig. 3 presents the MFCCAR extraction procedure for a speech signal consists of five frames that represents speech signal, where we extract MFCC and AR vector from each frame and reconstruct one matrix of MFCCAR as shown in this figure.

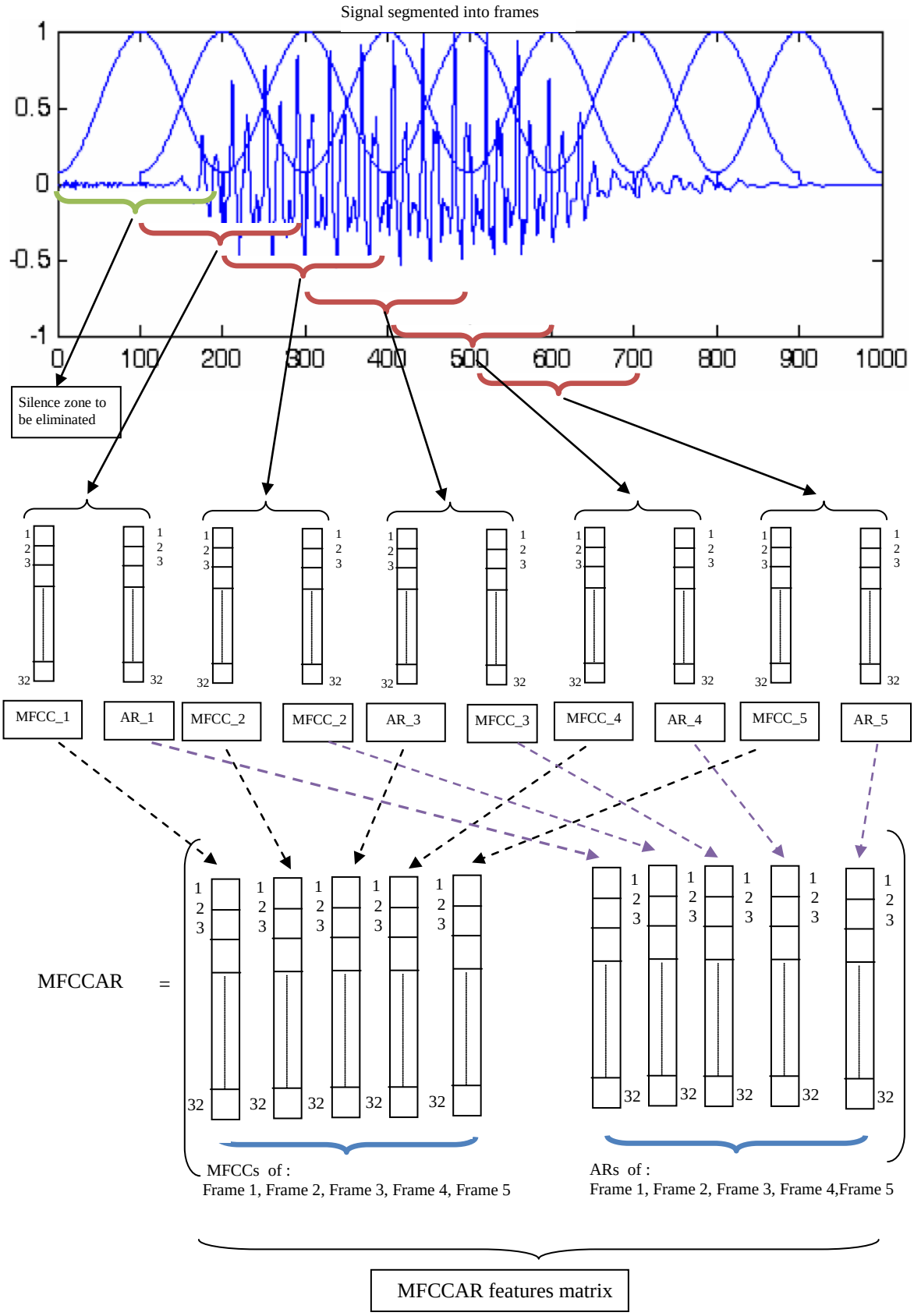

Fig.3. MFCCAR extraction procedure.

C. Modeling using GMM

In the training stage, pattern generation is the process of generating speaker specific models with collected data. Gaussian mixture models are commonly used for classification of varying length patterns represented as sets of feature vectors.Maximum likelihood (ML) based method is commonly used for estimation of parameters of a GMM for each class [1].

The system (remote identification system according to figure 1) was trained using speakers from the TIMIT database [17] 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), so that sixty-four coefficients are extracted (32 mel-frequency cepstral coefficients and 32 LP coefficients (autoregressive model)) and models characterization using GMM are formed.

D. Testing phase

Speech signal is coded using PCM code. A convolutional code [15], 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 transmitted through the transmission 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 and reconstruct MFCCAR.

E. Decision 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 [1]. Decision phase performed using GMM with maximum likelihood (ML) where 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 [1].

The speaker identification rate is given by:

$$Id = \frac{Number\ \ of\ \ utterance\ \ correctly\, identified}{Total\ \ number\ \ of\ \ utternace\ \ under\ \ test} \times 100\% \quad (15)$$

III. Various speech Enhancement Methods

Overview of seven methods of the state of the art has been evaluated in terms of robustness against real noise (babble, airport, car and street) where speech signals chosen from NOIZEUS noisy speech corpus developed in Hu and Loizou [18] that is suitable for evaluation of speech enhancement algorithms. We evaluate their performance by Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862) [19] before the study of their effects on our speaker identification accuracy. Besides, our aim is a study of the effects of speech enhancement algorithms on our text-independent speaker Identification performance based on MFCCAR. The methods that we have evaluated are:

- Tracking Of Non-Stationary Noise Based On Data-Driven Recursive Noise Power Estimation, this method was proposed by Erkelens and Heusdens, 2008 [19].
- Speech Enhancement Based On A Priori Signal To Noise Estimation. This method was proposed by Scalart, and Vieira, 1996 [20].
- Geometric approach to spectral Geometric Approach. This recent method was proposed by Yang Lu and al., 2008 [21].
- Harmonic Regeneration Noise Reduction (HRNR). This method was proposed by Plapous, and al., 2006 [22].
- Phase Spectrum Compensation (PSC). This work was proposed by Anthony and al., 2008 [23].
- Speech Enhancement Using a Non causal A Priori SNR Estimator. This technique was proposed by I. Cohen, 2004 [24].
- Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay. This technique was proposed by Gerkmann and Richard 2012 [25].

IV. Experiments, 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. The TIMIT database contains speech from 630 speakers. We 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 20ms, it takes 160 samples overlapping by 50% (10ms) of 80 samples. From each frame, 32 coefficients MFCC and 32 coefficients AR were calculated and 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 mixture in the beginning of training stage to model the features extracted from each speaker's voice sample.

A. Proposed MFCCAR versus MFCC, delta-MFCC and PLP.

At first, before features extraction, speech signals should passe through the SAD algorithm proposed in [9], detection of speech/non-speech of this algorithm depends on parameter α which depends on SNR level (α: is a real number in the interval of]0,1[ ). To have high identification accuracy, we should increase the value of α as noise level increase.

We compare the proposed MFCCAR with MFCC, ΔMFCC and perceptual linear predictive (PLP) [26] features in noisy conditions (Additive White Gaussian Noise). These results are reported in Fig. 4. From these results, it can be seen that the

proposed MFCCAR features provide good improvements of speaker identification in comparison with MFCC, ΔMFCC and PLP over AWGN channel.

We compare MFCCAR with MFCC, ΔMFCC and PLP in terms of runtime. Table 1 shows simulation results in terms of runtime, we can observe that MFCCAR consuming more time than MFCC (we have used a Laptop that is: Intel ®(TM) i5-3210M CPU @ 2.5GHZ 2.50 GHZ).

Otherwise, we have evaluated our speaker identification system based on MFCCAR in presence of different kind of noise (additive noise) like: WGN (White Gaussian Noise), Pink, Blue, 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. We have added different kind of noise to speech signals of TIMIT database (we used 200 speaker signals from TIMIT database). All results are reported in Table 2. Table 2 represents identification rate accuracy using MFCCAR, MFCC, ΔMFCC and PLP in presence of: WGN, Pink, Blue, and violet noise. Table 2 shows that MFCCAR provided a higher speaker identification rate. Otherwise, Table 3 represents the average of Identification rate accuracy for AR-MFCC, MFCC, ΔMFCC and PLP in presence of different kind of noise: WGN, pink, blue and violet noise (not over communication channel) where MFCCAR provided a higher average of speaker identification rate.

B. Effects of speech enhancement method on speaker identification accuracy

Our goal is to make our system more robust to noise by choosing speech enhancement technique. Hence, we start by a comparative study of speech enhancement methods. Table 4 represents the Average PESQ scores of different methods for speech contaminated by babble, airport, car, street and restaurant noise, speech signals chosen from NOIZEUS noisy speech corpus developed in Hu and Loizou laboratory [18]. From this table 5, we can conclude that the method of tracking of Non-Stationary Noise Based On Data-Driven Recursive Noise Power Estimation [19] provided the best Average PESQ. We study the effect of seven speech enhancement methods mentioned earlier (section 3) on our speaker identification system based on MFCCAR. The results of the Identification rate accuracy and the average using speech enhancement methods in presence of White Gaussian Noise reported in Table 6 where we have chosen 200 speakers. From table 5, the method of tracking of Non-Stationary Noise Based On Data-Driven Recursive Noise Power Estimation [19], provided the good Identification rate accuracy.

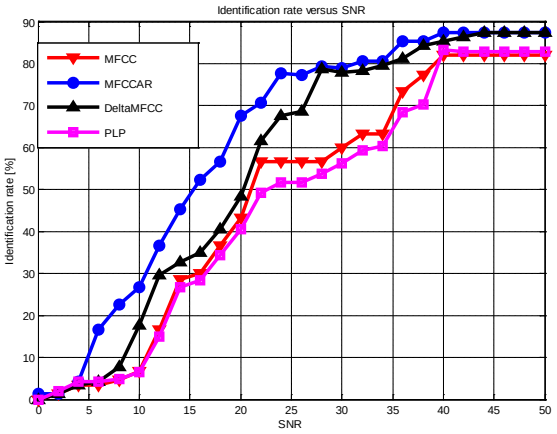

Fig 4. Identification rate accuracy over communication channel of: MFCCAR, MFCC MFCC and PLP versus SNR

	MFCCAR	MFCC	ΔMFCC	PLP
Elapsed time [sec] %	685.9909	384.8797	518.3525	322.1164

Table 1. Run time of: MFCCAR, MFCC, ΔMFCC and PLP (for 100 speakers).

Noise	SNR [dB]	Identification rate %			
		MFCCAR	MFCC	ΔMFCC	PLP
WGN	clean	99	85	90	80
	30	93	73	80	71
	20	92	70	80	68
	15	65	50	55	52
	10	50	27	35	30
	5	29	5	15	7
	0	10	5	5	5
Rose	30	95	85	90	78
	20	95	75	83	66
	15	70	55	60	50
	10	55	30	41	21
	5	35	5	15	5
	0	15	7	10	7
Blue	30	88	75	80	71
	20	45	30	32	30
	15	10	15	18	10
	10	6	5	5	5
	5	5	5	5	5
	0	5	5	5	5
violet	30	75	75	70	69
	20	40	30	45	26
	15	15	10	15	7
	10	9	5	5	5
	5	5	5	5	5
	0	5	5	5	5

Table 2. Identification rate accuracy for AR-MFCC, MFCC, ΔMFCC and PLP in presence of different kind of additive noise: WGN, pink, blue and violet noise (not over communication channel).

Noise	SNR [dB]	Average Identification rate %			
		MFCCAR	MFCC	ΔMFCC	PLP
WGN	clean 30 20 15 10 5 0	**62.57**	45.00	51.42	44.71
Rose	clean 30 20 15 10 5 0	**66.28**	48.85	55.57	43.57
Bleu	clean 30 20 15 10 5 0	**36.85**	31.42	33.71	29.14
violet	clean 30 20 15 10 5 0	**35.42**	30.85	33.75	27.85

Table 3: Average of Identification rate accuracy for MFCCAR, MFCC, ΔMFCC and PLP in presence of different kind of noise.

Method	SNR [dB]	Noise type				
		Babble	Airport	Car	Street	Restaurant
Erkelens, 2008	0 5 10 15	**2.3120**	**2.4704**	**2.3750**	**2.3979**	**2.2978**
Yang Lu, 2008	0 5 10 15	2.1820	2.2239	2.0707	2.1328	2.1908
Scalart, 1996	0 5 10 15	1.7774	1.8102	1.7617	1.9542	1.7062
Plapous , 2006.	0 5 10 15	1.7562	1.9336	2.0241	1.8911	1.7101
Anthony , 2008	0 5 10 15	1.1814	2.2256	2.0742	2.1366	2.1879
Gerkmann. 2012	0 5 10 15	1.7025	2.4206	2.2913	2.2700	2.2839
Cohen, 2004	0 5 10 15	2.0674	2.0812	2.1271	2.1164	2.0097

Table 4 Average PESQ scores of different methods for speech contaminated by babble, airport, car, street and restaurant noise

Method	SNR [dB]	Identification Rate [%]	IdenIdentification Average [%]
Erkelens, 2008	0 5 10 15 20 25 30 40	7 13 30 53 77 87 97 99	**57.87**
Yang Lu, 2008	0 5 10 15 20 25 30 40	7 20 30 47 70 73 77 82	50.75
Scalart, 1996	0 5 10 15 20 25 30 40	2 6 23 33 50 63 73 87	42.12
Plapous , 2006.	0 5 10 15 20 25 30 40	7 3 17 37 43 70 77 90	43.00
Anthony, 2008	0 5 10 15 20 25 30 40	5 7 10 13 43 60 77 99	39.25
Gerkmann.2012	0 5 10 15 20 25 30 40	7 13 37 53 63 83 97 98	56.37
Cohen, 2004	0 5 10 15 20 25 30 40	10 16 23 43 50 53 63 63	40.1250

Table 5. Effects of speech enhancement methods on Identification accuracy.

V. CONCLUSION

In this paper, we have tried to provide a robust feature extraction based on MFCC and AR vectors approach (MFCCAR) to enhance the performance of an Automatic

Speaker Recognition System over communication channel in noisy environment.

A comparison of MFCCAR (64 coefficients), MFCC (32), ΔMFCC (32) and PLP (32) for remote speaker identification in noisy environment done a best Identification accuracy for MFCCAR feature.

We have evaluated our Identification system based on MFCCAR and SAD [8] algorithm in presence of different kind of additive noise WGN, Pink, Blue and violet noise but without communication channel, where the maximum identification rate of 99% was found for MFCCAR. The results of experiments indicate that the performance of the speaker identification system is improved for MFCCAR feature. However, in term of runtime, MFCCAR requires more time to execute than the other methods of the state of the art (MFCC, delta-MFCC and PLP).

Also, in this paper we have done a comparative of seven speech enhancement methods, considering their effects on our remote text-independent speaker identification accuracy. The comparative showed the method of tracking of Non-Stationary Noise Based On Data-Driven Recursive Noise Power Estimation, proposed by Erkelens and R. Heusdens, 2008 [19] provided the good Identification rate accuracy, so we recommend this method to enhance speech signal.

Our system may be very strong if we decrease the run time of MFCCAR.The performance of this system can also be improved by improving the noise removing technique of the speech signal.

.

REFERENCES

[1] D.A. Reynolds, “Speaker Identification and Verification Using Gaussian Mixture Speaker Models,” Speech Commun., vol. 17, no. 1-2, Aug. 1995, pp. 91-108.

[2] W Togneri, Roberto et Pullella, Daniel. An overview of speaker identification: Accuracy and robustness issues. Circuits and Systems Magazine, IEEE, 2011, vol. 11, no 2, p. 23-61.

[3] Ahidullah, Md et Saha, Goutam. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication, 2012, vol. 54, no 4, p. 543-565

[4] 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.

[5] Naito, M., Deng, L., & Sagisaka, Y. (2002). Speaker clustering for speech recognition using vocal tract parameters. *Speech Communication*, *36*(3), 305-315.

[6] Ganapathy, S., Mallidi, S. H., & Hermansky, H. (2014). Robust feature extraction using modulation filtering of autoregressive models. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, *22*(8), 1285-1295.

[7] Chen, S. H., Guido, R. C., Truong, T. K., & Chang, Y. (2010). Improved voice activity detection algorithm using wavelet and support vector machine.*Computer Speech & Language*, *24*(3), 531-543.

[8] Marković, I., Jurić-Kavelj, S., & Petrović, I. (2013). Partial mutual information based input variable selection for supervised learning approaches to voice activity detection. *Applied soft computing*, *13*(11), 4383-4391.

[9] Riadh AJGOU, Salim SBAA, Ghendir Said, A. Chemsa, A. Taleb-Ahmed, *" Novel Detection Algorithm of Speech Activity and the impact of Speech Codecs on Remote Speaker Recognition System", WSEAS Transactions on Signal Processing*, 2014, vol. 10.

[10] Mandasari, M. I., Saeidi, R., & van Leeuwen, D. A. (2015). Quality measures based calibration with duration and noise dependency for speaker recognition. *Speech Communication*, *72*, 126-137.

[11] Sahidullah, Md et Saha, Goutam. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication, 2012, vol. 54, no 4, p. 543-565.

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

[13] Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. *Signal Processing Magazine, IEEE,13*(5), 58.

[14] Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press.

[15] Peinado, A. M., & Segura, J. C. (2006). *Front Matter* (pp. i-xvi). John Wiley & Sons, Ltd.

[16] Vaidyanathan, P. P. (2007). The theory of linear prediction. *Synthesis lectures on signal processing*, *2*(1), 1-184.

[17] 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.

[18] Yi Hu and Philipos C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement". IEEE transactions on audio, speech, and language processing, vol. 16, no. 1, January 2008.

[19] J.S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation", IEEE Trans. Audio, Speech & Lang. Proc., Vol. 16, No. 6, pp. 1112-1123, August 2008.

[20] P. Scalart, and J. Vieira Filho, “Speech Enhancement Based on a Priori Signal to Noise Estimation,” IEEE Intl. Conf. Acoust., Speech, Signal Processing , Atlanta, GA, USA, Vol. 2, pp. 629–632, May 1996.

[21] LU, Yang and LOIZOU, Philipos C. A geometric approach to spectral subtraction. *Speech communication*, 2008, vol. 50, no 6, p. 453-466.

[22] Plapous, C.; Marro, C.; Scalart, P., "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, Issue 6, pp. 2098 - 2108, Nov. 2006.

[23] A.P. Stark, K.K. Wojcicki, J.G. Lyons and K.K. Paliwal, "Noise driven short time phase spectrum compensation procedure for speech enhancement", Proc. INTERSPEECH 2008, Brisbane, Australia, pp. 549-552, Sep. 2008.

[24] I. Cohen .Speech Enhancement Using a Non causal A PrioriSNR Estimator. IEEE, signal processing letters, vol. 11, no. 9, september 2004.

[25] Gerkmann, T. & Hendriks, R. C. Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay. IEEE Trans Audio, Speech, Language Processing, 2012, 20, 1383-1393.

[26] Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., & O’Shaughnessy, D. (2013). Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech communication*, *55*(2), 237-251.