Improving Speaker Identification System Using Discrete Wavelet Transform and AWGN

Heba Maged
Department of Computer Engineering
College of Engineering and Technology,
AAST, Alexandria, Egypt
eng.hebamaged@yahoo.com

Ahmed AbouEl-Farag
Department of Computer Engineering
College of Engineering and Technology,
AAST, Alexandria, Egypt
abouelfarag@aast.edu

Saleh Mesbah
Department of Computer Science
College of Computing and Information
Technology, AAST, Alexandria, Egypt
saleh.mesbah@gmail.com

Abstract-**This paper presents a robust speaker identification method from degraded noisy speech signals. This method is based on Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction from the noisy speech signals and Discrete Wavelet Transform (DWT). A comparative analysis is carried out with the traditional MFCCs based feature extraction method from noisy speech signals with additive white Gaussian noise (AWGN). The implementation mainly incorporates MFCCs which used for feature extraction and Vector Quantization using the Linde-Buzo-Gray (VQLBG) algorithm. It aims to minimize the amount of data to be handled. Results show that feature extraction from DWT of the degraded signals adds more speech features from the approximation and detail components. This helps achieving higher identification rates. Results also show that the proposed method improves the recognition rates computed at different degradation levels using different values of SNR cases.**

Keywords; Speaker Identification, DWT, MFCCs, AWGN.

I. INTRODUCTION

Speaker recognition refers to recognizing every human from their voice. Each speaker has her/his voice characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on. Most recent state of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition rates [1, 2].

In speaker identification systems, the two major operations performed are feature extraction and classification [3]. The feature extraction can be considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate. There are various techniques for extracting speech features in the form of coefficients such as the linear prediction coefficients (LPCs), the Mel-Frequency Cepstral Coefficients (MFCCs) and the Linear Prediction Cepstral Coefficients (LPCCs) [3]. Classification is a process having two phases; speaker modeling and speaker matching. In the speaker modeling step, the speaker is enrolled to the system using features extracted from the training data. When a sample data from unknown speaker arrives, pattern matching techniques are used to map the features to a model corresponding to a known speaker. The combination of a speaker model and a matching technique is called a classifier. Classification techniques used in speaker identification systems include Gaussian Mixture Models (GMMs), Vector Quantization (VQ), HMMs and ANNs [3].

The MFCCs are the most popular acoustic features used in speaker identification. The use of MFCCs for speaker identification provides a good performance in clean environments, but they are not robust enough in noisy environments. They are based on the known evidence that the information carried by low frequency components of the speech signal is more than that carried by high frequency components. The MFCCs assume that the speech signal is stationary within a given time frame and may therefore lack the ability to analyze the localized events accurately [3]. Recently, a lot of research has been directed towards the use of wavelet based features [4]. The discrete wavelet transform (DWT) has a good time and frequency resolution and hence it can be used for extracting the localized contributions of the signal of interest. Wavelet de-noising can be used to suppress noise from the speech signal and lead to good representation of stationary as well as non-stationary segments of the speech signal.

In this paper, a new method for speaker identification is presented. This method is based on feature extraction from DWT. Features are extracted from DWT vector synthesized features in MFCC to generate more features. This increases and enhances the signal representation in Time-Frequency analysis by Wavelet Transform. It generates a model for features extracted by VQLBG model. The objectives of this method are to recognize the noisy speech signal and to enhance the performance of the MFCCs based method in the presence of noise by introducing more features from the signal wavelet transform.

The rest of the paper is organized as follows: Section 2 gives an overview on the structure of any speaker identification system. Section 3 discusses the proposed speaker identification method is introduced. In Section 4, the implementation and process of DWT, process of feature extraction in MFCC and Feature matching. In Section 5, gives the experimental results. Finally, Section 6: summarizes the concluding remarks.

978-1-4799-3279-5 /14/$31.00 ©2014 IEEE

II. SPEAKER IDENTIFICATION SYSTEMS

An Automatic speaker identification system comprises two stages; a feature extraction stage and a classification stage. This system operates in two modes; training and recognition modes. Both of them include a feature extraction step which is sometimes called the front end of the system. The feature extractor converts the digital speech signal into a sequence of numerical descriptors called feature vector [3]. The features exploited in this paper are the MFCCs that model the shape of the time waveform of MFCCs.

A. Classification of Speaker Recognition Methods

The problem of speaker recognition can be divided into two major parts: speaker identification and speaker verification.

1. *Speaker identification:* the task of determining who is talking from a set of known voices of speakers. It is the process of determining who has provided a given utterance based on the information contained in speech waves. The task is referred as closed set identification.

2. *Speaker Verification:* the process of accepting or rejecting the speaker claiming to be the actual one. Since it is assumed that imposters (those who fake as valid users) are not known to the system, this is referred as the open set task. Adding none of the above option to the closed set identification task would enable merging of the two tasks, called open set identification.

Both the text dependent and independent methods share the same problem. These systems can be deceived because someone who plays back a recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. Even the use of pre-determined set of words or digits that are randomly chosen every time can be reproduced in the requested order by advanced electronic recording equipment. Therefore a text prompted (machine driven text dependent) speaker recognition system could be considered. With the merge of speaker and speech recognition systems and improvement in speech recognition accuracy, the distinction between text dependent and independent applications will eventually decrease. The text dependent speaker recognition is the most commercially viable and useful technology, although there has been much research conducted on both the tasks. However, due to the possibilities offered, more attention is being paid to the text independent methods of speaker recognition irrespective of their complexity [5].

Speaker recognition techniques alongside with facial image recognition, fingerprints and retina scan recognition represent some of the major biometric tools for identification of a person. Each of these techniques carries its advantages and drawbacks. If these methods can provide unique identification, then it is still not clear what kind of parametric representations contain information which is essential for the identification process, and for how long and under what conditions, this representation remains valid. As long as these questions are unanswered, there is a scope for research and improvements [6]. Current speaker recognition systems face the challenge of performance degradation due to the speaker's aging, changing health conditions, mental state, the effect of different sources of distortion and all the factors effect on the variability in the speech signal which are (Background noise, Room reverberation, Microphone characteristics, Intra-speaker variability, and Inter-speaker variability) [7].

The exact effects of these factors on speaker recognition are not known. The development area in this paper focused on implementation and effort to understand how speaker recognition could be used as one of the best forms of biometric to recognize the identity of human voice. It briefly describe all the stages from enrollment of voice samples, power spectrum computation, Mel frequency wrapping to plotting of acoustic vectors and VQ code words which generate the highest percentage of matching score. Different standard techniques are also used at the intermediate stage of the processing [8].

The *drawbacks* of the previous design and implementation work of speaker identification system are investigated as follows; the previous traditional MFCCs for feature extraction based speaker identification system is not robust enough in the presence of noise environment, ex; Background noise…etc. So, the feature extraction from the wavelet transform of the degraded signals adds more speech features from the approximation and detail components of these signals which assist in achieving higher identification rates [9].

III. THE PROPOSED SPEAKER IDENTIFICATION METHOD

This research aims to design a system that minimizes the probability of identification errors using an algorithm with MFCC feature extraction technique to extract information content of speech signal. The proposed method transforms a signal to a different domain to get a better representation of the signal. The proposed algorithm uses DWT for signals degraded by additive white Gaussian noise signal (AWGN). DWT is a useful tool to overcome the poor signal representation problems that happen by AWGN. The Multilevel 1-D DWT of a speech signal decomposes the signal into approximation and detail coefficients. Features are extracted from the DWT of the speech signal. Wavelet de-noising is used to reduce the effect of noise prior to speaker identification. The database of this research consists of 13 speakers. Five decomposition levels are prepared for each 13 speakers in the training and test folders and using Daubechies1 (db1) wavelet. DWT is tested before and after adding AWGN to the original signal. The proposed approach for feature extraction in the presence of noise AWGN in the audio signal is illustrated in Fig. 1.

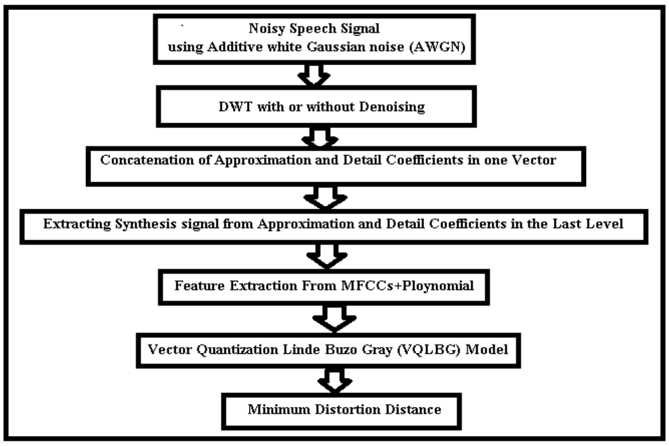

Noisy Speech Signal
using Additive white Gaussian noise (AWGN)
DWT with or without Denoising
Concatenation of Approximation and Detail Coefficients in one Vector
Extracting Synthesis signal from Approximation and Detail Coefficients in the Last Level
Feature Extraction From MFCCs+Ploynomial
Vector Quantization Linde Buzo Gray (VQLBG) Model
Minimum Distortion Distance

Figure 1. The proposed approach for feature extraction in the presence of AWGN in the original signal

A. *Design Description*

The text independent biometric speaker recognition system is accomplished by reading AWGN audio data in 6 different values for SNR in disable units (db.): 0db, 10db, 20db, 25db, 28db and 30db. The data are located in a training folder and clear data from test folder for each 13 speakers in the database. The data are used to compute DWT to get concatenation vector of approximation and detail coefficients and extracted the synthesis signal from the last level coefficients to be used in speech processing for both test and train folders. The MFCC is computed to extract features of the audio data and compute voice quantization of the audio data used to be used in speech processing for both test and train voices [8].

1. *Descriptions of the Wavelet-Based MFCCs feature extraction technique for representing speech signals:*

Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal [10]. An important problem is to determine a representation that is well adapted for extracting information content of speech signals. In general, transforming a signal to a different domain is done to get a better representation of the signal. For recognition, a better representation of the signal means having more ability to separate signals which belong to separate classes or categories in the new domain than in the original domain.

2. *Multilevel 1-D Discrete Wavelet Transform (DWT):*

The DWT is considerably easier to implement when compared to the CWT. DWT provides sufficient information both for analysis and synthesis of the original signal, with a significant reduction in the computation time [11]. In DWT, filters of different cutoff frequencies are used to analyze the signal at different scales. The signal is passed through a series of high pass filters to analyze the high frequencies, and it is passed through a series of low pass filters to analyze the low frequencies.

The resolution of the signal, which is a measure of the amount of detail information in the signal, is changed by the filtering operations, and the scale is changed by up-sampling and down-sampling (subsampling) operations. Subsampling a signal corresponds to reducing the sampling rate, or removing some of the samples of the signal. Subsampling by a factor *n* reduces the number of samples in the signal *n* times. Up-sampling a signal corresponds to increasing the sampling rate of a signal by adding new samples to the signal.

In DWT, scales and positions of powers of two are chosen. An efficient way to implement this scheme using filters is developed by Mallat in 1988 [12]. Given a signal S of length N, and N must be a strictly positive integer, the DWT consists of $\log_2^N$ stages at most with using 'wavelet name' like we use in our work is” Daubechies1 Wavelet” that is ‘db1’. Ingrid Daubechies1, one of the brightest stars in the world of wavelet research, invented what are called compactly supported orthonormal wavelets, thus making discrete wavelet analysis practicable. The first step in DWT produces, starting from S, two sets of coefficients: approximation coefficients CA_1 and the detail coefficients CD_1. These vectors are obtained by convolving S with a low pass filter for approximations, and with a high pass filter for details, followed by dyadic decimation. The next step splits the approximation coefficients CA_1 into two parts using the same scheme, replacing S by CA_1 and producing CA_2 and CD_2, and so on. The DWT of the 3 decomposition levels are illustrated in the Fig. 2. This technique is most effective when it is applied to the detection of short-time phenomena, discontinuities, or abrupt changes in the signal.

3. *Vector Quantization Recognition Techniques:*

The vector quantize encoder computes for a given input, the index of nearest code word based on Euclidean or weighted Euclidean distance measure. The Vector Quantize Encoder block compares each input column vector to the codeword vectors in the codebook matrix. Each column of this codebook matrix is a codeword. The block finds the codeword vector nearest to the input column vector and returns its zero-based index. This block supports real floating point and fixed-point signals on all input ports. The block finds the nearest codeword by calculating the distortion. The block uses two methods for calculating distortion: Euclidean squared error (unweight) and weighted Euclidean squared error. Consider the codebook, $CB=[CW_1\ CW_2\ ....\ CW_N]$. This codebook has N code words; each codeword has k elements. The ith codeword is defined as a column vector, $CW_i = [a_{1i}\ a_{2i}\ ...\ a_{ki}]$. The multichannel input has M columns and is defined as $U= [U_1\ U_2\ ....\ U_M]$, where the pth input column vector is $U_p= [U_{1p}\ U_{2p}\ ....\ U_{kp}]$. The squared error (un-weighted) is calculated as in (1):

$$D = \sum_{j=1}^{k}(aji - ujp)^2 \qquad (1)$$

The weighted squared error **w** is calculated as in (2):

$$D = \sum_{j=1}^{k} wj\ (aji - ujp)^2 \qquad (2)$$

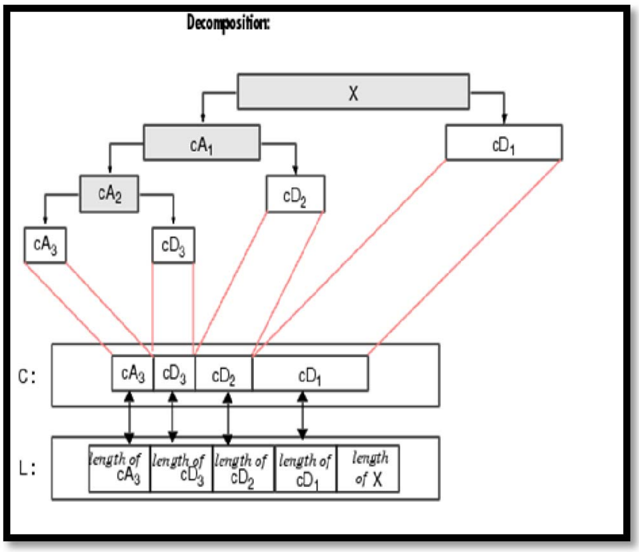


Figure 2. The DWT structure of level-3 decomposition

IV. IMPLEMENTATION

A. Speaker (Voice) Recognition Algorithm

A voice analysis is carried out after taking an input through microphone from a user. The design of the system is improved using DWT. It involves manipulation of the input audio signal at different levels. Different operations are performed on the input signal such as Pre-emphasis, discrete wavelets transform (DWT), Framing, Windowing, Mel-Cepstrum analysis, and recognition (matching) of the spoken word.

The speaker recognition system consists of two distinguished phases. the enrolment or training phase, and the recognition or testing phase as described in Fig. 3 [13].

1. *Training Phase:* Each speaker has to provide samples of their voice so that the reference template model can be built.
2. *Testing Phase:* To ensure that input voice matches with stored reference template model and make recognition decision.

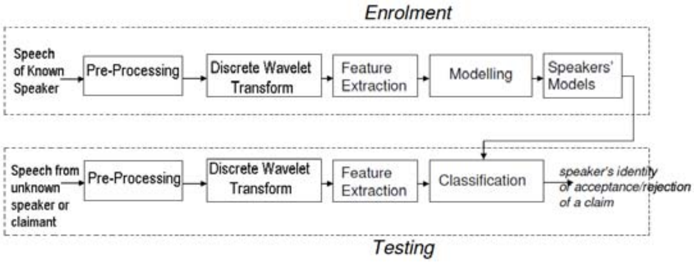

Enrolment
Speech of Known Speaker
Pre-Processing
Discrete Wavelet Transform
Feature Extraction
Modelling
Speakers' Models
Speech from unknown speaker or claimant
Pre-Processing
Discrete Wavelet Transform
Feature Extraction
Classification
speaker's identity or acceptance/rejection of a claim
Testing

Figure 3. A conventional speaker (voice) recognition system

B. The Multilevel 1-D DWT Steps

This section describes the steps to perform the Daubechies1 'db1' wavelet decomposition for each 13 speakers in database, where each audio is contaminated by AWGN. The steps are to decompose and de-noise audio signal and finally synthesize the signal. First, a signal S of length N samples is input DWT that consists of $\log_2^N$ stages using 'db1' wavelet. The transform returns the coefficients of all the components then extract the approximation and details coefficients CA_n, CD_1,……CD_n of the last n level that are concatenated in one vector. Then reconstruct the approximation and details in the last level.

C. MFCCs Features Extraction Technique

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1KHz. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech.

MFCC consists of seven computational steps. Each step has its function and mathematical approaches as follows.

Step 1: Pre-Emphasis

In this process, the speech signal is pre-emphasized to remove glottal and lip radiation effects. This step processes the passing of signal through a filter which emphasizes higher frequencies and is called a first order finite impulse response (FIR) filter. This process will increase the energy of signal at higher frequency as in (3).

$$Y[n] = X[n] - 0.95\ X\ [n-1] \qquad (3)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 ms. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N = 256.

Step 3: Hamming Windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines [8]. For each frame, a windowing function is usually applied to increase the continuity between adjacent frames. Common windowing functions include the rectangular window, the Hamming window, the Blackman window and flattop window. According to the convolution theorem, the windowing corresponds to a convolution between the short term spectrum and the window function frequency response. A good window function has a narrow main lobe and low side lobe levels in its frequency response. The most commonly used window function in speech processing is the Hamming window which defined as W_H (n) as in (4) [9]:

$$W_H (n) = 0.54\text{-}0.46 \cos\left(\frac{2n\pi}{N-1}\right), \qquad (4)$$

Where, *N* is the number of samples in each frame, n = 0, 1… N-1.Then, the result of windowing is the signal can be given as in (5):

$$Y_t(n)= X_t(n)*W_H(n),\ 0\leq n \leq N\text{-}1 \qquad (5)$$

Step 4: Fast Fourier Transform (FFT)

FFT converts each frame of N samples from time domain into frequency domain. Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports as in (6):

$$Y (w) = FFT\ [h (t)*X (t)] = H (w)*X (w) \qquad (6)$$

Where, X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Step 5: Mel-Frequency Spacing Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decrease linearly to zero at center frequency of two adjacent filters [14]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for a given frequency **f** in Hz as in (7):

$$F (Mel) = [2595 *\log_{10} (1+\left(\frac{f}{700}\right)] \qquad (7)$$

Step 6: Discrete Cosine Transform

This is the process that converts the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector. Therefore if we denote those energy of the k^{th} Mel-filter output or k^{th} Mel power spectrum coefficients that are the result of the last step are $\tilde{S}$(k), k = 0, 2,….., k-1, the MFCC's , $\hat{c}_{n,}$ calculated in (8).

$$\hat{c}_n = \sum_{k=1}^{k}(\log \tilde{S}(k)) \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{k}\right],\ n = 0,1 ,\ldots, k\text{-}1 \qquad (8)$$

Step 7: Delta Energy and Delta Spectrum

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal X in a window from time sample t1 to time sample t2, is represented as in (9) [33].

$$Energy = \Sigma\ X^2[t] \qquad (9)$$

D. Speech Samples Matching

All the 13 voice samples data values are loaded without using DWT part. The voice samples using DWT are loaded in sound database. Training samples with or without added noise are tested one by one to database stored in another file without using DWT. Trained samples with or without added noise are tested one by one to de-noised database stored using DWT. The matching results are given on the basis of minimum distortion distance between the corresponding sound files in both test and train files to eventually determine precisely the best match and identify the speaker [17].

V. RESULTS

A comparison between the MFCCs system and the current recognition system using wavelet-based MFCCs is made. An evaluation of noisy signal using AWGN with 6 different values of SNR:0db, 10db, 20db, 25db, 28db, and 30db for 8 and 13 speakers.

The first part is the test patterns for 8 speakers and as the results show in the Table (1) for clean speech, an identification rate of 100% of the time is achieved by the wavelet-based MFCCs feature extraction technique and using the MFCCs technique. When the test patterns are corrupted with White Gaussian noise, the performance of the system using MFCCs and wavelet-based MFCCs features is affected significantly by the added noise signal with *S/N=0dsb* and there is no improvement. Increasing the S/N from *10db* to *30db* with White Gaussian noise shows an improvement in the identification rate and the error rate. When increasing the S/N with additive white Gaussian noise, better identification rate with low error rate using MFCCs.

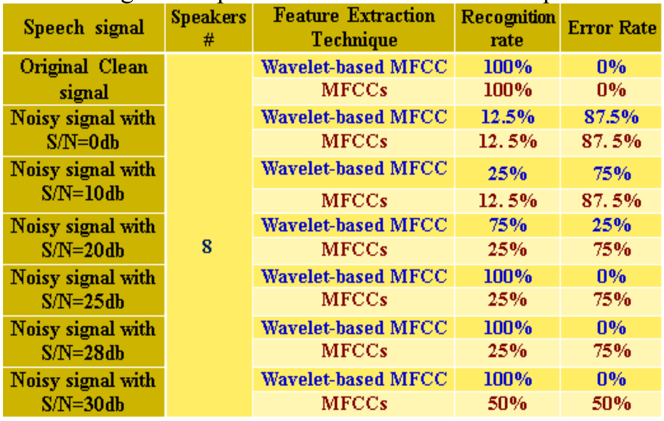
The second part is the test patterns for 13 speakers and as the results show in the Table (2) for clean speech, an identification rate is improved in the proposed feature extraction technique over the MFCCs technique. Increasing the S/N with additive white Gaussian Noise is demonstrated in Table (2). There is much improvement in identification rate with low error rate of the proposed system. The proposed technique demonstrates the same best results for 13 speakers as wavelet-based MFCC of 8 speakers.

I. CONCLUSION

In this paper, DWT is used to represent the voice features. The proposed method used power spectrum computation. The system tests gave high percentage of matching score. Different standard techniques are used at the intermediate stage of the processing. Multilevel approximations and details resolutions channels are obtained. The MFCCs of the DWT channels are calculated for capturing the characteristics of the speech signals with decrease the rate of losing features during the feature extraction. The proposed method is used in feature extraction with noise contaminations in the speech signals.

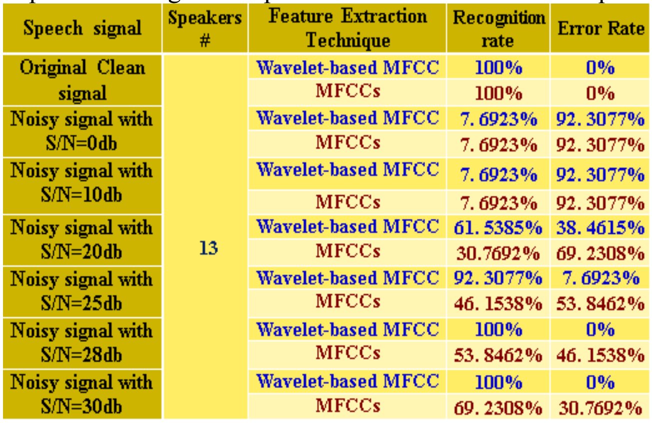
Results showed that the proposed technique gives better performance than MFCCs. In addition, this technique reduces the noise effect and improves the recognition rate when dealing with noisy speech signals like AWGN with different values of SNR compared to the MFCCs which operate only in clean environment.

TABLE1. Recognition Rates and Error Rates of 8 Speakers Using the Proposed and the MFCCs Techniques



Speech signal	Speakers #	Feature Extraction Technique	Recognition rate	Error Rate
Original Clean signal	8	Wavelet-based MFCC	100%	0%
		MFCCs	100%	0%
Noisy signal with S/N=0db		Wavelet-based MFCC	12.5%	87.5%
		MFCCs	12. 5%	87. 5%
Noisy signal with S/N=10db		Wavelet-based MFCC	25%	75%
		MFCCs	12. 5%	87. 5%
Noisy signal with S/N=20db		Wavelet-based MFCC	75%	25%
		MFCCs	25%	75%
Noisy signal with S/N=25db		Wavelet-based MFCC	100%	0%
		MFCCs	25%	75%
Noisy signal with S/N=28db		Wavelet-based MFCC	100%	0%
		MFCCs	25%	75%
Noisy signal with S/N=30db		Wavelet-based MFCC	100%	0%
		MFCCs	50%	50%

TABLE2. Recognition Rates and Error Rates of 13 Speakers Using the Proposed and The MFCCs Techniques



Speech signal	Speakers #	Feature Extraction Technique	Recognition rate	Error Rate
Original Clean signal	13	Wavelet-based MFCC	100%	0%
		MFCCs	100%	0%
Noisy signal with S/N=0db		Wavelet-based MFCC	7. 6923%	92. 3077%
		MFCCs	7. 6923%	92. 3077%
Noisy signal with S/N=10db		Wavelet-based MFCC	7. 6923%	92. 3077%
		MFCCs	7. 6923%	92. 3077%
Noisy signal with S/N=20db		Wavelet-based MFCC	61. 5385%	38. 4615%
		MFCCs	30.7692%	69. 2308%
Noisy signal with S/N=25db		Wavelet-based MFCC	92. 3077%	7. 6923%
		MFCCs	46. 1538%	53. 8462%
Noisy signal with S/N=28db		Wavelet-based MFCC	100%	0%
		MFCCs	53. 8462%	46. 1538%
Noisy signal with S/N=30db		Wavelet-based MFCC	100%	0%
		MFCCs	69. 2308%	30.7692%

II. REFERENCES

[1] Gonzalez-Rodriguez, J., Garcia-Gomar, D. G.-R. M., Ramos-Castro, D., Ortega-Garcia, J. “Robust likelihood ratio estimation in Bayesian forensic speaker recognition”. In: Proc. 8th European Conf. on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, September 2003, pp. 693–696.

[2] Thiruvaran, T., Ambikairajah, E., Epps, J., 2008. “FM features for automatic forensic speaker recognition”. In: Proc. Interspeech 2008, Brisbane, Australia, September 2008, pp. 1497–1500.

[3] D. Pullella, "Speaker Identification Using Higher Order Spectra", Dissertation of Bachelor of Electrical and Electronic Engineering, University of Western Australia, 2006.

[4] B. C. Jong, "Wavelet Transform Approach For Adaptive Filtering With Application To Fuzzy Neural Network Based Speech Recognition", PhD Dissertation, Wayne State University, 2001.

[5] S. K. Singh, Prof P. C. Pandey, “Features and Techniques for Speaker Recognition”, M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03.

[6] Memon S., “Automatic Speaker Recognition: Modelling, Feature Extraction and Effects of Clinical Environment,” School of Electrical and Computer Engineering Science, Engineering and Technology, RMIT University, June 2010.

[7] E. Moore, M. Clements, J. Peifer, and L. Weisser, “Analysis of prosodic variation in speech for clinical depression,” in Proceedings, 25th Annual Conference on Engineering in Medicine and Biology, 2003, pp. 2925–2928.

[8] Gbadamosi L.,”Text Independent Biometric Speaker Recognition System,” International Journal of Research in Computer Science, *Computer Science Department, Lagos State Polytechnic, Lagos, Nigeria,* eISSN 2249-8265 Volume 3, Issue 6 (2013), pp. 9-15.

[9] A. Shafik, Elhalafawy, Diab, Sallam and Abd El-samie, “A Wavelet Based Approach for Speaker Identification from Degraded Speech” International Journal of Communication Networks and Information Security (IJCNIS), Vol. 1, No. 3, December 2009.

[10] A.K. Jain, R. Bolle, and S. Pankanti, “Biometrics: personal identification in networked society”, Springer, 1999.

[11] *POLIKAR, R.,”* wavelet tutorial,” The Wavelet Tutorial is hosted by Rowan University, College of Engineering Web Servers Dept. of Electrical and Computer Engineering, March 07,1999.

[12] S Mallat, “A theory for multiresolution signal decomposition: the wavelet representation”, IEEE Trans. On Pattern Analysis and Machine Intelligence, pp. 674-693, 1989.

[13] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques” Journal of computing, Vol 2, Issue 3, March 2010.

[14] Palden Lama and Mounika Namburu, “Speech Recognition with Dynamic Time Warping using MATLAB”, CS 525, SPRING 2010-PROJECT REPORT.