DWT and MFCCs based Feature Extraction Methods for Isolated Word Recognition

Mahmoud I. Abdalla
Department of Electronics and Communications,
Zagazig University, Egypt.

Haitham M. Abobakr
Department of Computer and Systems,
Zagazig University, Egypt.

Tamer S. Gaafar
Department of Computer and Systems,
Zagazig University, Egypt

ABSTRACT

A new method for feature extraction is presented in this paper for speech recognition using a combination of discrete wavelet transform (DWT) and mel Frequency Cepstral Coefficients (MFCCs)**.** The objective of this method is to enhance the performance of the proposed method by introducing more features from the signal. The performance of the Wavelet-based mel Frequency Cepstral Coefficients method is compared to mel Frequency Cepstral Coefficients based method for features extraction. Wavelet transform is applied to the speech signal where the input speech signal is decomposed into various frequency channels using the properties of wavelet transform. then Mel-Frequency Cepstral Coefficients (MFCCs) of the wavelet channels are calculated. A new set of features can be generated by concatenating both features. The speech signals are sampled directly from the microphone. Neural Networks (NN) are used in the proposed methods for classification. The proposed method is implemented for 15 male speakers uttering 10 isolated words each which are the digits from zero to nine**.** each digit is repeated 15 times.

General Terms

Speech Recognition, Isolated Word Recognition.

Keywords

Speech Recognition, Feature Extraction, Mel-Frequency Cepstral Coefficients, Discrete Wavelet Transforms, Neural Networks.

1. INTRODUCTION

Speech recognition system is divided into two parts, feature extraction and classification. Feature extraction method plays a vital role in speech recognition task.

Isolated word/sentence recognition requires the extraction of features from the recorded utterances followed by a training phase.The most widely used feature extraction techniques are the Perceptual Linear Predictive (PLP), the Linear Prediction Coefficients (LPC), the Linear Prediction Cepstral Coefficients (LPCC) and Coefficients. Mel Frequency Cepstral Coefficients (MFCC) and various forms of the Mel Frequency Cepstral Coefficients (ΔMFCC, ΔΔMFCC) [1]. A relatively new technique used in the field of signal processing for feature extraction is the wavelet technique.

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. 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. Recently, a lot of research use the wavelet based features. The discrete wavelet transform (DWT) has a good time and frequency resolution. Wavelet denoising can also be used to suppress noise from the speech signal and it can lead to a good representation of stationary and non-stationary segments of the speech signal.

There are two phases for speech recognition, training and testing. The classification process is done using the trained data where the parameters of the classification model are estimated using the Training Data and then during the testing phase the feature of the test pattern is compared to that of the trained model for each class. When a matching occurs the test pattern is classified to belong to that class (the matched one).[2]

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]

Automatic Speech Recognition (ASR) systems based on the Hidden Markov Model (HMM) have been used extensively in the mid 1980's. HMM is a well-known and widely used statistical method for characterizing the spectral features of speech frame. The use of the HMM depends on that the speech signal can be well characterized as a parametric random process, and the parameters of the stochastic process can be predicted in a precise, well defined manner. [4]

Artificial Neural Networks (ANNs) have been investigated for many years in automatic speech/speaker recognition. These ANNs consists of many neurons divided in two or more layers. These neurons are non-linear computational elements operating in parallel in patterns similar to the biological neural networks. ANNs have been used extensively in speech recognition during the past two decades. The most important advantages of ANNs for solving speech/speaker recognition problems are their error tolerance and non-linear property.[5]

In this paper, a new method for speaker recognition is presented. This method is based on the extraction of the wavelet parameters from the original speech signal, then the Mel-Frequency Cepstral Coefficients MFCCs) of the wavelet channels are calculated. A Multi layer perceptron architecture is used for classification and recognition. The objective of this method is to enhance the better performance of the new method for feature extraction over using MFCCs based method.

The rest of the paper is organized as follows. Section 2 discusses the Discrete Wavelet Transform process In Section 3 Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction is discussed. Section 4 discusses the Artificial Neural Network for pattern recognition. The Database used is discussed in section 5. In Section 6 the proposed speaker identification method is introduced and also gives the experimental results. Finally, Section 7 summarizes the concluding remarks.

2. DISCRETE WAVELET TRANSFORM

The wavelet technique is considered a relatively new technique in the field of signal processing for feature extraction. Wavelet transforms have been used by several ireseraches for automatic speech recognition, speech coding and compression, speech denoising and enhancement and other processes.[6] It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allows better time resolution at high frequencies than Fourier Transform.[2]

The term **wavelet** means a **small wave**.The smallness refers to the condition that this (window) function is of finite length. And as this function is oscillatory so it is called wave.[7]

Wavelet transform was introduced as it is more suitable to deal with non-stationary signals like speech. [8] The advantage of using the Wavelet over the Fourier Transform is that it provides at what time which frequency is present. So it provides time frequency information of the signal. The Fourier Transform (FT) is not suitable for the analysis of such non stationary signal because it provides only the frequency information of signal but does not provide the information about at what time which frequency is present.[9]

So Wavelets have the ability to analyze different parts of a signal at different scales. The wavelet transform (WT) is a transformation that provides time-frequency representation of the signal. [10]

The continuous one dimensional wavelet transform (CWT) is a decomposition of f (t) into a set of basis function $\psi_{a,b}$(t) called wavelets:

$$w(a,b) = \int f(t)\ \psi^*_{a,b}(t)\, dt \qquad (1)$$

The wavelets are generated from a single mother wavelet called

$\psi_{a,b}$ (t) by dilation and translation[8] As the functions with different region of support that are used in the transformation process are derived from one main function, so it is called the mother wavelet.[7]

$$\psi_{a,b} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \qquad (2)$$

Where: $f(t)$ is the signal to be analyzed, a is the scale, and b is the translation factor. ψ (t) is the transforming function and is called the mother wavelet. Filters of different cut off frequencies are used to analyze the signal[5]

As CWT is a function of two parameters, scale and translation parameters as $a=2^j$ and $b=2^j k$. So DWT theory requires two sets of related functions called scaling function and wavelet function given by

$$\phi(t) = \sum_{n=0}^{N-1} h[n]\sqrt{2}\,\phi(2t-n) \qquad (3)$$

And

$$\psi(t) = \sum_{n=0}^{N-1} g[n]\sqrt{2}\,\phi(2t-n) \qquad (4)$$

where, function $\psi(t)$ is called scaling function, $h[n]$ is an impulse response of a low pass filter and $g[n]$ is an impulse response of a high pass filter.[9]

The basic operation principles of DWT are similar to the CWT however the difference between them is that the scales used by the wavelet and their positions are sampled up or down by a factor of two. This is called the dyadic scales.[11] The discrete wavelet transform (DWT) uses filter banks for the construction of the multi resolution time-frequency plane.

A filter bank consists of filters which separate a signal into frequency bands.[12] for a two stage filter it consists of a low pass filter L(z) and a high pass one H(z). For many signals, the low-frequency content is the most important part. It is what gives the signal its identity. In wavelet analysis, the high-scale ,low-frequency components of the signal are called the approximations, while the low-scale, high-frequency components are called the details. [13]

The filtering process, at its most basic level, looks like this

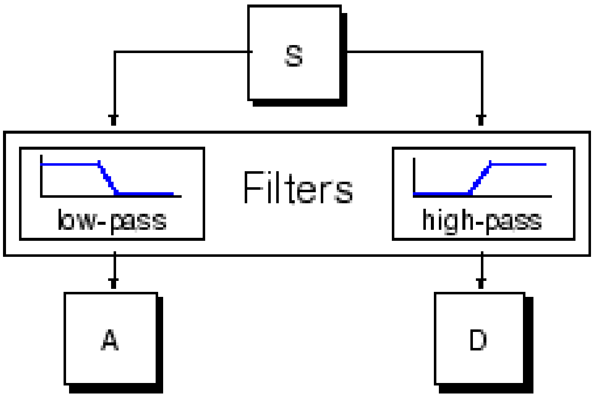


Fig 1: A Two Stage Filter

The output of the filters each contain half the frequency content, but an equal amount of samples as the input signal. The two outputs together contain the same frequency content as the input signal, however the amount of data is doubled. Therefore downsampling by a factor two, denoted by ↓ 2, is applied to the outputs of the filters in the analysis bank. The process of downsampling is that what produces the DWT coefficients.[12]

Given a signal S of length N, the DWT consists of $\log_2$ N stages at most. The first step produces, starting from S, two sets of coefficients: approximation coefficients CA1 and the detail coefficients CD1. These vectors are obtained by convolving S with a low pass filter for approximations, and with a high pass filter for details.[10]

The decomposition process can be repeated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower resolution components. This is called the wavelet decomposition tree.

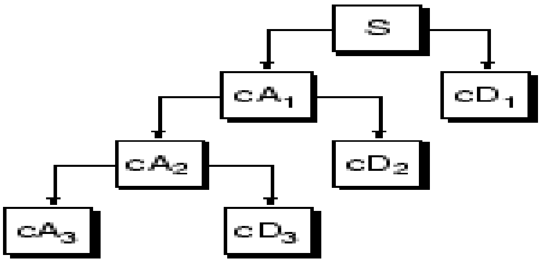


Fig 2: Three Level Wavelet Decomposition Tree

There are a lot of wavelet families, the Daubechies are the most widely used in speech recognition problems. The daubechies names are written as dbN where N is the order of the family[13].

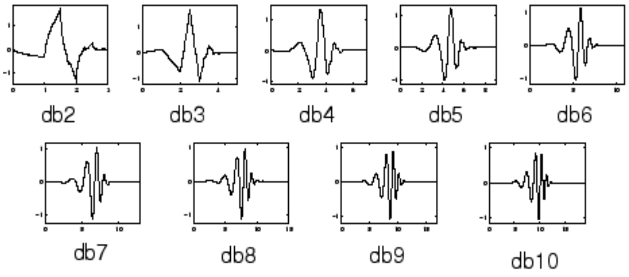


Fig 3: The Daubechies Wavelet Functions

3. MFCCs

Mel-frequency Cepstrum (MFC) is the representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.[14]

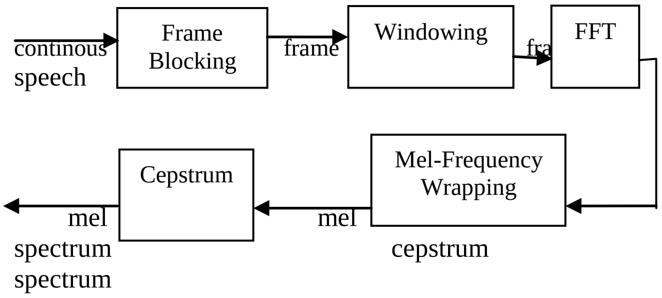


Fig 4: MFCC Extraction Process

The first step is the frame blocking step where the continuous speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* (*M* < *N*). The first frame consists of the first *N* samples. The second frame begins *M* samples after the first frame, and overlaps it by *N* - *M* samples and so on. The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Hamming window is the most commonly used which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

(5)

The next step is the Fast Fourier Transform (FFT), which converts each frame of *N* samples from the time domain into the frequency domain. The output of this step is known as *spectrum*. then The magnitude spectrum is frequency warped in order to transform the spectrum into the Mel-frequency scale. The Mel-frequency warping is performed using a Mel-filter bank composed of a set of bandpass filters with constant bandwidths and spacings on the Mel-scale .The bank consists of one filter for each desired Mel-frequency component, where each filter has a triangular filter bandpass frequency response. The triangular filters are spread over the entire frequency range from zero to the Nyquist frequency.[15]

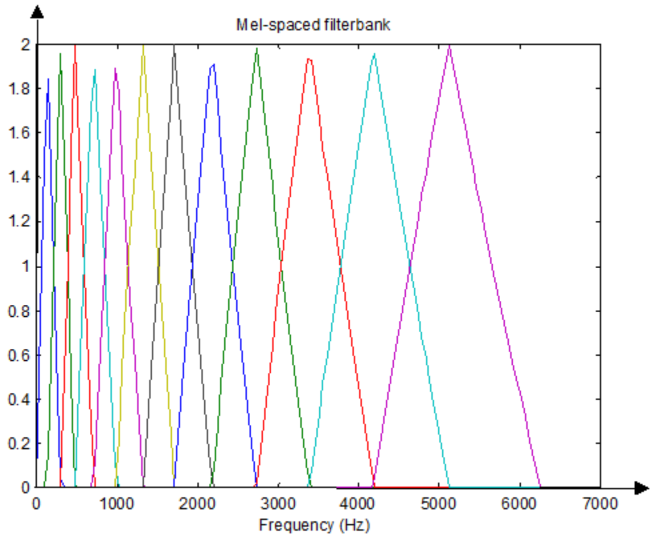


Fig 5: Mel Filter Bank

One mel is defined as one thousandth of the pitch of a 1 kHz tone (Huang et al.2001). Mel-scale frequency can be approximate by Eq. (6):

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f}{700}\right)$$

(6)

This non-linear transformation can be seen in Figure. It shows that equally spaced values on mel-frequency scale correspond to non-equally spaced frequencies.

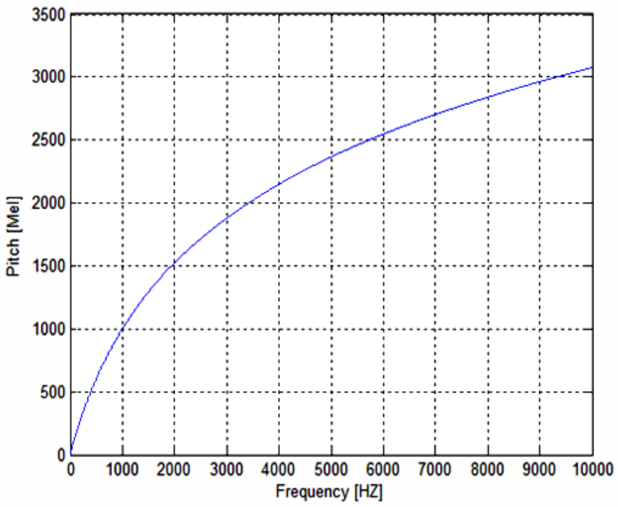


Fig 6: Mel-to-Linear Frequency scale transformation

So, the mel scale can model the sensitivity of the human ear more closely than a purely linear scale, and provides for greater discriminatory capability between speech segments.[16]

In the final step, the log mel spectrum is converted back to time. Where we take the Discrete Cosine transform (DCT) of the mel-scaled log-filter bank energies to calculate MFCCs. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis[15].

The number of the resulting MFCCs is chosen between 12 and 20, since most of the signal information is represented by the first few coefficients.[3]

4. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) is a powerful pattern recognition technique, Neural Networks are widely used in the last two decades in the hope of achieving human-like performance in automatic speech/speaker recognition. It is used for feature matching (classification process). The feature matching process between the features of a new speaker (testing data) and the features saved in the database (training data) in automatic speaker identification systems is called the classification step [3].The most important advantages of ANNs for solving speech/speaker recognition problems are their error tolerance and non-linear property. In several studies, a wavelet neural network was used for speech recognition. [5]

Neural Networks are used for speech recognition problems. neural network models attempt to mimic the human brain by using some organizational principles such as learning, generalization, adaptively, fault tolerance etc..[10]

A neural network (NN) is a massive processing system that consists of many processing elements called neuron. An artificial neuron is the smallest unit that constitutes the artificial neural network.[17] Each neuron in the neural network is characterized by an activation and a bias functions, and each connection between two neurons by a weight factor.[3] A Multilayer Perceptron (MLP) network consists of an input layer, one or more hidden layers, and an output layer. Each layer consists of multiple neurons. The actual computation and processing of the neural network happens inside the neuron [17].

The perceptron model is shown in Figure 7, where *x* is an input vector, *w* is a weight vector, w_0 is the bias and the activation function is a step function.

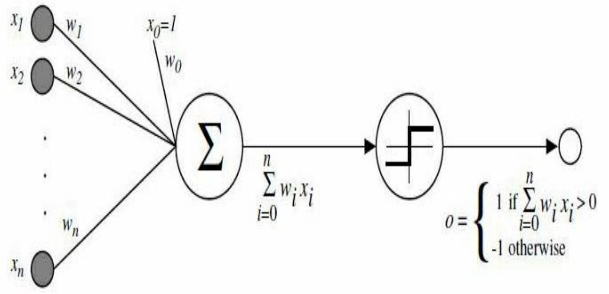

x_1
w_1
$x_0=1$
w_0
x_2
w_2
w_n
x_n
Σ
$\sum_{i=0}^{n} w_i x_i$
$o = \begin{cases} 1 \text{ if } \sum_{i=0}^{n} w_i x_i > 0 \\ -1 \text{ otherwise} \end{cases}$

Fig 7: Model of a perceptron

The hidden layers act as a feature extractor and use a nonlinear function such as sigmoid or a radial-basis function to functions of input. The outputs of all the neurons in the hidden layer serve as input to all of the neurons on the next layer. The output layer acts as a logical net that chooses an index to send to the output on the basis of inputs it receives from the hidden layer, so that the classification error is minimized.[18]

During the training process, network architecture and connection weights are updated for proper classification. Training a neural network is accomplished by adjusting its weights using a training algorithm. The training algorithm adapts the weights by attempting to minimize the sum of the squared error between a desired output and the actual output of the output neurons given by

$$E = \frac{1}{2} \sum_{o=1}^{O} D_{o} - Y_{o} \qquad (7)$$

where D_o and Y_o are the desired and actual outputs of the oth output neuron. O is the number of output neurons. Each weight in the neural network is adjusted by adding an increment to reduce E as rapidly as possible.[3]

The main advantage of using neural networks is that they have the ability to learn complex nonlinear input-output relationships by using training procedures and adapting themselves to the data. Algorithms based on neural networks are well suitable for addressing speech recognition tasks.[19]

5. DATABASE

An indoor database was created from all ten English digits from zero to nine. A number of 15 individual male persons Arabic native speakers were asked to utter all digits fifteen times. Hence, the database consists of 15 repetitions of every digit produced by each speaker, totaling of 2250 tokens. All the 2000 tokens were used for training and testing phases such that 66.67% of the data were used for training and 33.33% were used for testing.

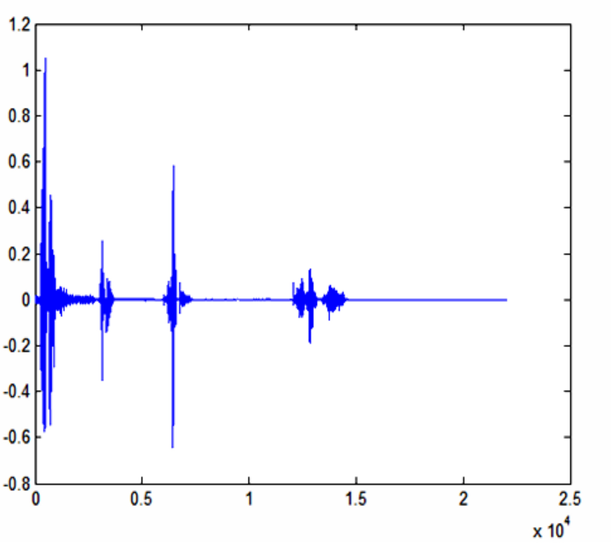
6. EXPERIMENTAL RESULTS

In this paper, a robust feature extraction algorithm for speech signals is applied. This algorithm introduces a recently used method for feature extraction stage depends on combining both the wavelet transform and the MFCCs. First, the speech signal is decomposed into two different frequency channels using the wavelet transform. These two frequency channels are the approximations which are the components of the low frequency channel, while the high frequency channel components are the details. The decomposition process can be successively iterated for further approximations being decomposed.

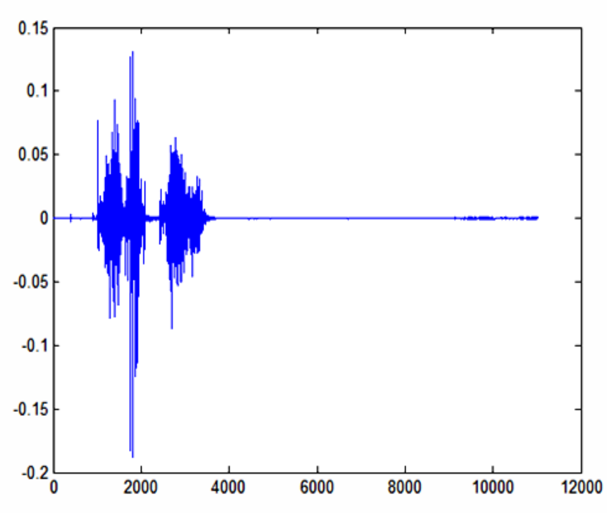
Second, the MFCCs of the approximations and detail channels are calculated, based on this mechanism, the multi-resolution features of the speech signal can easily be extracted using the wavelet decomposition and calculating the related coefficients.

In this work, the signals are decomposed at level 3 using db7 wavelet which is suitable for the problem of ASR. As more decomposition processes leads to computational complexity and useless increase in the number of features without such a significant information. Fig.(8) shows a sample speech signal and its wavelet.

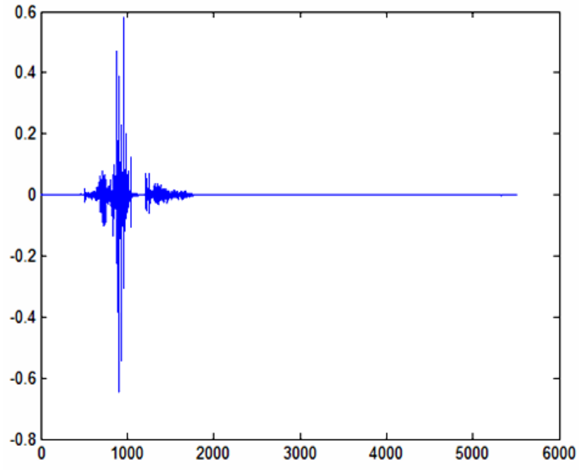
a) Original Sample



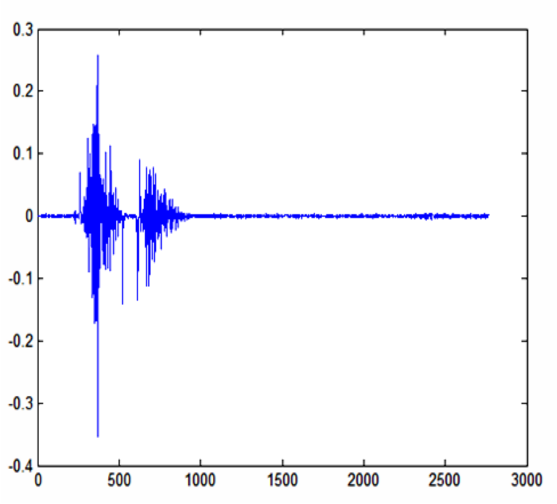
b) Approximation A3



c) Detail D1



d) Detail D2



e) Detail D3

Fig.(8) a sample of speech signal and its wavelet

In this approach, Dwt for the original voice of all speakers is taken gives the approximate and detail coefficients then MFCC is applied to all of the obtained features (both approximate and detail) and then data is partitioned for the train and test phase.

In the second approach MFCC is applied to the original speech signal, data is also partitioned for the training and testing phase. A comparison is taken between the two approaches.

NN is used for the recognition process. The number of layers and neurons in each layer is determined by just trial and error as there is no direct method till now to determine them. In our work the neural network consists of 3 layers an input layer, one hidden layer and the output layer where the number of neurons in each layer is 50, 30, and 15 respectively

The results showed that a recognition rate of 99.6% is achieved using the wavelet based MFCCs features compared to 99.2% using MFCCs only as shown in the table below. These results were obtained using Matlab.

Table 1. Recognition Rates using The Proposed and The Mfccs Techniques

Feature Extraction Technique	**Recognition Rate**
Wavelet-based MFCCs	99.6 %
MFCCs	99.2 %

7. CONCLUSION

This paper proposes the idea that the speaker Recognition system performance can be improved by using a new feature extraction technique A new set of features can be generated by concatenating both DWT and MFCCs features. The MFCCs of the wavelet channels are calculated for capturing the characteristics of the speech signals. Neural Network is used for feature extraction. The results showed an improvement of the recognition rate of this new method over using MFCCs feature only where the new method gives a recognition rate of 99.6 % versus 99.2% using mfccs only.

8. REFERENCES

[1] M. C. Shrotriya, Omar Farooq, and Z. A. Abbasi, " Hybrid Wavelet based Lpc Features for Hindi Speech Recognition ", International Journal of Information and Communication Technology, March 2008.

[2] Santosh k. Gaikwad, Bharti W. Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", international journal of computer applications, November 2010**.**

[3] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam and F. E. Abd El-samie, "A Wavelet based Approach for Speaker Identification from Degraded Speech", International Journal of Communication Networks and Information Security (IJCNIS), December 2009.

[4] Yousef Ajami Alotaibi, "Comparative Study of ANN and HMM to Arabic Digits Recognition Systems", JKAU: Eng. Sci., Vol. 19 No. 1, pp: 43-60 (2008 A.D. / 1429 A.H.)

[5] Engin Avci, "An Automatic System for Turkish Word Recognition Using Discrete Wavelet Neural Network based on Adaptive Entropy", The Arabian Journal for Science and Engineering, October 2007

[6] Daniel Motlotle Rasetshwane, "Identification of Transient Speech Using Wavelet Transforms", MS, thesis submitted to University of Pittsburgh, 2005.

[7] Robi Polikar, " The Wavelet Tutorial ", 2006.

[8] N. S. Nehe and R. S. Holambe, " DWT and LPC based Feature Extraction Methods for Isolated Word Recognition" , EURASIP Journal on Audio, Speech, and Music Processing, 2012

[9] Sonia Sunny, David Peter S and K Poulose Jacob, Recognition of Speech Signals: An Experimental Comparison of Linear Predictive Coding and Discrete Wavelet Transforms ", Sonia Sunny et al. / International Journal of Engineering Science and Technology (IJEST), April 2012.

[10] Mahmoud I. Abdalla and Hanaa S. Ali, " Wavelet-based Mel – Frequency Cepstral Coefficients for Speaker Identification using Hidden Markov Models ", Journal of telecommunications , March 2010.

[11] G. Madhavilatha , A .Mallaiah, and T.Venkata Lakshmi, "Advanced Speaker Verification System Using Wavelets" International Journal of Engineering Research and applications (IJERA), ISSN: 2248-9622 www.ijera.com Vol. 1, Issue 3, pp.891-898.

[12] R.J.E. Merry, " Wavelet Theory and Applications " , Alexandria . tue . nl , Eindhoven, June 7 , 2005

[13] " Wavelet Toolbox ", Help Quide, Matlab 2012.

[14] Amruta Anantrao Malode and Shashikant Sahare, " Advanced Speaker Recognition ", International Journal of Advances in Engineering & Technology, July 2012.

[15] " An Automatic Speaker Recognition System ", http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition.

[16] Noelia Alcaraz Meseguer, " Speech Analysis for Automatic Speech Recognition ", Thesis submitted to Norwegian University of Science and Technology Department of Electronics and Telecommunications, 2009.

[17] Meysam Mohamad pour and Fardad Farokhi, " An Advanced Method for Speech Recognition ", World Academy of Science, Engineering and Technology 25 2009.

[18] Veera Ala-Keturi, " Speech Recognition based on Artificial Neural Networks ", Helsinki hnology Institute of Tec, 2004.

[19] Sonia Sunny, David Peter S and K Poulose Jacob, "Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words ", International Journal of Computer Applications, January 2012.