ENHANCED NEURAL NETWORK CLASSIFIER FOR PATTERN RECOGNITION OF SPEECH SIGNAL BASED ON DISCRETE WAVELET TRANSFORMS

B.Bhagyasri
Asst Prof
CMRIT
bhagyasri38@gmail.com

Abstract

Automatic Emotion Recognition (AER) from discourse finds more prominent importance in better man machine interfaces and apply autonomy. Discourse feeling based investigations firmly identified with the databases utilized for the examination. We have made and broke down three enthusiastic discourse databases. Discrete Wavelet Transformation (DWT) was utilized for the component extraction and Artificial Neural Network (ANN) was utilized for example characterization. Highlight extraction in discourse preparing is one of the primary stages to create discourse handling applications. An extensive arrangement of highlight extraction techniques is accessible to actualize on discourse handling approaches, anyway the decay through Wavelet bundles is a standout amongst the most mainstream these days for its strength. This paper portrays the improvement and usage of the WPD method utilizing discourse tests of the expressions of/cero/and/uno/. The trademark coefficients that after effect of the WPD are entered in an example acknowledgment dependent on neural systems to characterize information and perceive between the expressed words. The outcomes demonstrate an order above 75%, which shows the appropriateness of the strategy for acknowledgment.

Keywords: *Wavelet Packet Decomposition, Wavelet Transform, Non-Audible Murmur, Neural Networks*

1.0 INTRODUCTION:

Speech processing systems have allowed developing many applications; from speech coding, text-to-speech synthesis, speaker identification/verification, to automatic speech recognition systems from which the development of natural language processing can be achieved. The fundamental basis of each of the applications listed above is the same despite of the processing result between them. The fundamental step is known as the feature extraction stage of the system, and its objective is to reduce and characterize the data of a signal in a condensed way to be treatable in further processing stages. Classical approaches for feature extraction are based on temporal and frequency representations (i.e. correlation, FFT) which are the basis of more robust approaches like Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), or cepstral representations (Cepstrum). The techniques listed above have been developed through the years taking into account different models of voice production. Such models, have demonstrated that the voice as the output of the vocal tract, has a non-stationary behavior which means that the vocal tract system characteristics also varies over time The Wavelet Packet Decomposition is used in to build an ASR system for spoken words in Malayalam, this approach uses the WP for extracting characteristics coefficients and an information cost function to estimate relevant information of the signals which used for recognition. Taking into account the advantages of the multi resolution analysis using Wavelets, in this paper is proposed a methodology

for characterization and pattern recognition using the robust characteristics extracted by the Wavelet Packet Transform and neural networks to do a recognition system. Since speech signals are non stationary in nature, many parameters affect the speech recognition process. A lot of research work has gone into speech recognition. But there is requirement of much more research and development in this field. Speech recognition system usually involves some kind of classification or recognition based upon speech features. The speech features are usually obtained via time-frequency representations

2.0 LITERATURE REVIEW:

Khaled Daqrouq, (2009) In this work, an average framing linear prediction coding (AFLPC) procedure for content free speaker distinguishing proof frameworks is introduced. Customarily, linear prediction coding (LPC) has been connected in discourse acknowledgment applications. Be that as it may, in this examination the blend of adjusted LPC with wavelet transform (WT), named AFLPC, is proposed for speaker ID. The examination technique depends on highlight extraction and voice grouping. In the period of highlight extraction, the recognized speaker's vocal tract attributes were separated utilizing the AFLPC system. The measure of a speaker's element vector can be improved in term of a satisfactory acknowledgment rate by methods for genetic algorithm (GA). Henceforth, a LPC request of 30 is observed to be the best as per the framework execution.

Birsel Ayrulu-Erdem, (2011) We extract the informative features of gyroscope signals using the discrete wavelet transform (DWT) decomposition and provide them as input to multi-layer feed-forward artificial neural networks (ANNs) for leg motion classification. Since the DWT is based on correlating the analyzed signal with a prototype wavelet function, selection of the wavelet type can influence the performance of wavelet-based applications significantly. We also investigate the effect of selecting different wavelet families on classification accuracy and ANN complexity and provide a comparison between them.

Sonia Suuny, (2013) Discourse is the most regular methods for correspondence among people and discourse preparing and acknowledgment are escalated regions of research throughout the previous five decades. Since discourse acknowledgment is an example acknowledgment issue, arrangement is a critical piece of any discourse acknowledgment framework. In this work, a discourse acknowledgment framework is created for perceiving speaker autonomous spoken digits in Malayalam. Voice signals are inspected legitimately from the amplifier. The proposed technique is actualized for 1000 speakers articulating 10 digits each. Since the discourse signals are influenced by foundation clamor, the signs are tuned by expelling the commotion from it utilizing wavelet denoising technique dependent on Soft Thresholding. Here, the highlights from the signs are removed utilizing Discrete Wavelet Transforms (DWT) in light of the fact that they are well appropriate for handling non-stationary signs like discourse.

3.0 Methodology:

Discrete Wavelet Transforms:

A wavelet can be thought of as an extension of the classic Fourier transform to overcome the resolution problem. A wavelet transform works on a multi-scale

basis instead of a single scale time or frequency basis. Wavelet transform decomposes a signal into a set of basic functions called wavelets. The Discrete Wavelet Transform (DWT) is any wavelet transform for which the wavelets are discretely sampled and is a special case of the wavelet transform that provides a compact representation of a signal in time and frequency that can be computed efficiently. It is a relatively recent and computationally efficient technique for extracting information about non-stationary signals like audio. The wavelet transform is a multi- resolutional, multi-scale analysis, which has been shown to be very well suited for speech processing. The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. Pattern recognition rate is improved by this method.

$W(j, K) = \Sigma_j \Sigma_k X(k)\ 2^{-j/2} \Psi\ (2^{-j}n-k)$ (1)

Where Ψ (t) is the basic analyzing function called the mother wavelet. The functions with different region of support that are used in the transformation process are derived from the mother wavelet. DWT is used to obtain a time-scale representation of the signal by means of digital filtering techniques. The original signal passes through two complementary filters, namely low-pass and high-pass filters. In speech signals, low frequency components known as the approximation coefficients h[n] are of greater importance than high frequency signals known as the detail coefficients g[n] as the low frequency components characterize a signal more than its high frequency components. The wavelet decomposition tree is shown in figure 1.

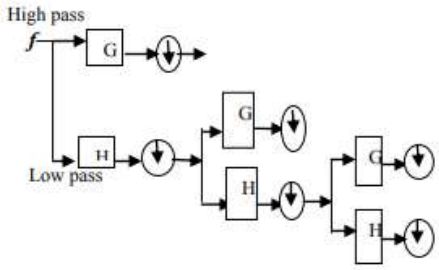


Figure 1. Wavelet decomposition tree

The low frequencies sequence of the first level forms as an input to the second stage. The discrete time domain signal is subjected to successive low pass filtering and high pass filtering to obtain DWT. This algorithm is called the Mallat algorithm. At each decomposition level, the half band filters produce signals spanning only half the frequency band. The filtering and decimation process is continued until the desired level is reached. The main advantage of the wavelet transforms is that it has a varying window size, being broad at low frequencies and narrow at high frequencies, thus leading to an optimal time–frequency resolution in all frequency ranges. The DWT of the original signal is then obtained by concatenating all the coefficients starting from the last level of decomposition. Though it is possible to decompose the high frequency and low frequency components as in the case of wavelet packet decomposition, decomposing only the low frequency components gives better recognition accuracy.

$Y_{high}[k] = \Sigma_n x[n] g[2k-n]$ (2)

$Y_{low}[k] = \Sigma n x[n] h[2k-n]$ (3)

Where Y high (detail coefficients) and Y low (approximation coefficients) are the outputs of the high pass and low pass filters obtained by sub sampling by 2. Here the time resolution is halved, but since the output has half the frequency band of the input, the frequency resolution has been

doubled. The filter analysis block diagram is given in figure 2.

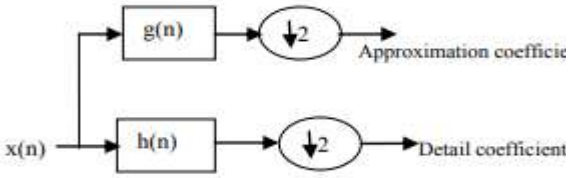


Figure 2. Filter Analysis block diagram

DWT is also related to a multi-resolution framework. Now, DWT is more popular in the field of Digital Signal Processing (DSP) due to its multi resolution capability. Also it has the property of constant Q, which is one of the demands of many signal processing applications, especially in the processing of the speech signals as human's hearing system is constant Q perceptional.

Speech Classification Module: Speech recognition is basically a pattern recognition problem. Pattern recognition deals with mathematical and technical aspects of classifying different objects. Pattern recognition is becoming increasingly important in the age of automation and information handling and retrieval. Since neural networks are good at pattern recognition, many early researchers applied neural networks for speech pattern recognition. During classification stage, decision is taken based on all the similarity measures after trained using information relating to known patterns and the similarity measured from the pattern. In this study also, neural networks are used as the classifier. Neural networks can perform pattern recognition; handle incomplete data and variability well.

Artificial Neural Networks: Neural networks are an artificial intelligence method for modeling complex non-linear functions. Neural networks can be viewed as massively parallel computing systems consisting of an extremely large number of simple processors called nodes with many interconnections. 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. If x1, x2, x3, ……xn are the inputs and w1, w2, w3…wn are the corresponding weights, then the total input to the next neuron or the output neuron I is calculated by the summation function.

The ANN processes information in parallel with a large number of processing elements called neurons and uses large interconnected networks of simple and non linear units. The computational intelligence of neural networks is made up of their processing units, characteristics and ability to learn. During learning the system parameters of NN vary over time and are characterized by their ability of local and parallel computation, simplicity and regularity

$$I = w_1x_1 + w_2x_2 + \ldots\ldots\ldots + w_nx_n = \sum_{i=1}^{n} w_ix_i \qquad (4)$$

The result of the summation function, which is the weighted sum, is transformed to a working output through an algorithmic process called the activation function or the transfer function. The feed-forward network is the most commonly used type of neural network used in the area of pattern classification, which includes multilayer perceptron. In this work, we use architecture of the Multi Layer Perceptron (MLP) network, which consists of an input layer, one or more hidden layers, and an output layer. The algorithm used in this case is the back propagation training algorithm. In this type of network, the input is presented to the network and

moves through the weights and nonlinear activation functions towards the output layer, and the error is corrected in a backward direction using the error back propagation correction algorithm.

Speech Recognition System: The aim of the work presented in this paper is to describe the design criteria and the implementation steps taken into account in the construction of a speech recognition system, based on a wavelet extraction system and neural networks for identification. The processing stages for speech recognition are shown in Figure 3.

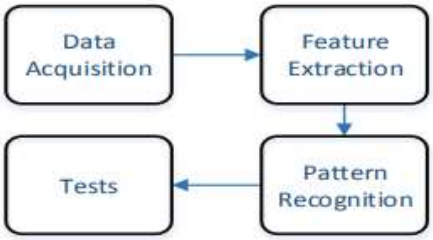


Figure 3: Sub-processing stages for speech recognition.

Figure 1 shows the four fundamental processing phases for the speech recognition. The first three correspond to the signal processing for extracting characteristics and the training of a feedforward neural network for patern recognition. A Non-Audible Murmur microphone, is the transductor used to measure the NAM signals. A STM32F4-Discovery development board is used to perform the data acquirement. Finally, the 13 coefficients are entered to a multilayer perceptron neural network to identify and recognize the uttered words. The recognition word vocabulary (lexicon) in this case corresponds to isolated digits of the Spanish language (zero /cero/, one /uno/). The phoneme units for such task, are defined as complete words since the lexicon is reduced to two isolated digits.

A. Data acquisition: A NAM transductor is a modified electret microphone used to acquire Non-Audible Murmur signals. the primary structure of this device is shown in Figure 4.

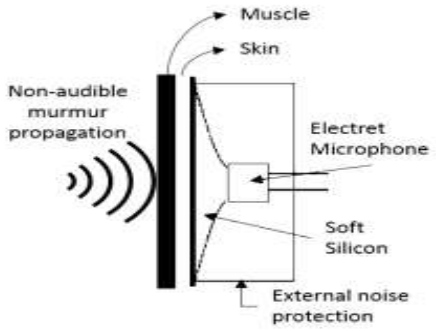


Figure 4: NAM microphone.

The NAM microphone depicted in Figure 4, is connected to the STM32F4-Discovery board to perform the acquisition stage. The STM32F407VG MCU, is configured to perform the analog-to-digital conversion at a frequency rate of 8000 Hz. The voltage resolution of the ADC in this case, is calculated from the operative voltage range of the MCU i.e. 0 V to 3 V. Eq (5) shows how to calculate the ADC resolution.

$$Q = \frac{V_{ref+} - V_{ref-}}{255} \quad (5)$$

Where Q is the resolution in volts, $Vref+$ and $Vref-$ are the upper and lower limits of operative voltage range of the MCU respectively. In this case the resolution Q is 11.76 mV, using a positive voltage of reference equal to 3 V and a negative reference of 0 V. The ADC bit resolution is 255, and was configured in this value to ease the UART data transmission. Figure 5 shows the NAM location in the mastoid process.

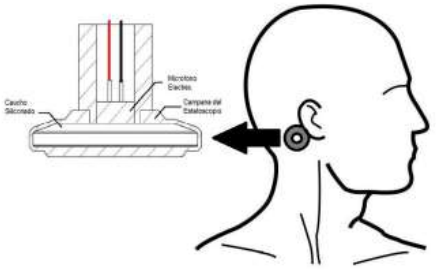


Figure 5: NAM microphone location.

The data of the digitalized signal, is transmitted to the Matlab interface using the UART protocol. The transmission baud rate was set in 9600 bits per second, ensuring a standardized rate. The summarized processing scheme for signal acquisition is depicted in Figure 6.

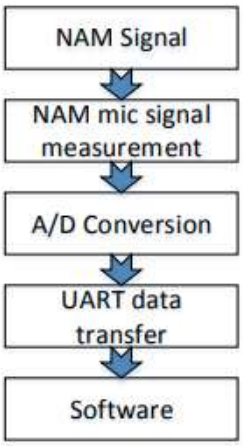


Figure 6: Acquisition and data transmission sub-processes.

B. Pre-processing The pre-processing stage of the recognition system is implemented to isolate speech signals. The result of the isolating process is shown in Figure 7

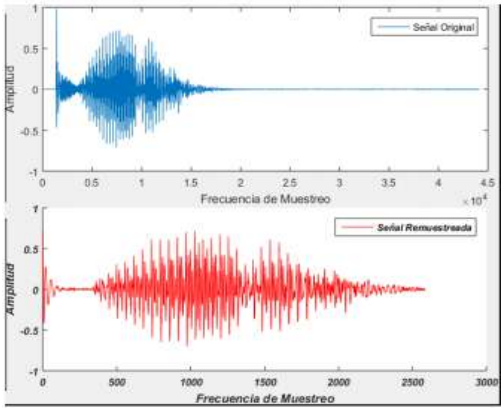


Figure 7: Original signal vs. Isolated signal.

Wavelet Based Feature Extraction System: In general, a feature extraction system for speech processing involves several phases to estimate unique characteristics of the analyzed signals. In this case, the feature extraction phase is carried out with a Wavelet approach. The multi resolution analysis of Wavelets is initiated with the Wavelet transform (WT) (Eq. (6)). The Wavelet transform is defined as the sum in time of the multiplication of the scaled version with the shifted version of the original signal.

$$S(\tau,a)=\int_{-\infty}^{+\infty} S(t)\frac{1}{\sqrt{a}}\varphi*\left(\frac{t-\tau}{a}\right)dt$$

(6)

Where φ is the mother Wavelet conjugate, which is scaled and shifted point to point to estimate the comparation levels with the analyzed signal (t). The value of $a=\frac{f}{f_0}$ from the dilatation of the Wavelet, using $f0$ as the fundamental frequency and τ as the time shift. The WP method is summarized in Figure 8, and is based on the WT definition.

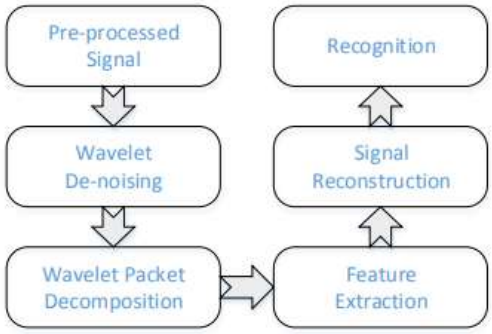


Figure 8: Wavelet Feature Extraction phases.

C. Denoising and Wavelet Packet Decomposition: Denoising of a signal is the rejection of disturbance components that are present on a waveform. Using the filtering characteristics of the Wavelet method, is possible to reject the noise of a signal according to the environmental source and the frequency ranges of the speech (0, 8000 Hz).

The schematic diagram depicted in Figure 8, shows the signal decomposition through high-pass and low-pass filters. For each scale index, a corresponding number of filters which varies in a power of two basis (2) are defined. The original signal $s[n]$, is

splitted in two filtered parts (A1 and A2), which are filtered again to extract new Wavelet Packet vectors, which in turn are filtered until reach a final scale ($j = m$). In Figure 8, the functions $H_{j,}$ and $G_{j,i}$ denote the high-pass and low-pass filters respectively, and produce the new WP vectors for the next scale. Each vector in each depth scale depending of the filtering process is known as an Approximation or Detail (A, D) set of coefficients.

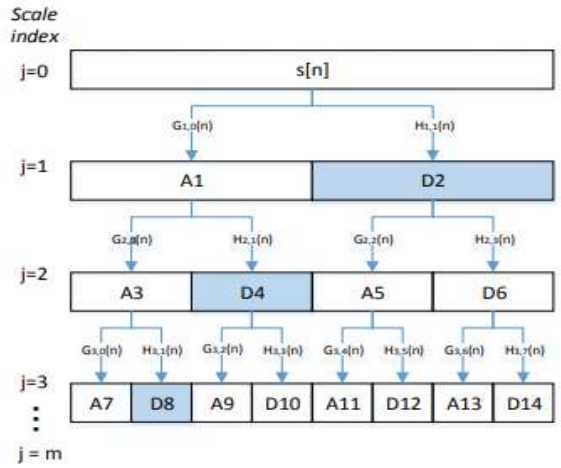


Figure 9: Diagram of Wavelet Decomposition.

As Figure 9 shows, at each scale level the WP vectors are reduced in size at a power of two rate, this leads to extract specific characteristics in specific time intervals, which is similar to the multiresolution analysis performed using the discrete Wavelet transform.

D. Feature Extraction: The feature extraction process for extracting unique data using the WP approach is performed by implementing an information cost function. For this case, such data retention is carried out by the Shannon Entropy measure, which is defined for a discrete sequence si as shown in Eq. 7.

$$E(p) = -\sum_{i} p_i^2 \log(p_i^2)$$
$$i = 0,1, \dots, N-1 \quad (7)$$

Where (s) is the resulted Shannon Entropy, $p_i{}^2$ is the set of normalized energies from the discrete Wavelet packet analyzed and N is the length of the WP vector

The Shannon entropy for the detail WP and the las approximation set is estimated for the speech signals in the process described in this paper. The information measure is the maximum limit of compression of a signal without loss of information, which is a suitable method to acquire feature information of the signal. The final coefficients represented using the Shannon entropy for 6 speech samples are depicted in Figure 9.

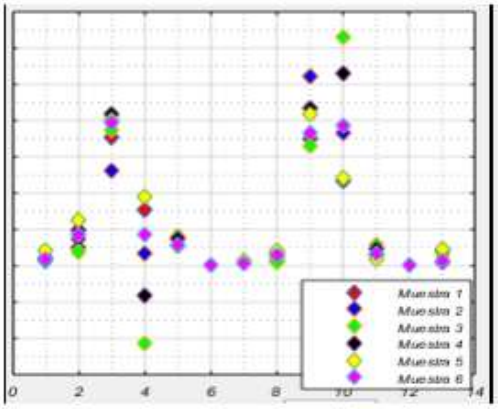


Figure 10: Coefficient representation of 6 speech signals of the utterance of /uno/ (one).

Figure 10, shows 13 coefficients of 6 different signals for the utterance of the word /one/. A correspondence between coefficients is observed and represent unique data of the signals. Taking into account the Eq. 8, the reconstruction of the signal for performed to estimate how well the extracted coefficients represent the signal.

$$\tilde{x} = \tilde{a}_j + \tilde{d}_{j-1} + \cdots + \tilde{d}_1 \quad (8)$$

Where j is the decomposition level and $\tilde{a}$, $\tilde{d}$ are the approximation and detail coefficients respectively.

Pattern Recognition: Pattern recognition is a derivation of the methodologies used

in machine learning, whose objective is to identify regular behavior in data series. These techniques search for unique patterns to classify certain events or categories in order to recognize information. In the development of a recognition system of speech signals for utterances of the "uno" and "cero" digits, the classifier is based on neural networks, and the categories are defined as the digits to recognize with a defined set of characteristics extracted from the WP method.

A feed-forward neural network was used as the classifier. Such network is composed by three layers, the first layer receives the WP characteristic vectors as inputs, the second layer processes the data, and the last layer adjusts the processed values to be used as outputs. A generic representation of the neural network is shown in Figure 11.

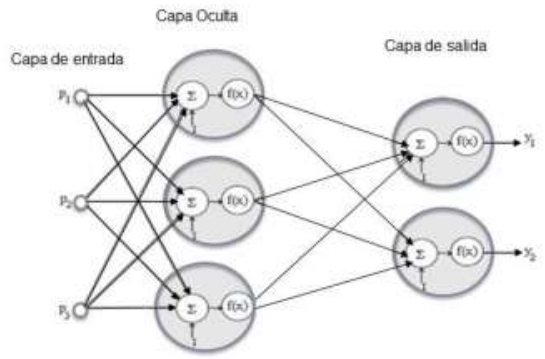


Figure 11: Generic representation of a multilayer neural network

The activation function for the hidden layer is a sigmoidal function, which is represented by Eq. (9).

$$f(n) = \frac{1}{1+e^{-n}} \quad (9)$$

The activation function for the output layer is based on a softmax function, as represented by Eq. (10).

$$g(n) = \frac{e^{n}}{\sum_{l=1}^{k} e^{n}} \quad (10)$$

4.0 RESULTS:

There are different types of wavelet families such as Daubechies, Symmlet, Coiflet etc. Selection of the suitable wavelet and the number of decomposition levels play an important role in obtaining good recognition accuracy in speech recognition. Among the various wavelet bases, the most popular wavelets that represent foundations of Digital Signal Processing called the Daubechies wavelets are used because of its orthogonality property and efficient filter implementation. In this paper, we are using db4 type of mother wavelet feature extraction purpose. The speech samples in the database are successively decomposed into approximation and detailed coefficients. The approximation coefficients from eighth level are used to create the feature vectors for each spoken word and the number of approximation coefficients obtained at the eighth level is twelve.

The feature vectors obtained using DWT are given as the input to the ANN classifier. Here we have divided the database into three. 70% of the data is used for training, 15% for validation and 15% for testing. MLP architecture is used for the classification scenario. Using this network, the classifier could successfully recognize the spoken words. After testing, the corresponding accuracy of the isolated spoken words is obtained. The results obtained clearly shows the efficiency of neural networks in classifying the extracted coefficients. Results obtained using DWT and ANN is given below. The original signal and various decomposition level coefficient values of 5 Malayalam words geetham, thamara, maram, vellam and amma are shown in figure 12 and the performance analysis based on error percentage is given in table 1.

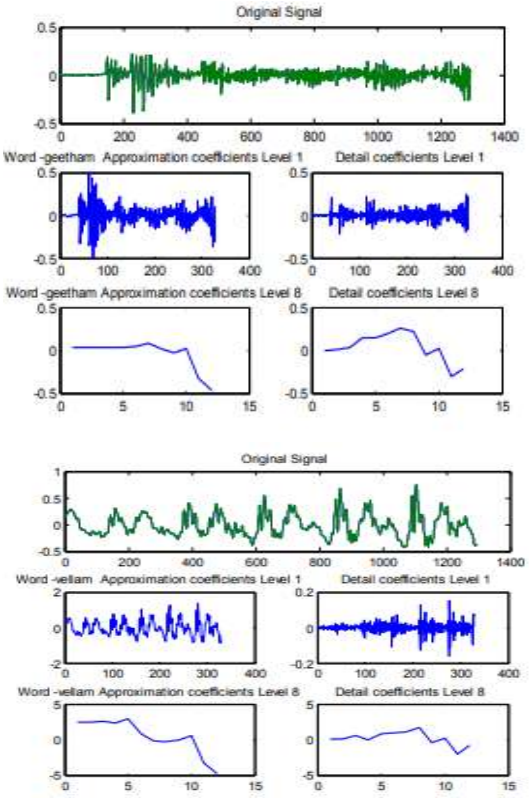
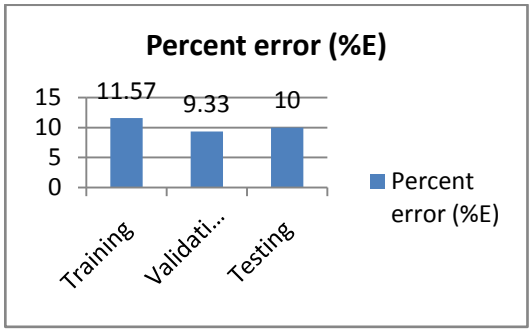


Figure 12. Different decomposition levels of spoken words

Table 1. Performance Analysis based on Error Percentage

	Mean Squared Error (MSE)	Percent error (%E)
Training	5.29349 e-3	11.57
Validation	4.0099 e-3	9.33
Testing	4.67599 e-3	10.00



Graph 1: Error Percentage

By using Discrete Wavelet Transforms as the feature extraction tool and Artificial Neural Networks as classifier, the overall recognition accuracy obtained is 90%.

The feature extraction phase was performed using 6 speech samples of one test subject for each of the categories to recognize; the phoneme units treated as words, and the categories as /cero/ and /uno/ utterances. To validate the speech recognition system, 10 samples were recorded for the utterances of /uno/ and /cero/, the results of recognition are condensed in the Tables 2 and 3.

Table 2: Recognition results for the first five samples.

Utterance	Test 1	Test 2	Test 3	Test 4	Test 5
/cero/	81.15	81.96	68.91	72.10	71.28
/uno/	76.95	75.90	74.24	79.79	75.16

Table 3: Recognition results for the last five samples.

Utterance	Test 6	Test 7	Test 8	Test 9	Test 10
/cero/	71.13	74.99	78.18	74.18	70.86
/uno/	77.59	75.84	76.85	72.84	79.26

From the Table 2 and 3, it is shown that the percentage of recognition has an average of 75.45 %, which is sufficient taking into account the amount of samples used for training and the number of categories that the neural network has to recognize.

4.0 CONCLUSION:

In this paper was portrayed the improvement of a discourse acknowledgment framework utilizing Wavelet techniques. The utilization of recurrence and fleeting portrayals mutually with neural systems, showed to be an

appropriate technique for example acknowledgment. The decay approach dependent on the usage of high-pass and low-pass channels in every one of the profundity dimensions of the multi goals examination, permits performing investigation in time and recurrence to extricate variety attributes of the discourse signals, and the conduct that they have between the oscillatory changes after some time. The example acknowledgment period of the framework, work from the plan of neural systems is an appropriate methodology which brings high acknowledgment rates, that together with a strong component extraction procedure can result in high acknowledgment rates as appeared in Tables 3 and 4. The computational unpredictability and highlight vector estimate is effectively diminished, all things considered, by utilizing discrete wavelet changes. Along these lines a wavelet change is a rich device for the investigation of non-stationary signs like discourse. These analysis results demonstrate that this cross breed design utilizing discrete wavelet changes and neural systems could viably extricate the highlights from the discourse motion for programmed discourse acknowledgment.

REFERENCES:

[1] Lawrence R., 1997, Applications of Speech Recognition in the Area of Telecommunications, Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding.

[2] Kuldeep Kumar, R. K. Aggarwal, 2011, Hindi Speech Recognition System Using Htk, International Journal of Computing and Business Research, Volume 2 Issue 2.

[3] Evandro B. Gouva, Pedro J. Moreno, Bhiksha Raj, Thomas M. Sullivan, Richard M. Stern, 1996, Adaptation and Compensation: Approaches to Microphone and Speaker Independence in Automatic Speech Recognition, Proc. DARPA Speech Recognition Workshop.

[4] Jiang Hai, Er Meng Joo, 2003, Improved Linear Predictive Coding Method for Speech Recognition, ICICS-PCM, Singapur.

[5] S. Mallat, A, 1999, Wavelet Tour of Signal Processing (second edition),Academic Press.

[6] S. Kadambe, P. Srinivasan, 1994, Application of Adaptive Wavelets for Speech, Optical Engineering Vol 33(7).

[7] S .G. Mallat, 1989, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol.11.

[8] Elif Derya Ubeyil, 2009, Combined Neural Network model Employing wavelet Coefficients for ECG Signals Classification, Digital signal Processing, Vol 19.