

# Language Learning Using Feature Extraction Based ANN Techniques

Priyanka V. Pote<sup>1</sup>, Prof. V. R. Ingle<sup>2</sup>, Prof. Y. A. Sadawarte<sup>3</sup>

<sup>1</sup> M.Tech, Electronics Department, Bapurao Deshmukh College of Engineering, Sewagram, Maharashtra, India

<sup>2</sup> Professor, Electronics Department, Bapurao Deshmukh College of Engineering, Sewagram, Maharashtra, India

<sup>3</sup> Professor, Electronics Department, Bapurao Deshmukh College of Engineering, Sewagram, Maharashtra, India

## ABSTRACT

All speech recognizers include an initial signal processing front end that converts a speech signal into its more convenient and compressed form called feature vectors. Feature extraction method plays a vital role in speech recognition task. There are two dominant approaches of acoustic measurement. First is in temporal domain approach (parametric) like Linear Prediction, which is developed to closely match the resonant structure of human vocal tract that produces the corresponding sound. Second is frequency domain approach (nonparametric) known as Mel-Frequency Cepstral Coefficients (MFCC). In another approach wavelet transform and wavelet packet tree have been used for speech feature extraction in which the energies of wavelet decomposed sub-bands have been used in place of Mel filtered sub-band energies.

**Keyword:-** Speech samples, Discrete Wavelet Transform, Linear Predictive Coding.

## 1. INTRODUCTION:

With the ever increasing power and falling cost of the digital signal processors, and availability of cheap memory chips, speech processing systems are mostly used for voice communication and recognition. Voice recognition system includes hands free input system for voice dialing, voice activated security systems etc. Presence of background noise and other types of disturbances also makes a speech processing system complex and difficult. The performance of a speech processing system is usually measured in terms of recognition accuracy. All speech recognizers include an initial signal processing front end that converts a speech signal into its more convenient and compressed form called feature vectors. Feature extraction method plays a vital role in speech recognition task. There are two dominant approaches of acoustic measurement. First is in temporal domain approach (parametric) like Linear Prediction, which is developed to closely match the resonant structure of human vocal tract that produces the corresponding sound. Second is frequency domain approach (nonparametric) known as Mel-Frequency Cepstral Coefficients (MFCC). In another approach wavelet transform and wavelet packet tree have been used for speech feature extraction in which the energies of wavelet decomposed sub-bands have been used in place of Mel filtered sub-band energies. However, the time information is lost due to use of wavelet sub-band energies. The speech is a non-stationary 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. The windowed short-time FT (STFT) provides the temporal information about the frequency content of signal. A drawback of the STFT is its fixed time resolution due to fixed window length. The wavelet transform, with its flexible time-frequency window, is an appropriate tool for the analysis of non-stationary signals like speech which have both short high frequency bursts and long quasi-stationary components also. In speech signal, high frequencies are present very briefly at the onset of a sound while lower frequencies are present latter for long period. DWT resolves all these frequencies well. The DWT parameters

contain the information of different frequency scales. This helps in getting the speech information of corresponding frequency band. In order to parameterize the speech signal, the signal can be decomposed into four frequency bands uniformly or in dyadic fashion. Artificial Neural Network (ANN) is an efficient pattern recognition mechanism which simulates the neural information processing of human brain. 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. Multi Layer Perceptron (MLP) architecture is used for pattern classification in this work. The MLP architecture consists of one or more hidden layers. A signal is transmitted in the one direction from the input to the output and therefore this architecture is called feed forward. The MLP networks are learned with using the Backward Propagation algorithm and is widely using in machine learning applications. MLP uses hidden layers to classify successfully the patterns into different classes. The inputs are fully connected to the first hidden layer, each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs. A wavelet transform is an elegant tool for the analysis of non-stationary signals like speech. The results have shown that this hybrid architecture using discrete wavelet transforms and neural networks could effectively extract the features from the speech signal for automatic speech recognition.

### **1.1 Motivation:**

The predominant mode of human communication for every day interaction is speech and it will also be the preferred mode for human computer interaction. Spoken language has been focused on its use as a human computer interaction mostly for information access and extraction. There is a need in spoken language not as a access of information but also source of information that would make language more important in respect of accessing, sorting, editing, translation etc. As speech signals are non-stationary in nature, speech recognition is a complex task due to the differences in gender, emotional state, accent, pronunciation, articulation, nasality; pitch, volume, and speed variability in people speak. Presence of background noise and other types of disturbances also makes a speech processing system complex and difficult. The performance of a speech processing system is usually measured in terms of recognition accuracy. Speech processing is useful for various applications such as mobile applications, healthcare, automatic translation, robotics, video games, transcription, audio and video database search, household applications, language learning applications etc.

## **2. LITERATURE REVIEW:**

### **2.1 Simulation Voice Recognition System for controlling Robotic Applications**

[1]In this paper, support vector machines (SVMs) have proven to be a powerful technique for pattern classification. SVMs map inputs into a high dimensional space and then separate classes with a hyperplane. A critical aspect of using SVMs successfully is the design of the inner product, the kernel, induced by the high dimensional mapping. We consider the application of SVMs to speaker and language recognition. A key part of the approach is the use of a kernel that compares sequences of feature vectors and produces a measure of similarity. Sequence kernel is based upon generalized linear discriminants, shows that this strategy has several important properties. First, the kernel uses an explicit expansion into SVM feature space, this property makes it possible to collapse all support vectors into a single model vector and have low computational complexity. Second, the SVM builds upon a simpler mean-squared error classifier to produce a more accurate system. Finally, the system is competitive and complimentary to other approaches, such as Gaussian mixture models (GMMs).

### **2.2 A novel voice conversion approach using admissible wavelet packet decomposition**

[2]In this paper, automatic Emotion Recognition (AER) from speech finds greater significance in better man machine interfaces and robotics. Speech emotion based studies closely related to the databases used for the analysis. We have created and analyzed three emotional speech databases. Discrete Wavelet Transformation (DWT) was used for the feature extraction and Artificial Neural Network (ANN) was used for pattern classification. We can find that recognition accuracies vary with the type of database used. Daubechies type of mother wavelet was used for the experiment. Overall recognition accuracies of 72.05 %, 66.05% and 71.25% could be obtained for male, female and combined male and female databases respectively.

### **2.3 DWT and LPC based feature extraction methods for isolated word recognition**

[3]In this paper a new efficient feature extraction methods for speech recognition have been proposed. The features are obtained from Cepstral Mean Normalized reduced order Linear Predictive Coding (LPC) coefficients derived from the speech frames decomposed using Discrete Wavelet Transform (DWT). LPC coefficients derived from wavelet-decomposed sub-bands of speech frame provide better representation than modelling the frame directly. Experimentally it has been shown that, the proposed approach provides effective (better recognition rate), efficient (reduced feature vector dimension) features. The speech recognition system using the Continuous Density Hidden Markov Model (CDHMM) has been implemented. The proposed algorithms were evaluated using isolated Marathi digits database in presence of white Gaussian noise.

#### **2.4 Context Dependent State Tying For Speech Recognition Using DeepNeural Network Acoustic Models**

[4] This paper proposes an algorithm to design a tied-state inventory for a context dependent, neural network-based acoustic model for speech recognition. Rather than relying on a GMM/HMM system that operates on a different feature space and is of a different model family, the proposed algorithm optimizes state tying on the activation vectors of the neural network directly. Experiments show the viability of the proposed algorithm reducing the WER from 36.3% for a context independent system to 16.0% for a 15000 tied-state system.

#### **2.5 Convolutional Neural Networks for Distant Speech Recognition**

[5]In this paper, authors investigated using CNNs for DSR with single and multiple microphones. A CNN trained on a single distant microphone is found to produce a WER approaching these of a DNN trained using beam forming across 8 microphones. In experiments with multiple microphones, we compared CNNs trained on the output of a delay-sum beam former with those trained directly on the outputs of multiple microphones. In the latter configuration, channel-wise convolution followed by a cross channel max-pooling was found to perform better than multichannel convolution. We have explored different weight sharing approaches and propose a channel-wise convolution with two-way pooling. Our experiments, using the AMI meeting corpus, found that CNNs improve the word error rate (WER) by 6.5% relative compared to conventional deep neural network (DNN) models and 15.7% over a discriminatively trained Gaussian mixture model (GMM) baseline. For cross-channel CNN training, the WER improves by 3.5% relative over the comparable DNN structure. Compared with the best beam formed GMM system, cross-channel convolution reduces the WER by 9.7% relative and matches the accuracy of a beam formed DNN.

#### **2.6 Neural Networks used for Speech Recognition**

[6]This paper is showing that neural networks can be very powerful speech signal classifiers. A small set of words could be recognized with some very simplified models. The pre-processing quality is giving the biggest impact on the neural networks performance. In some cases where the spectrogram combined with entropy based endpoint detection is used we observed poor classification performance results, making this combination as a poor strategy for the pre-processing stage. On the other hand we observed that Mel Frequency Cepstrum Coefficients are a very reliable tool for the pre-processing stage, with the good results they provide. Both the Multilayer Feed forward Network with back propagation algorithm and the Radial Basis Functions Neural Network are achieving satisfying results when Mel Frequency Cepstrum Coefficients are used.

### **3. OVERALL ANALYSIS OF RESEARCH WORK:**

The objective of our project is to implement speech recognition system using wavelet transform and artificial neural network (ANN) for isolated English digits. The objective of our project is to implement speech recognition system using wavelet transform and artificial neural network (ANN) for isolated English digits. Our contribution includes firstly deriving features from the frequency sub-bands of the frame using discrete wavelet transform. Secondly each frame of speech signal is decomposed into different frequency sub-bands using discrete wavelet transform. Thirdly classification of each sub-band using artificial neural network (ANN). Finally determination of accuracy of speech recognition system.

#### **3.1 PROBLEM DEFINITION / FORMULATION:**

Best recognition was found from the DWT decomposition when compared to the MFCCs for speaker independent and speaker dependent tasks respectively. Wavelet transform approaches provided good results in clean, noisy and reverberant environments and also has a much lower computational complexity. Wavelet decomposition results in a logarithmic set of bandwidths, which is very similar to the response of human ear to frequencies. Wavelet transform efficiently locates the spectral changes in speech signal as well as beginning and end of the sounds can also be located. Results show that hybrid architecture using discrete wavelet transforms and neural networks could

effectively extract the features from the speech signal for automatic speech recognition. Artificial neural network performance depends on the size and quality of training samples. The simplification of the ANN architecture without reducing the recognition rate can also speed up the recognition time.

### 3.2 OBJECTIVES

- To derive features from the frequency sub-bands of the frame using discrete wavelet transform.
- Each frame of speech signal is decomposed into different frequency sub-bands using discrete wavelet transform.
- Classification of each sub-band using artificial neural network (ANN).
- Determination of accuracy of speech recognition system.

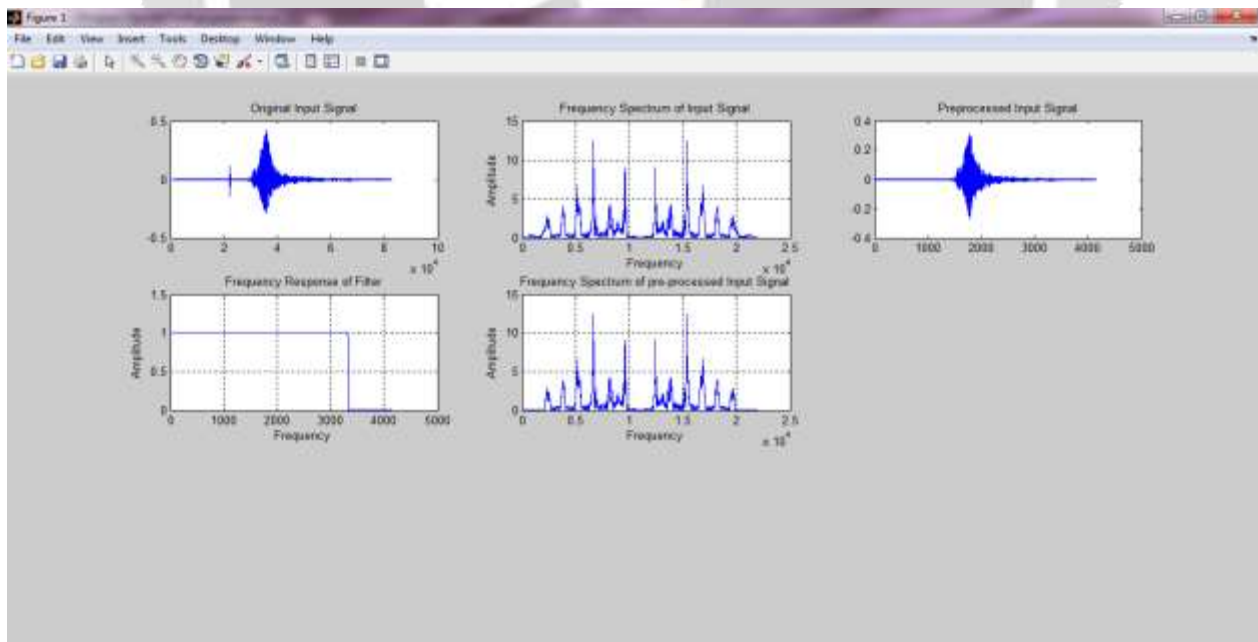
### 3.3 PROPOSED METHODOLOGY

- Recording of speech signal.
- Pre-processing of input speech signal for enhancement.
- Decomposition of speech signal using wavelet transforms.
- Feature extraction of speech signal.
- Classification of features using ANN.
- Performance evaluation.
  - Accuracy.

### RESULT:

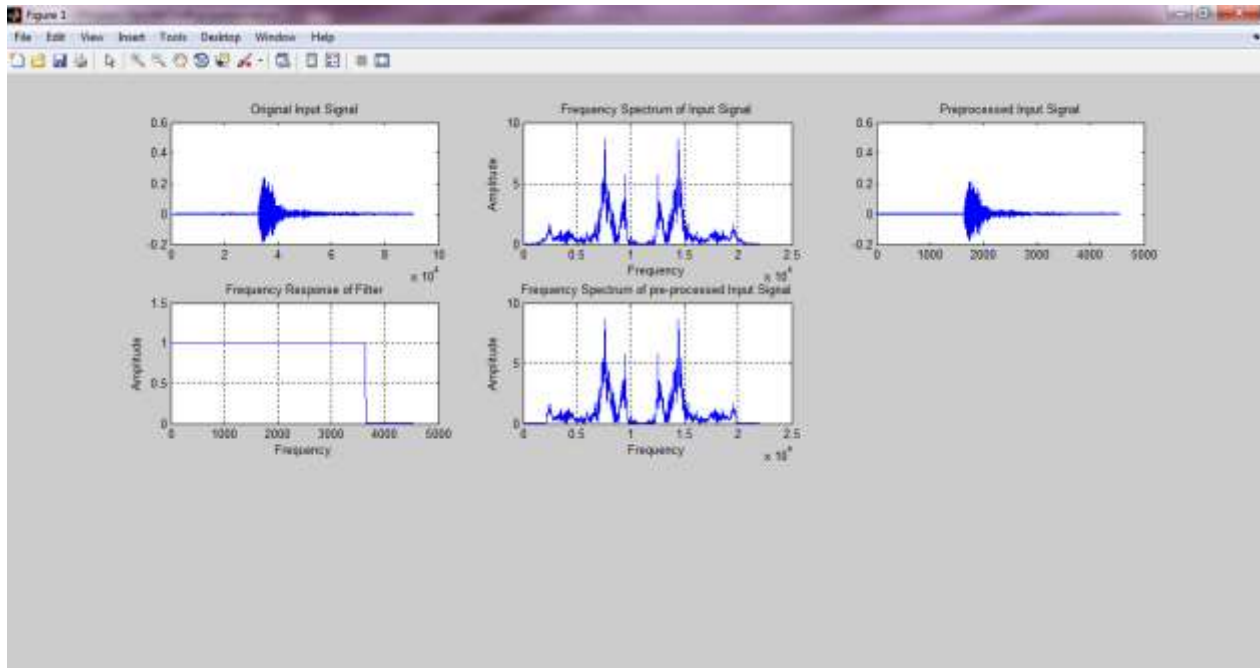
Results can be tested for the English digits from one to nine. Following is the outputs of the above explained methodology. The output results shows that the input speech signal is first pre-processed. In pre-processing input signal is decimated by 20 and Low Pass Filter is applied. After pre-processing 3-level DWT is applied. After applying Discrete Wavelet Transform the Linear Predictive Coding is applied. For better results Auto Regression Model is also applied.

OUTPUT FOR “ONE”:

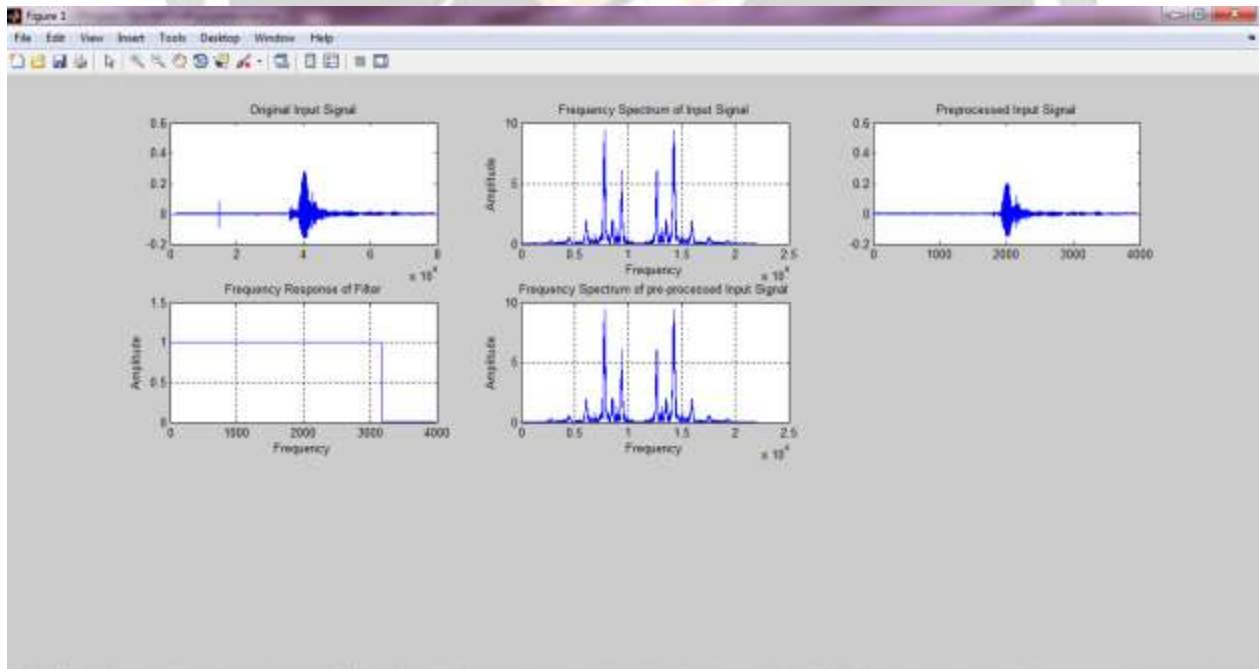


OUTPUT FOR “TWO”:





OUTPUT FOR “THREE” :




#### 4. CONCLUSIONS

The effective speech features extraction and classification would improve quality of the speech, represent speech signal in terms of frequency and bandwidth and improve speech recognition.

## 5. REFERENCES

- [1]. Wahyu Kusuma R., Prince Brave Guhyapati V., “Simulation Voice Recognition System for controlling Robotic Applications”, Journal of Theoretical and Applied Information Technology, vol.39, no.2, pp. 188-196, May 2012.
- [2]. Jagannath H Nirmal, Mukesh A Zaveri, Suprava Patnaik and Pramod H Kachare, “A novel voice conversion approach using admissible wavelet packet decomposition”, Springer EURASIP Journal on Audio, Speech, and Music Processing, pp 1 – 10, 2013.
- [3]. N. S. Nehe, R. S. Holambe, “DWT and LPC based feature extraction methods for isolated word recognition”, Springer EURASIP Journal on Audio, Speech, and Music Processing, vol.2012, pp.1-7, 2012.
- [4]. Michiel Bacchiani, David Rybach, “Context Dependent State Tying For Speech Recognition Using DeepNeural Network Acoustic Models”, inIEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.
- [5]. Pawel Swietojanski, Arnab Ghoshal, Steve Renals, “Convolutional Neural Networks for Distant Speech Recognition”, in IEEE Signal Processing Letters, Vol. 21, No. 9, September 2014, pp 1120 – 1124.
- [6]. Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, “Neural Networks used for Speech Recognition”, in Journal Of Automatic Control, University Of Belgrade, Vol. 20, 2010, pp 1 – 7.

## BIOGRAPHIES

	<p>Priyanka V. Pote</p> <p>Student: M. Tech (Electronics)</p> <p>Bapurao Dekhmukh College Of Engineering, Sewagram, Wardha.</p> <p>Email-id: <a href="mailto:potepriyanka@gmail.com">potepriyanka@gmail.com</a></p>
--	---