# Speech Recognition using Neural Network

Pankaj Rani
BGIET, Sangrur

Sushil Kakkar
BGIET, Sangrur

Shweta Rani
BGIET, Sangrur

## ABSTRACT
Speech recognition is a subjective phenomenon. Despite being a huge research in this field, this process still faces a lot of problem. Different techniques are used for different purposes. This paper gives an overview of speech recognition process. Various progresses have been done in this field. In this work of project, it is shown that how the speech signals are recognized using back propagation algorithm in neural network. Voices of different persons of various ages in a silent and noise free environment by a good quality microphone are recorded. Same sentence of duration 10-12 seconds is spoken by these persons. These spoken sentences are then converted into wave formats. Then features of the recorded samples are extracted by training these signals using LPC. Learning is required whenever we don't have the complete information about the input or output signal. At the input stage, 128 samples of each sentence are applied, then through hidden layers these are passed to output layer. These networks are trained to perform tasks such as pattern recognition, decision making and motoric control.

## Key words
Neural network, speech recognition, back propagation, training algorithm.

## 1. INTRODUCTION
Speech could be a useful interface to interact with machines. To improve this type of communication, researches have been for a long time. From the evolution of computational power, it has been possible to have system capable of real time conversions. But despite good progression made in this field, the speech recognition is still facing a lot of problems. These problems are due to the variations occurred in speaker including the variations because of age, sex, speed of speech signal, emotional condition of the speaker can cause the difference in the pronunciation of different persons. Surroundings can add noise to the signal. Sometimes speaker causes the addition of noise itself [4]. In speech recognition process, an acoustic signal captured by microphone or telephone is converted to a set of characters. A view about automatic speech recognition (ASR) is given by describing the integral part of future human computer interface. Hence for the interaction with machines human could use speech as a useful interface. Human always want to achieve natural, possessive and simultaneous computing. Elham S. Salam [13] compared the effect of visual features on the performance of Speech Recognition System of disorder people with audio speech recognition system. Comparison between different visual features methods for selection is done and English isolated words are recognized. The recognition of simple alphabet may be taken as a simple task for human beings. But due to the occurrence of some problems like high acoustic similarities among certain group of letters, speech recognition may be a challenging task [11]. The use of conventional neural network of Multi-Layer Perceptron is going to increase day by day. Work is well done as an effective classifier for vowel sounds with stationary spectra by those networks. Feed forward multi-layer neural network are not able to deal with time varying information like time-varying spectra of speech sounds. This problem can be copied by incorporated feedback structure in the network.

### 1.1 Procedure of speech recognition process
Speech recognition is mainly done in two stages named as training and testing. But before these, some basic techniques that are necessary are applied to these speech signals.

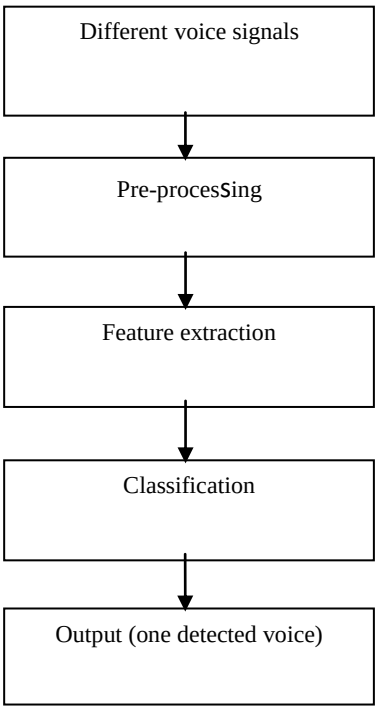



**Fig.1: Block diagram of speech recognition process**

In this process the voice of different persons is recorded by a good quality microphone in such an environment where no noise is present. These speech signals are then pre- processed by using suitable techniques like filtering, entropy based end point detection and Mel Frequency Cestrum Coefficient etc. this type of technique makes the speech signal smoother and helps us in extracting only the required signal that is free of noise.
Samples are recorded with a microphone. Besides speech signals, they contain a lots of distortion and noise because of the quality of microphone. First of all low and high frequency noise is eliminated by performing some digital filtering. The situation of speech signals is mainly between 300Hz to 750Hz.Identical waveforms never produced by recorded samples and the background noise, length and amplitude may vary. 128 samples are applied with sampling rate 11 KHz, this makes possible to represent all speech signals.

### 1.2 Speech Classification
Classification of speech signal is very important phenomenon in speech recognition process. Different models are introduced by different authors to classify the speech. But in this work of project, neural network is to be used for

classification. A neural network consists of small no of neurons. A Number of neurons are interconnected. A Number of processing units which are used for the processing of speech signals. The very simple techniques like pre-processing, filtering are processed by these types of units. A non-linear weight is computed simply by each unit and the result over its outgoing connection to other units is broadcast. Learning is a process in which value of the appropriate weights is settled. It is necessary whenever we don't have the complete information about the input and output signal. The weights are adjusted by the proposed algorithm to match the input and output characteristics of a network with the desired characteristics. The desired response has to be assumed by our self with the help of teacher. In this work of project, features of the pre- processed speech signals are extracted by using MFCC, LPC. This is called training. The networks are usually trained to perform tasks such as pattern recognition, decision making, and motoric control. Training of the unit is accomplished for the adjustment of the weights and threshold for the classification SVM classifier is used. The feature extraction may be of two types as temporal analysis and spectral analysis. In temporal analysis, the wave formats of the speech signals are analyzed by it. In spectral analysis, the wave format of the speech signal is analyzed by the spectral representation. Except all this, there are some other tools that are necessary to study out are linear predictive coding (LPC) and Mel Frequency Cestrum Coefficients (MFCC). LINEAR Predictive Coding is a tool that is used for the processing of audio signal and speech for representation of spectral envelope and digital signal in compressed form. LPC is based on the idea that expression of each sample of signal in a linear combination of the previous samples. Mel frequency cestrum coefficient is preferred to extract the feature of speech signal. It transforms the speech signal into frequency domain, hence training[3] vectors are generated by it. Another reason of using this method is that human learning is based on frequency analysis. Before obtaining the MFCC of a speech signal the pre emphasis filtering is applied to the signal with finite impulse response filter given by

$$H_{pre}(Z) = \sum_{k=0}^{n} a_{pre}(\mathrm{K})\, Z^{-K}$$

Its Z-Transform is

$$H_{pre} = 1 + a_{pre}\ z^{-k}$$

The value of $a_{pre}$ is usually taken between -1.0 to 0.4.

Testing is the process, in which different speech signals are tested by using special type of neural network. This is the main step in the speech recognition process. Testing of the speech signals is done after training.

## 1.3 Speech Recognition Process

Recognition of speech is more difficult than the recognition of the printed versions. Various techniques are to be used for the recognition of speech. Basic procedure is shown by the block diagram. It is shown that how speech can be recognized using different processes.

Speech is used effortlessly by humans as a mode of communication with one another. Same type of easy and natural communication is wanted with machines by people. So, speech is preferred as an interface rather than using any other interfaces like mouse and keyboard. The speech recognition process is somewhere difficult and complicated phenomenon. The speech recognition system can further be divided into various classes. It may be classified based on the model of speaker and type of vocabulary.

This figure shows the general procedure of the speech recognition process. Typical speech sentence consist of two main parts; speech information is carried out by one part and silent and noise sections between the utterances without any verbal information is carried out by the other part. At the input side, different voice signals are applied. Before applying these signals to the neural network, pre- processing of the signals is done by using filtering; Entropy based end point detection and MFCC. The audio signals are converted into particular waveforms. The next step is to extract the features of the voice signals by the of special kind of neural network. Neural Network acts as the brain of human. Trained neural networks trains these networks and at the last testing of voice signals is done. Tested signal is detected as the output. All the working procedure is shown in the block diagram of speech recognition process how the steps take place.

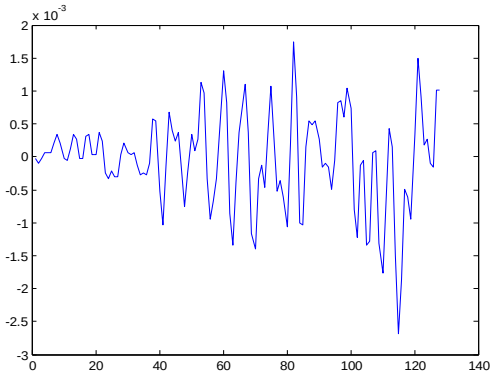
## 1.4 Voice Individuality

Before trying to solve the problem described in the goal of project, we must understand the characteristics of the different voice signals. Acoustic parameters have the greatest influence on the voice individuality. Acoustic parameters may be divided into two types: time dimensions that represent the pitch frequency or fundamental frequency and in frequency dimensions that represent the vocal tract resonance. We can consider the voice signals as quasi periodic signals. Pitch may be defined as the fundamental frequency of the voice signal. The average pitch speed, time pattern, gain and fluctuation change from one individual to another and also within the speech of the same speaker.in actual the frequency response of the vocal tract filter is the shape and gain of the spectral envelope of the signal. From some researches on voice individuality, it has been concluded that pitch fluctuation that gives the second place to the format frequencies is the most important factor in the voice individuality. From many other studies it also be concluded that the spectral envelope has the greatest influence on the voice individuality perception.

From the above discussion it is concluded that there is no single parameter that can alone define a speaker. A group of parameters that depend on the nature of speech materials vary from one individual to another having their respective importance.

## 2. CONVERSION OF SPEECH SIGNALS INTO WAVES

The samples of the speech signals are converted into wave formats. This is the most general way for the representation of the signal. A disadvantage of this method is also there which is that it cannot represent speech related information. This problem may be solved by the technique pre-processing. This representation shows the change in amplitude spectra over time. There are three dimensions. X-axis represents time in meter per second. Y-axis represents frequency and z-axis represents the color intensity that represents the magnitude of the signal. It is not possible to start samples exactly at the same time because different persons pronounce sentences differently i.e. slowly or fast and as the result intensities at the different times might be different.




**Time (m/sec)**

**Fig.2 Wave format of speech signal of a male**

This figure shows the wave format of a speech signal obtained in the implementation in the MATLAB. The change in amplitude spectra over time is shown by the domain representation. The complete sample is split into two time frames with almost 50% overlap. We calculate the short term frequency for each time. Although good visual representation of speech signal is provided by the spectrogram, verification between samples is still there. Samples never start exactly at the same time, sentences are pronounced differently by different persons slower or faster and as a result that might have different intensities at different times.

### 2.1 Algorithm Used

Here the algorithm used is the back propagation. The back propagation algorithm was originally introduced in the 1970s. Several neural networks are described here, in which back propagation works faster than the earlier approaches used for the learning phenomenon. It makes possible to use neural network to solve problems which have not been solved for a long time. In today's world, the back propagation algorithm is the work horse of learning in neural network. Besides this, it gives us detailed insights into how changing the weights and how the overall behavior of a network is changed by the biases.

### 2.2 Roll Played by Neural Network in Speech Recognition Process

Neural network works as a human brain. These networks perform learning phenomenon. Neural network is a computational model inspired by an annual central nervous system which is capable of machine learning as well as pattern recognition the artificial neural networks are generally presented as systems in which number of neurons are interconnected which have been used to solve a wide variety of tasks that are difficult to solve using ordinary rule based programming including computer vision, speech recognition. This type of network potentially contains a large number of simple processing units, roughly analogous to neurons in the brain. All these units are operated simultaneously. Except neural network, there is no other processor that oversees their activity. These units perform all computations in the system. A scalar function is computed simply by each unit and the result is broadcasted to its neighboring units. The course of dimensionality problem that many attemptations to model non- linear functions with large number of variables is also kept in check by neural network. Representative data is collected by neural network users and then training algorithms are invoked because they have to learn the structure of the data automatically [6]. Lakshami Kanaka [6] described a method for estimating a continues target for training patterns of neural networks that are based on the generalized regression neural network and they compared the performance with the performance of linear and multilayer perceptron.

There are two input units in a network by which data is received from the environment, hidden layers y which transformation of data is represented internally output units whose function is to take decision. It is possible train recurrent neural networks for sequence labeling problems where the input and output alignment is not known by end to end training method such as connectionist temporal classification. Hence neural network plays a great role in recognizing the speech in this work of project.

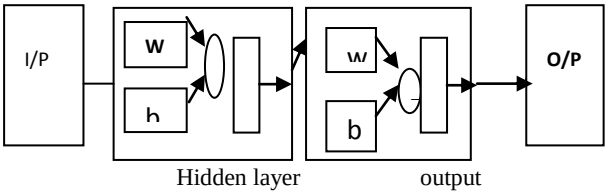



**Fig.3: A Basic Neural Network Used for Training**

This is the neural network with an input, ten hidden layers and one output stage. 128 samples of each sentence are applied. Out of which 70 are used for training and 58 are for testing. All the performance is carried out in MATLAB with coding.When number of words which are to be recognized increases, the number of neurons in hidden layers also have to be increase. The number of neurons required are almost equal to the number of words are too be recognized. Whenever we increase the number of hidden layers, the training time grows sensitively. The quality of the signal pre-processing should be good because performance of the network is mainly dependent on this unit.

## 3. NEURALNETWORK IMPLEMENTATION

Neural networks have been used by many of the authors in the past. For our work of project's implementation, MATLAB neural network toolbox has been used to create, train, and simulate the network. For each sentence, 128 samples are used. From these 128 samples, 70 are used for training while the other 58 are used for testing the network. The trained network can also be tested with real time input from a good quality microphone. Setup of MFCC and neural network for experiment are presented by T.B. Adam. [11]. They took speech data from T146 database isolated alphabet called T1ALPHA. They set the output nodes to nine in order to recognize the nine letters of E-set.

The hidden layers can be calculated by using the formula h=$\sqrt{n * m}$, where n is the number of input nodes and m is the number of output nodes [11].

## 4. RESULTS & DISCUSSION

From the above data, it can said that communication can be module and efficient by speech. Beyond efficiency, speech helps human in making comfortable and familiar with speech. More concentration and restrict movement is required by other modalities due to unnatural positions. Speech is identified by machines by the use of process Automatic Speech Recognition. Feature vector helps in representing each word by conventional method of speech recognition. Artificial Neural Networks (ANN) are biological inspired tools that process the information. Priori information of speech process is not required by artificial neural networks.

Best result is obtained at epoch 4 in this work. 100% accuracy is not achieved in any of the case. The best training performance rate is 2.2596-20 at epoch 4.

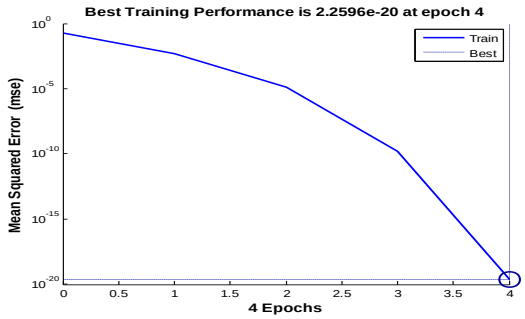



**Fig.4: Best Training Performance Obtained**

It is shown that, in our work of project best training performance is obtained at the epoch 4. We handle the adjustment with a learning rule from which we can also derive a training algorithm for a specific task. Data is trained using neural network toolbox and remaining of the 70 samples are simulated against this trained neural network. The performance of neural network is seen when it runs. The mean square error (MSE) is a network performance function. The performance of the network is measured according to the mean of squared errors.
Mean Square Error is defined as the average squared difference between the output and targets. If zero is obtained, it means no error is there, lower value means result is better. In the above graph, it is shown that the mean square error of the network is starting at large value and decreasing to a small value.

**Table 1. Result obtained of different samples**

| Age group | No of tested samples | passed | Failed | %age |
|---|---|---|---|---|
| 3-10 | 5 | 4 | 3 | 90% |
| 10-20 | 4 | 2 | 2 | 80% |
| 20-30 | 6 | 5 | 1 | 95% |
| 30-40 | 5 | 4 | 1 | 97.23% |
| 40-50 | 5 | 3 | 2 | 86.5% |

Above table shows the performance of speech signals of different persons of various ages including male and female. From the result it is concluded that the better percentage of accuracy is obtained on the recognition of speech signals that are recorded in a closed room than those are recorded in an open room.

## 5. CONCLUSION

From the presented work, it is concluded that neural networks can be very powerful models for the classification of speech signals. Some types of very simplified models can recognize the small set of words. The performance of the neural networks is being impacted largely by the pre- processing technique. On the other hand, it is observed that Mel Frequency Cestrum Coefficients are very reliable tool for the pre-processing stage. Very good results are provided by these coefficients. Satisfying results are achieved by the use of both the Multilayer Feed Forward and Radial basis function neural network with the back propagation algorithm when Mel Frequency Cestrum Coefficients are used.

## 6. REFERENCES

[1] John Paul Hosom, Ram Ad Mark Fanty, "Speech Recognition using Neural Networks" volume.1, July 6 1999.

[2] Antanas Lipeika, Joana Lipeika, Loimutis Telksnys, "development of Isolated Word Speech Recognition System" volume.30, No.1, PP 37-46, 2002.

[3] Ben Gold and Nelsom Margan, "Speech and Audio Processing" Willey addition New Delhi, 2007.

[4] Wouter Geuaert, Georgi Tsenav, Valeri Mladenov, "Neural Network used for Speech Recognition" Journals Automatic Control, volume.20.1.7, 2010

[5] Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yannawar, "A Review of Speech Recognition Technique" IJCA, volume.10, No.3, November 2010.

[6] Lakshami Kanaka, Venkateswarlu Revada, Yasautcha Kumari Rambatla and Koti Verra NagayaAnde, volume.8, issue. 2, March 2011, IJCST.

[7] Nidhi Srivastava, "Speech Recognition using Artificial Neural Networks" volume.3, issue.3, may 2014, IJEST.

[8] Dr. R. L. K. Venkates, Dr. R. Vasantcha Kumari, G. Vani Jayasatu, "Speech Recognition using A Radial Basis Function Neural Networks" volume.3, PP 441-445, April 2011, 3rd INC on E computer technique IEEE.

[9] Vansantha Kumari, G.Vani, Dr. R. L. K. Vankateswarlu, Dr. R. Jayasar, "Speech Recognition by using Recurrent Neural Network" IJSER volume.2, issue.6, June 2011.

[10] Abdul Syapiq B Abdul Sukor, "Speaker Identification System using Mel Frequency Cestrum Coefficient Procedure and Noise Reduction Method" Master Thesis, University Tun Hussein Onn Malaysia, January 2012.

[11] T. B. Adam, Md Salam, "Spoken English Alphabet Recognition with MFCC AND Back Propagation Neural Network" IJCA, volume.42, No.12, March 2012.

[12] R. B. Shinde, Dr. V. P. Pawar, "Vowel Classification Based on LPC and ANN" IJCA, volume.50, No.6, July 2012.

[13] Elhan S. Salam, Reda A. El-Khoribi, Mahmoud E. Shoman, "Audio Visual Speech Recognition For People with Speech Disorder" volume.96, No.2, June 2014.