A Novel Approach for Multilingual Speech Recognition with Back Propagation Artificial Neural Network

Rajat Haldar
Electronics & Telecommunication Department
RCET Bhilai(C.G.) India
haldarrajat12@gmail.com

Dr. Pankaj Kumar Mishra
Electronics & Telecommunication Department
RCET Bhilai(C.G.) India
pmishra1974@yahoo.co.in

Abstract- "Speech Recognition" of audio signal is important for telecommunication, language identification and speaker verification. Robust Speech Recognition can be applied to automation of houses, offices and telecommunication services. In this paper Speech Recognition & Language Identification have done for Bengali, Chhattisgarhi, English and Hindi speech signals. The Bengali, Chhattisgarhi, English, Hindi speech signals are "Ekhone Tumi Jao", "Ae Bar Teha Ja", "Now This Time You Go" and "Ab Is Bar tum Jao" respectively. This method is mainly applied in two phases, in the first phase Speech Recognition and Language identification have done with Back Propagation Artificial neural Network (BPANN) and in the second phase Speech Recognition and Language Identification have done with the combination of the Particle Swarm Optimization (PSO) feature selection technique and BPANN. For the feature extraction Mel Frequency Cepstral Coefficients (MFCC) & Linear Predictive $^{\text{Coding}}$ (LPC) is used. MFCC and LPC are the most widely used feature extraction method. BPANN is a feed forward type neural network, it can trace back the error signal for weight modification, error signal generates when the actual output value differs from the target output value. The system accuracy and performance is measured on the basis of "Recognition Rate" and amount of error. Multilingual Speech Recognition and Language Identification with PSO feature selection technique gives the better Recognition Rate as compare to the without PSO feature selection technique.

Keywords— Multilingual Speech Recognition, Language Identification, Linear Predictive Coding, Mel Frequency Cepstrum Coefficients, Artificial Neural Network, Back Propagation Artificial Neural Network, Automatic Speech Recognition, Particle Swarm Optimization

*****

1.INTRODUCTION

1.1 Speech Recognition

Speech recognition (SR) is the method in which the speech signals of some languages and from multiple speakers is recognized with soft computing techniques. In the soft computing technique mainly ANN is used. For the speech recognition preprocessing, feature extraction, ANN training and ANN testing is the necessary process. It is also known as "automatic speech recognition" (ASR). Speech Recognition has done for many languages in the past. Automatic Speech Recognition (ASR) is a wide research area and it is used by the research community. The interest on "Multilingual Speech Recognition" Systems arises because of there are many languages in India for example Hindi, English, Telugu etc. So for recognition of different audio signals and to use in a proper way we have to adapt the method of Multilingual Speech Recognition. Hence the scholars are taking so much interest in this research area. The percentage of error is difference of actual output and desired output computed on the basis of different technique.

The workability of the system is calculated on the amount of recognition rate. There are several methods are available for Speech Recognition like LPC (linear predictive coding) with back propagation feed forward neural network, SOM (Self Organizing Map) with hybrid ANN/HMM, LVQ (Learning Vector Quantization) with MFCC features, LPCC with ANN etc. The Recognition Rate varies for the different signals and for the different language.

1.2 Multilingual Speech Recognition

"Multilingual Speech Recognition" could be a wide space of research and analysis. In present days there are several languages are utilized in the world wide. English is commonest language and largely use language within the world. Similarly in India many languages are used for communication and to interact with each other. In the education system there are many languages used. This is the reason that Multilingual Speech Recognition is very useful. Speech recognition of more than two languages can be done with soft computing technique like ANN, Fuzzy Logic etc this process is called "Multilingual Speech Recognition." Two or more languages are first analyzed by the system which is depending on system training. After analyze the signal it recognizes by the system in the testing process. The basic process in Multilingual Speech Recognition is feature extraction, training and testing of the speech signals, training and testing is done by ANN. In this paper Back Propagation ANN is used for the training and recognition phase.

1.3 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are the learning models which are inspired by Biological Neural Networks (nervous systems of human being) and are used for function approximation. The ANN is consists of many nodes which is also called the processing elements. The nodes are interred connected with weighted connection. The ANN can be two layer or multilayer. In the two layers ANN there are input unit and the output unit. In the multilayer ANN there are input unit, hidden unit and output unit. At first the input fed into the hidden unit further it goes to the output unit for final processing step. No direct connection is available between input and output in multilayer ANN. In the single layer ANN there is a direct connection between the input and output layer. The ANN is widely used for Speech Recognition, Pattern Recognition, and Computer Applications etc. There are three essentials part of the ANN which are ANN architecture, ANN learning and ANN testing. In the ANN architecture there are many types of ANN available, they are Back propagation ANN, Radial Basis Function (RBF) ANN, Recurrent Artificial Neural Network, Self Organizing Map (SOM) etc. The ANN learning methods are supervised learning and unsupervised learning. In the unsupervised learning the output is not known and in the supervised learning the output is known. In the ANN testing we have to compare the ANN output with desired output. If the ANN output is not same to the desired output then training of ANN is further required and vice versa.

The simple ANN is consists of three layers which are input layer, hidden layer and output layer. Many inputs can be fed to the input unit. The input layer has a unidirectional connection with hidden layer and hidden layer has a unidirectional connection with the output layer. The ANN can be consists of many nodes or processing elements. The diagram of ANN is shown in Figure 1.

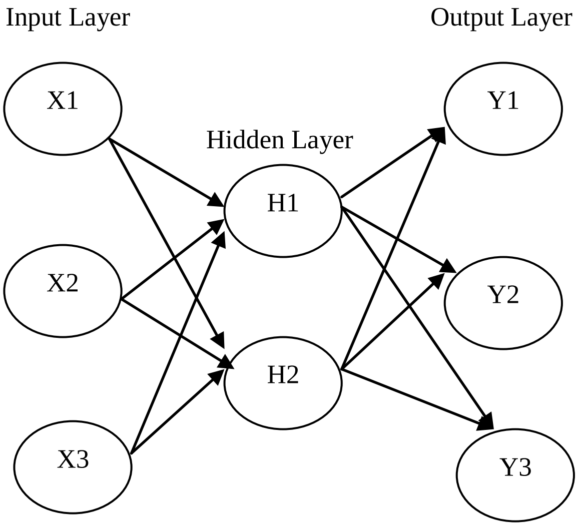

Figure 1.Artificial Neural Network

Various methods have been used for the speech recognition in literature; similarly for language identification many methods are adopted. Back Propagation Neural Network [1] with LPC [1] feature is used for English Alphabet recognition, Spiking Recurrent Self Organizing [2] used for the speech recognition with MFCC [2, 11] features. Time Delay Neural Network (TDNN) [3], hybrid ANN/HMM [4] model is also used for speech recognition which is the combination of Artificial Neural Network (ANN) and Hidden Markov Model (HMM) , in this both method MFCC features is used. Convolution Neural network [5] and Perceptron network [7] is also used for speech recognition. For speech recognition and speaker identification ANN [11] is used. Language Identification [12] has been done for the Bosque context by applying hybrid ANN/HMM model.

This paper proposed Multilingual Speech Recognition and Language Identification with BPANN, PSO feature selection & without PSO feature selection. The paper is summarized as follows methodology is given in section 2, section 3 is result and discussion, section 4 is conclusion and future scope.

2. METHODOLOGY

After observing the importance of the Multilingual Speech Recognition in many types of fields it is necessary that the research work should be expand for some more languages. In past the speech recognition has done for single language, two languages, English alphabets, English digits and number etc. The main objective proposed in this work is divided into two phases and comparison between these two. First phase is speech recognition of Bengali, Chhattisgarhi, English and Hindi speech signal with Artificial Neural Network and the second phase is speech recognition of Bengali, Chhattisgarhi, English and Hindi speech signal with the combination of Particle Swarm Optimization (PSO) feature selection and Artificial Neural Network. After applying these two methods the comparison has done based on recognition rate and error. In this proposed work Back Propagation Artificial Neural Network (BPANN) is used for speech recognition.

Language Identification is also performed for Bengali, Chhattisgarhi, English and Hindi speech signal. The database is same for Language Identification and speech recognition. In the proposed work Language Identification has done in two phases and comparison between these two phases. First phase is Language Identification with feature extraction, ANN training and testing is performed and the second phase is Language Identification with feature extraction, particle swarm optimization technique, training and testing is performed. Based on these two methods the results are compared. In language Identification the different signal is loaded to MATLAB for processing, after the processing it gives the output that the input signal belongs to which language. Performance of the system is measured on the basis of Recognition Rate and the error. The complete methodology is shown in following flow chart.

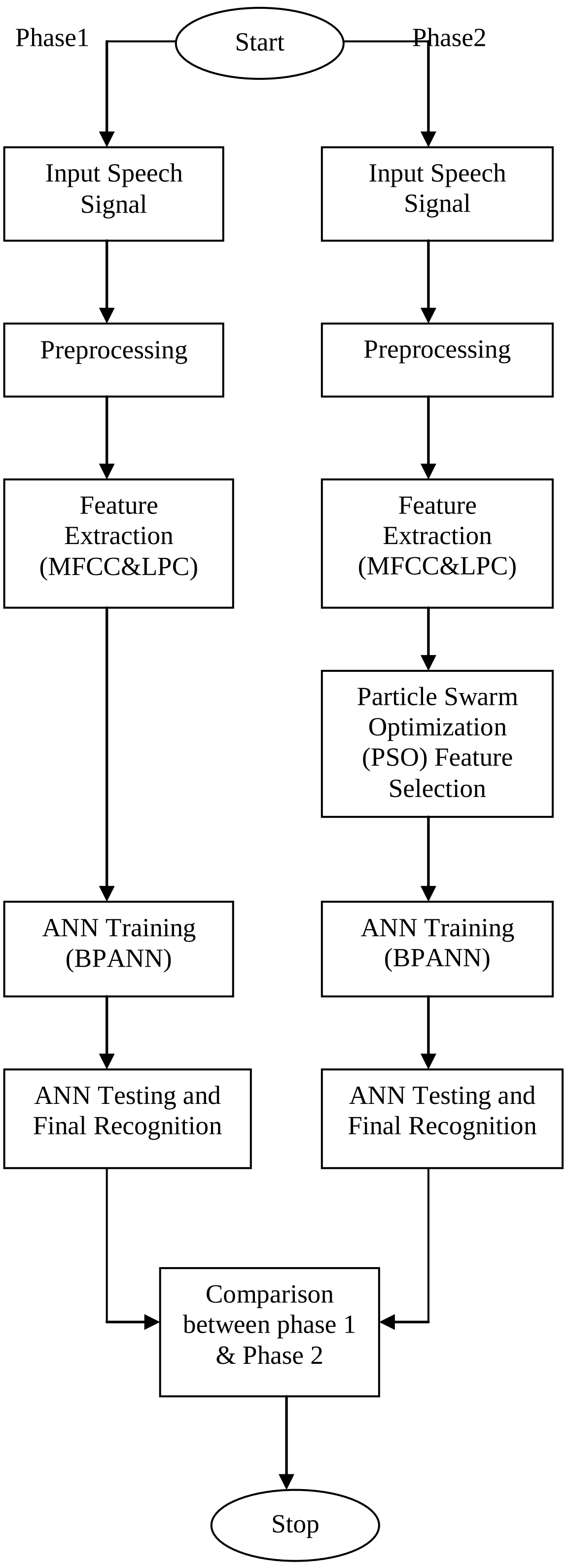
Phase1
Start
Phase2
Input Speech Signal
Input Speech Signal
Preprocessing
Preprocessing
Feature Extraction (MFCC&LPC)
Feature Extraction (MFCC&LPC)
Particle Swarm Optimization (PSO) Feature Selection
ANN Training (BPANN)
ANN Training (BPANN)
ANN Testing and Final Recognition
ANN Testing and Final Recognition
Comparison between phase 1 & Phase 2
Stop

Fig2. Flow chart of the Methodology

The flow chart of the methodology in given in Fig2, the methodology is divided in mainly two phases. In first phase MFCC and LPC feature extraction is used of input speech signals after preprocessing, then the training and recognition is done by the BPANN. In second phase MFCC and LPC feature extraction is used of input speech signals after preprocessing, after extraction of the features value Particle Swarm Optimization (PSO) feature selection technique is applied, then the training and final recognition is done by BPANN. At last the comparison is done between this two phases on the basis of Recognition Rate and amount of error. Same methodology is applied for Language Identification also. The flow chart description of phase 1 and phase 2 is given below.

2.1 Feature Extraction

For the feature extraction process MFCC and LPC is used. LPC and MFCC is most widely used feature extraction technique. Feature Extraction gives the compact view of the input speech signal and it also converts the audio signal into numerical values. The features values are applied to the Back Propagation Artificial Neural Network (BPANN) for the training process. When the training completed then the final testing and recognition is done of the test signal by BPANN.

2.1.1. MFCC Feature Extraction

Mel-frequency cepstral coefficients (MFCC) are coefficients that collectively make up an MFC. They are derived from a kind of cepstral illustration of the audio clip. The distinction between the cepstrum and also the mel-frequency cepstrum is that within the MFC, the frequency bands are unit equally spaced on the mel scale, which approximates the human additive system's response more closely than the linearly-spaced frequency bands utilized in the traditional cepstrum. MFCC features are obtained by applying following steps which are read the input signals, preprocessing and frame blocking of the input signals, windowing, taking the Fast Fourier Transform (FFT) of the given signal, mel frequency wrapping, taking the log of the signal and Inverse Discrete Fourier Transform (IDFT), after applying these steps we get the MFCC coefficients. After applying the inputs signal preprocessing is done, in preprocessing noise is removed and sampling is done at the frequency of 8 KHz. After preprocessing the signals are send for frame blocking, in this process input signals are divided into frames and frame overlapping is avoided. Further frame blocking process windowing is applied on the signal, for this Hamming Windowing is used, then the Fast Fourier Transform (FFT) is taken of the windowing signals. The FFT signals are wrapped into the mel scale then we get the mel spectrum. Further log is taken of the signals and Inverse Discrete Fourier Transform (IDFT) is applied, finally we get the MFCC coefficients. These coefficients are in the form of numerical value and it gives the precise view of the input audio signals. The flow chart of MFCC analysis is given in Fig3.

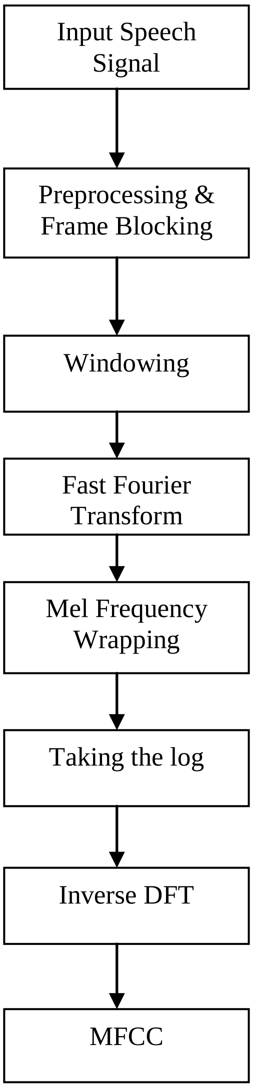

Fig.3 Flow Chart of MFCC Analysis

2.1.2 LPC Feature Extraction

LPC analysis is considers as a strong Feature Extraction process of the input signal analysis to compute the main parameters of speech signals. It is passive feature extraction technique and it encodes speech at low bit rate and also provides the accurate estimates of speech parameters of input Speech signal. The basic idea of LPC feature extraction process is to estimate an input speech sample as a linear Combination of past speech sample. LPC analysis consists of Pre-emphasis; frame blocking, Hamming Window, Auto Correlation analysis. On the basis of best auto correlation value the LPC coefficient are selected. LPC feature extraction is also a mostly used feature extraction technique for the speech recognition system, when LPC feature extraction technique applied on the input speech signals then the signal converted into the numerical value. The block diagram for the LPC analysis is given in Fig4.

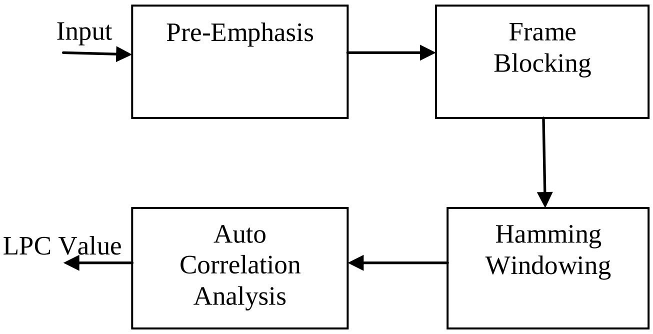

Fig4. Block Diagram of LPC Analysis

2.2 Training, Testing and Final Recognition by BPANN

For the training and testing of input speech signals Back Propagation Artificial Neural Network (BPANN) is used. For training there are total 1020 samples for speech recognition there are total 800 samples for Language Identification, the total numbers of testing samples are 340 for both process. It is a multi layer feed forward type neural network, it used delta learning for training of the neurons. The main property of this network is back propagation of the errors. The training process of this network includes four steps which are initializing the weight of the neurons, feed forward the signal, and propagate the error in the back direction and updating the weighted and bias connection between the neurons. The architecture of the BPANN is shown in Fig5, where X1 and X2 are input layer neurons, Z1 and Z2 are hidden layer neurons, Y1 and Y2 are output layer neuron and bias is 1. Bias is connected to the hidden and output layer neurons and straight line is showing weighted connection between the layers. The numbers of neurons can be increase in each layer of the BPANN, in hidden layer the number of neurons can be changed for weight modification.

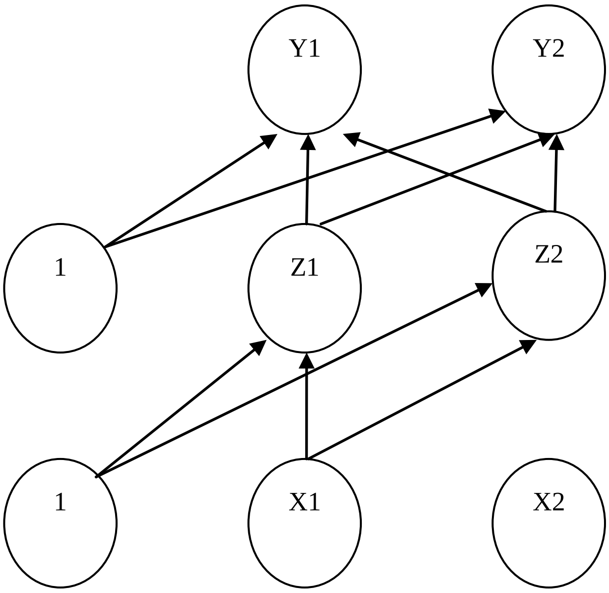

Fig5. BPANN Architecture

In the flow chart of the methodology we can see that phase1 and phase2 is similar except that in phase2 Particle swarm Optimization (PSO) feature selection process is also used. After the PSO feature selection the training and testing is done in phase2.

2.3 Particle Swarm Optimization (PSO) techniques

Particle Swarm Optimization (PSO) is a stochastic optimization technique which is developed in 1995, this technique is inspired by bird flocking. This technique is developed by Dr. Ebehart and Dr. Kennedy. PSO is very much similar to the Genetic Algorithm (GA), PSO has some advantages over GA which are it is easy to implement and in PSO there are very few parameters to adjust. PSO has been applied in function optimization, ANN training, fuzzy system control etc. The flow chart for PSO technique is given in Fig6 which consists of many steps:

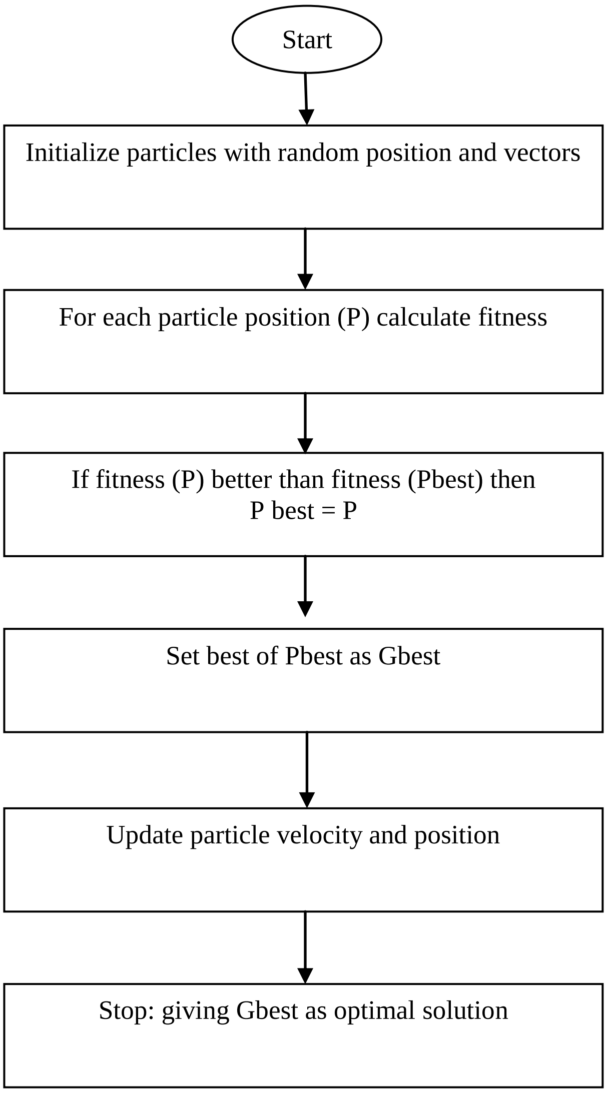
Start
Initialize particles with random position and vectors
For each particle position (P) calculate fitness
If fitness (P) better than fitness (Pbest) then
P best = P
Set best of Pbest as Gbest
Update particle velocity and position
Stop: giving Gbest as optimal solution

Fig6. Flow Chart of PSO technique

3. RESULT & DISCUSSION

3.1 Experimental Setup

This research work requires the Database of the Bengali, Chhattisgarhi, English and Hindi language. These database or speech signals have recorded with the microphone at the sampling frequency of 44.1 KHz. Database has collected of 20 persons for these four languages. The sentence which is recorded by the each person's is "Ekhone Tumi Jao", "Ae Bar Teha Ja", "Now This Time You Go" and "Ab Is Bar Tum Jao", these sentences are of Bengali, Chhattisgarhi, English and Hindi languages respectively. After that each word of these sentences has separated with the help of "AUDACITY" tool, now the each word is ".wav" file and it can be easily loaded to the MATLAB for further processing. In the first phase when these signals are loaded in MATLAB preprocessing, Feature Extraction, ANN training and ANN testing have performed. In the second phase when these signals are loaded in MATLAB preprocessing, Feature Extraction, Particle Swarm Optimization (PSO) of the features value, ANN training and ANN testing have performed. For feature extraction the combination of MFCC and LPC is used. The performance of this work is measured on the basis of the recognition rate and the percentage of error. Language Identification is also performed for Bengali, Chhattisgarhi, English and Hindi speech signal. The database is same for Language Identification and speech recognition. For language Identification all the processes are same as speech recognition.

3.2 Multilingual Speech Recognition Result

The performance of various methods can be evaluated by considering the "Recognition Rate" and the "Percentage of error" of different Speech signals. For the training and testing of input speech signals Back Propagation Artificial Neural Network (BPANN) is used. For training there are total 1020 samples for speech recognition there are total 800 samples for Language Identification, the total numbers of testing samples are 340 for both process.

1. Recognition Rate (RR): Recognition Rate is the ratio of total numbers of recognized signals to the total numbers of applied signals for speech recognition. It can be given by the following expression-

$$RR- \frac{\text{Number of recognized signals}}{\text{Total Number of signals}} *(100)$$

2. Percentage of error (PE): If the actual output is different from the desired output then the error occurs. Percentage of error can be defined as how much it different from the desired or target output. For a good speech recognition system the recognition rate should be high and the percentage of error should be very less.

On the basis of "Recognition Rate" and "Percentage of error" the result of the both phases is given in Table 1.

Table 1: Comparison of both phases of Speech Recognition

Methods	Recognition Rate	Percentage of error
1.Multilingual Speech Recognition without PSO (Phase1)	88%	10 to 12%
2. Multilingual Speech Recognition with PSO (Phase2)	90%	8 to 10 %

The Performance graph of Epochs value is shown in Fig7:

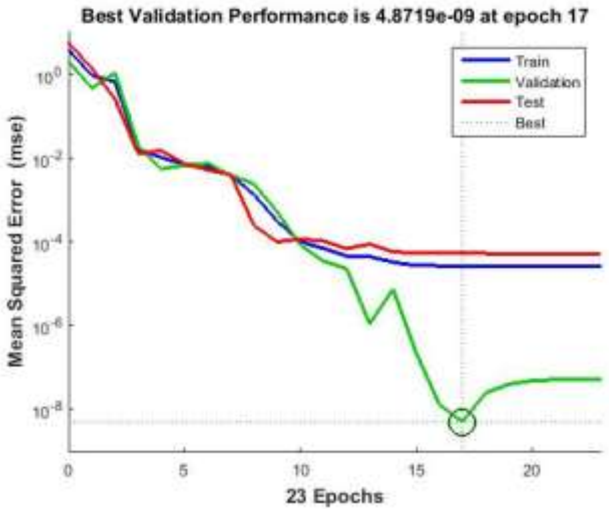

Fig.7Performance Graph of Epochs value

3.3 Language Identification (LID) Result

On the basis of "Recognition Rate" and "Percentage of error" the result of the Language Identification of both phases is given in Table 2. The recognition rate is reaches up to sufficient level.

Table 2: Comparison of both phases of Language Identification

Methods	Recognition Rate	Percentage of error
Language Identification(LID) without PSO (phase1)	85%	10 to 15 %
Language Identification (LID)with PSO (Phase2)	88%	10 to 12 %

The Performance graph of Epochs value is shown in Fig8.

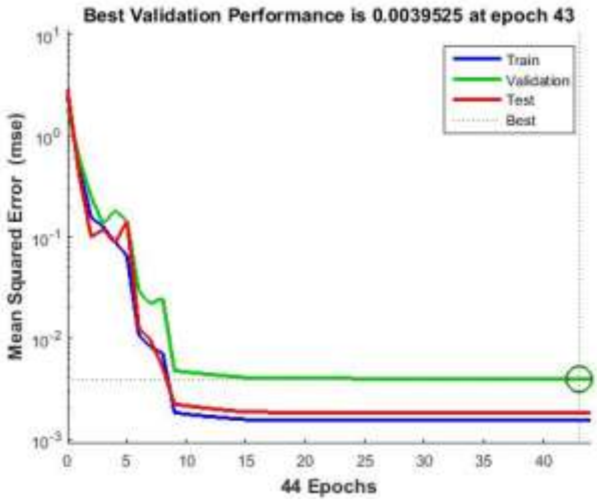

Fig.8Performance Graph of Epochs value

4. CONCLUSION & FUTURE SCOPE

As we discussed above in this research work Multilingual Speech Recognition and Language Identification has done with and without PSO technique. Each of this work is done in two phases. First phase is speech recognition of Bengali, Chhattisgarhi, English and Hindi speech signal with Artificial Neural Network and the second phase is speech recognition of Bengali, Chhattisgarhi, English and Hindi speech signal with the combination of Particle Swarm Optimization (PSO) feature selection and Artificial Neural Network. Speech recognition gives recognition rate of 88% without PSO technique and Speech recognition gives recognition rate of 90% with PSO technique. The percentage of error varies from 8 to 12%.

Language Identification is also performed for Bengali, Chhattisgarhi, English and Hindi speech signal. LID gives recognition rate of 85% without PSO technique and LID gives recognition rate of 88% with PSO technique. The percentage of error varies from 10 to 15%.

So we can conclude that Multilingual Speech Recognition with PSO gives good result as compare to without PSO, similarly in Language Identification with PSO gives good result as compare to without PSO. The performance is also good for these two methods which is shown in Fig.7 and Fig.8.

This work used Back Propagation ANN for training and testing process, for feature extraction MFCC and LPC is used. The current work is done for only four languages, further it can be done for some more languages. In future some other ANN like RBF, LVQ, and Self Organizing Map etc can be used for increasing the recognition rate of the system and to reduce the error of the system. In feature extraction some variant of MFCC and LPC can be also used.

REFERENCES

[1] Neelima Rajput and S.K. Verma, “Back Propagation Feed forward neural network approach for Speech Recognition”, Department of C.S.E, GBPEC, Pauri Gharwal, Uttrakhand, India, IEEE 2014

[2] Behi Tarek, Arous Najet, Ellouze Noureddine , “Hierarchical Speech Recognition system using MFCC feature extraction and dynamic speaking RSOM”, Laboratory of Signal, Image and Information Technologies National Engineering school of tunis, Enit Université Tunis El Manar, Tunisia, IEEE 2014

[3] Burcu Can, Harun Artuner, “A Syllable-Based Turkish Speech Recognition System by Using Time Delay Neural Networks (TDNNs)”, Burcu Can, Harun Artuner Department of Computer Engineering Hacettepe University Ankara, Turkey, IEEE 2013

[4] Mohamed ETT AOUIL Mohamed LAZAAR Zakariae EN-NAIMANI, “A hybrid ANN/HMM models for arabic speech recognition using optimal codebook”, Modelling and Scientific Computing Laboratory, Faculty of Science and Technology, University Sidi Mohammed ben Abdella Fez, MOROCCO, IEEE2013

[5] Ossama Abdel-Hamid Abdel-rahman Mohamed Hui Jiang Gerald Penn, “Applying Convolutional Neural Networks Concepts To Hybrid NN-HMM Model For Speech Recognition”, Department of Computer Science and Engineering, York University, Toronto, Canada, IEEE 2012

[6] Anup Kumar Paul ,Dipankar Das, Md. Mustafa Kamal, “Bangla Speech Recognition System using LPC and ANN”, Dhaka City College, Dhaka, Bangladesh, IEEE 2009

[7] Md Sah Bin Hj Salam, Dzulkifli Mohamad, Sheikh Hussain Shaikh Salleh, “Temporal Speech Normalization Methods Comparison in Speech Recognition Using Neural Network.”, Comp. Science and Info. System University Technology Malaysia 81300 Skudai, Johor, Malaysia, IEEE 2009

[8] Purva Kulkarni, Saili Kulkarni, Sucheta Mulange, Aneri Dand, Alice N Cheeran, “Speech Recognition using Wavelet Packets, Neural Networks and Support Vector Machines.”, Department of Electrical Engineering Veermata Jijabai Technological Institute Mumbai, India, IEEE 2014

[9] Javier Gonzalez-Dominguez, David Eustis, Ignacio Lopez-Moreno, Francoise Beaufays, and Pedro J. Moreno ,“A Real-Time End-to-End Multilingual Speech Recognition Architecture.”, IEEE 2014

[10] Niladri Sekhar Dey, Ramakanta Mohanty, K. L. chugh et al [10] proposed “Speech and Speaker Recognition System using Artificial Neural Networks and Hidden Markov Model.”, IEEE 2014

[11] Pialy Barua, Kanij Ahmad, Ainul Anam Shahjamal Khan, Muhammad Sanaullah “Neural Network Based Recognition of Speech Using MFCC Features.”, 2Department of Electrical and Electronic Engineering, Chittagong University of Engineering and Technology, Chittagong-4349, Bangladesh,IEEE 2014

[12] Oscal T.-C. Chen, Chih-Yung Chen ,“A Multi-lingual Speech Recognition System Using a Neural Network Approach.”, Computer & Communication Research Laboratories, Industrial Technology Research Institute,Hsinchu, Taiwan, R.O.C., IEEE 1996

[13] G. Rigoll, c. Neukirchen “A New Approach to Hybrid HMM/ANN Speech Recognition Using Mutual Information Neural Networks.”, Gerhard-Mercator-University Duisburg Faculty of Electrical Engineering, Department of Computer Science Bismarckstr. 90, Duisburg, Germany, IEEE 1995