ISOLATED WORD RECOGNITION SYSTEM FOR TAMIL SPOKEN LANGUAGE USING BACK PROPAGATION NEURAL NETWORK BASED ON LPCC FEATURES

Dr.V.Radha[1], Vimala.C[2], M.Krishnaveni[3]

Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

[1]radharesearch@yahoo.com
[2]vimalac.au@gmail.com
[3]krishnaveni.rd@gmail.com

ABSTRACT

Speech recognition has been an active research topic for more than 50 years. Interacting with the computer through speech is one of the active scientific research fields particularly for the disable community who face variety of difficulties to use the computer. Such research in Automatic Speech Recognition (ASR) is investigated for different languages because each language has its specific features. Especially the need for ASR system in Tamil language has been increased widely in the last few years. In this paper, a speech recognition system for individually spoken word in Tamil language using multilayer feed forward network is presented. To implement the above system, initially the input signal is preprocessed using four types of filters namely preemphasis, median, average and Butterworth bandstop filter in order to remove the background noise and to enhance the signal. The performance of these filters are measured based on MSE and PSNR values. The best filtered signal is taken as the input for the further process of ASR system. The speech features being the major part of speech recognition system, are analyzed and extracted via Linear Predictive Cepstral Coefficients (LPCC). These feature vectors are given as the input to the Feed-Forward Neural Network for classifying and recognizing Tamil spoken word. Experiments are done with sample Tamil speech signals and its performance are measured based on Mean Square Error (MSE) rate. The adopted network with the above specified parameters has produced the best result for limited vocabulary.

KEYWORDS

Bandstop Filter, Linear Predictive Cepstral Coefficients, Feed-forward neural networks, Tamil Speech Recognition, Mean Square Error, and Peak signal to Noise Ratio

1. INTRODUCTION

In recent years, with the new generation of computing technology, speech technology becomes the next major innovation in man-machine interaction. Obviously such interface would yield great benefits which can be accomplished with the help of Automatic Speech Recognition (ASR) system. It is a process by which a machine identifies speech. It takes a human utterance as an input and returns a string of words as output. Such research on ASR systems is primarily developed for English language but for Indian languages it is still in earlier stage. Tamil is one of the widely spoken languages of the world with more than 77 million speakers. Hence, there is an

DOI : 10.5121/cseij.2011.1401

urgent need for the system to interact with Tamil language. Recently, Neural Network (NN) [1] is considered as one of the most successful information processing tool that has been widely used in speech recognition [2][3]. The Multi-Layer Perceptron (MLP) have been increasingly used for word recognition and also for other speech processing applications. The main objective of this paper is to implement classification and recognition system for isolated Tamil [6][11]spoken words. To carry out this task two important preprocessing [4][5] steps are done before feature extraction[2][11] which includes filtering, [4] and windowing. Among the four filters implemented, the best filtered speech signal is chosen and fed as an input. These filtered outcomes are evaluated based on MSE and PSNR values. The popular feature extraction [2][10] technique of LPCC is used for extracting specific features from the speech signal. Using LPCC, 24 feature vectors are extracted and they are given as the input to the network. Finally, the speech recognition [2][3] is implemented using feed-forward neural networks [1]and its performances are measured based on its default parameter MSE.

The paper is organized as follows. Section 2 gives details about the system overview. Section 3 explains the preprocessing [5] steps involved in this system. Section 4 gives details about feature extraction technique based on LPCC. Section 5 deals with Feed forward neural network techniques and its performance in Tamil [6][11] speech signal. Section 6 explores the performance evaluation of the adopted method. Finally, the conclusion is summarized in section 7 with future work.

2. SYSTEM OVERVIEW

There are variety of speech recognition [2][3] approaches available such as Neural Networks, Hidden Markov Models, Bayesian networks and Dynamic Time Warping etc. Among these approaches Neural Networks (NNs) [1] have proven to be a powerful tool for solving problems of prediction, classification and pattern recognition. Rather than being used in general-purpose speech recognition applications it can handle low quality, noisy data and speaker independence applications. Such systems can achieve greater accuracy than HMM based systems, as long as there is training data and the vocabulary is limited.

One of the most commonly used networks based on supervised learning algorithm is multilayer feed forward network which is implemented in this paper for classifying and recognizing Tamil spoken words [6][11]. In Tamil language, the pronunciation of independent letters and group of letters forming words are not different. Tamil speech recognizing system [11][11] does not require the support of a dictionary. Thus the recognizing process in Tamil speech [11] is fairly simple compared to English.

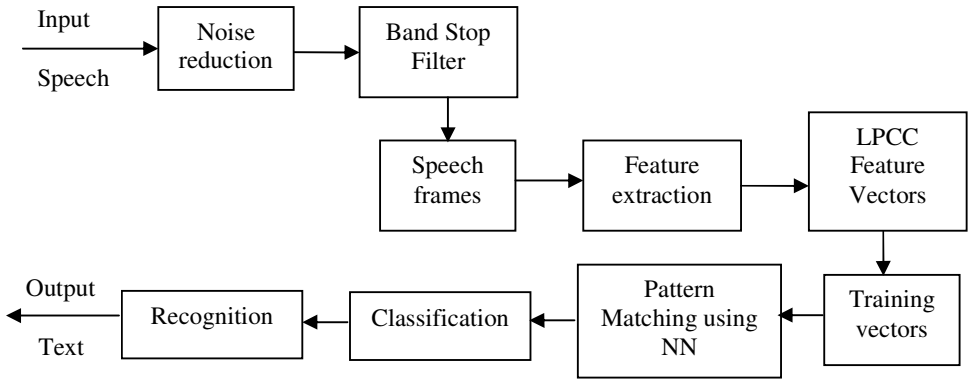

Figure 1. System overview of speech recognition based on NN

To implement the system, initially the speech data is preprocessed using filtering, framing [4] and windowing techniques. Subsequent to that, the enhanced signal is given as the input to the LPCC algorithm to extract features. These feature vectors also called cepstral coeffients are given as the input to the network. After that the network is trained with these input vectors and the target vectors. Finally classification and recognition is done based on pattern matching. The above figure 1 demonstrates the overall structure of the system.

3. PREPROCESSING

To enhance the accuracy and efficiency of the speech signals, they are generally pre-processed before further analysis. The selected dataset usually needs to be pre-processed prior to being fed into a Neural Network. In this research work, filtering, [4] and windowing are used as the important preprocessing steps.

3.1. Filtering Techniques

One of the foremost preprocessing steps in speech processing is to apply filters to the signal. It is essentially used for speech enhancement by reducing the background noise and to improve the quality aspects of speech. To reduce noise in the signal four types of filters are used namely pre emphasis, median, average and Bandstop. All these filters are implemented and based on the experimental results, it was found that the Bandstop filter performs well among the other filters.

A Band-Stop filter [7] works to screen out frequencies that are within a certain range, giving easy passage only to frequencies outside of that range. It is also called as band-elimination, band-reject, or notch filters. Placing a low-pass filter in parallel with a high-pass filter can make it as band-stop filter. The range of frequencies that a band-stop filter [7] blocks is known as the 'stop band', which is bound by a lower cut-off frequency and a higher cut-off frequency. The frequency of maximum attenuation in it is called the notch frequency [7]. Hence, apart from these filtes, specifically band stop filter works better for Tamil speech signal.

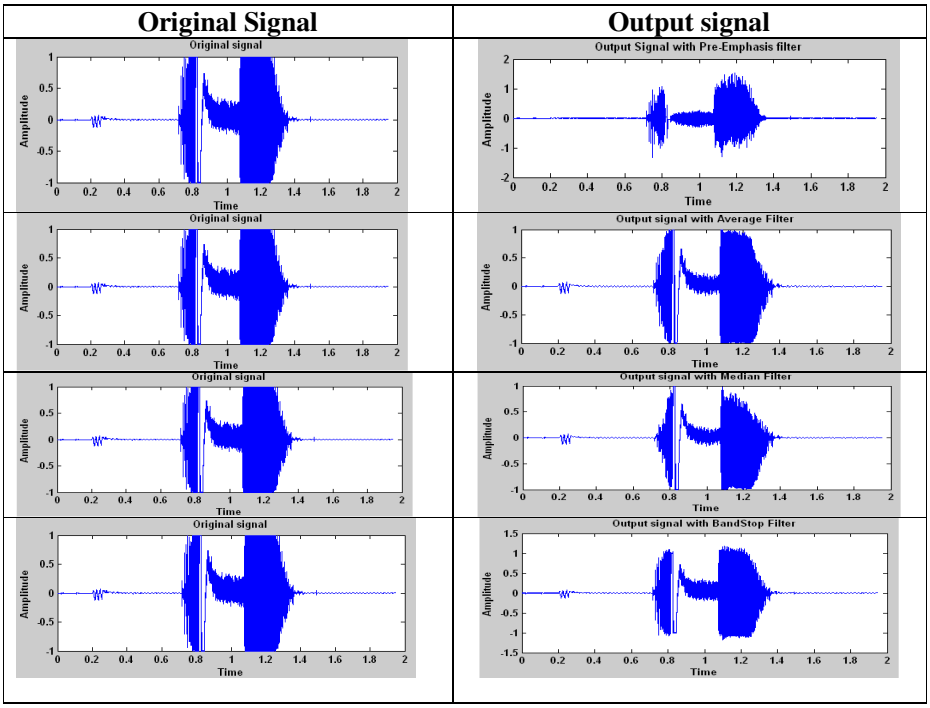

Figure 2. Original signals and the resulted signals after applying filters

Both subjective and objective performance evaluations are done for the above taken filters. This resulted signal is used for further processing of a system. The figure 2 shows the subjective evaluation of the above four filters.

3.2. Framing and windowing

Framing [4] and windowing are the important preprocessing [4] steps in speech signal processing. In this step, the speech signal is divided into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are 256.

The next step in the preprocessing [5] is to applying window for each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. It is used for minimizing the spectral distortion by using the window to taper the signal to zero at the beginning and the end of each frame. It is defined as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame and then the result of windowing is the desired signal. Windowing is done by using the equation(1).

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1 \qquad \text{-------} \quad (1)$$

Window choice is an important task in any signal processing applications. Among the different types of windows like triangular, Blackman etc, the hamming window best suited for speech signal processing. Typically the Hamming window is defined as in the equation (2).

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

4. FEATURE EXTRACTION USING LPCC

The motivation of feature extraction [2] [11] is to convert speech waveform to a parametric representation at a lower information rate for further analysis. Speech signals have non-stationary characteristics. As a result, the speech waveforms are commonly small frames (typically 5 ms to 40 ms) in which the signal characteristics are considered quasi-stationary to allow for short-term spectral analysis and feature extraction. A wide range of feature extraction techniques are available such as Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC). In this research work the most frequently used LPCC parameters are considered to determine the best feature set for the Tamil[6][11] speech database.

LPC has been widely used in speech recognition systems. Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. The basic idea behind linear predictive analysis is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. These coefficients form the basis for linear predictive analysis of speech. The analysis provides the capability for computing the linear prediction model of speech over time.

Feature vector sequence for each frame is usually obtained from the LP parameters. To extract the LPC features, the speech signal is blocked into frames of N samples and each frame is multiplied

by hamming window. Short term autocorrelation analysis is performed to find the important parameter of frame energy for speech-detection. After that coefficients based on Levinson-Durbin recursion are extracted and then converted to Q cepstral coefficients, which are weighted by a raised sine window. Finally the observation vectors are derived. In this work, the feature vector is composed of 12 cepstral coefficients and 12 difference cepstral coefficients. These feature vectors are fed as an input to the further processing of classification and feature matching technique. The steps involved in LPCC can be seen in the following figure 3 which illustrates the functioning process of deriving the Linear Predictive Cepstral Coefficients.

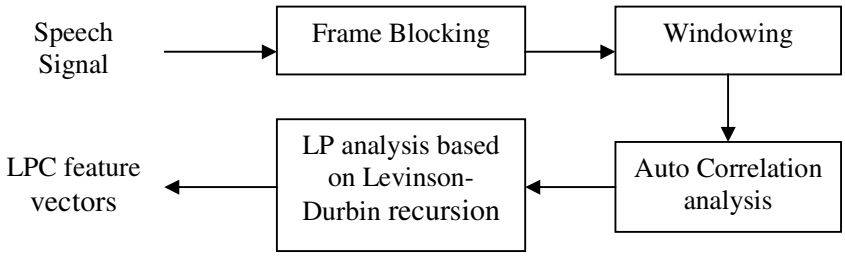

Figure 3. Steps involved in LPCC Feature Extraction

The following figure 4 and table 1 shows the 24 feature vectors derived from LPCC for the sample Tamil [11][11]speech signals. These coeffients are taken as the input vector for the network.

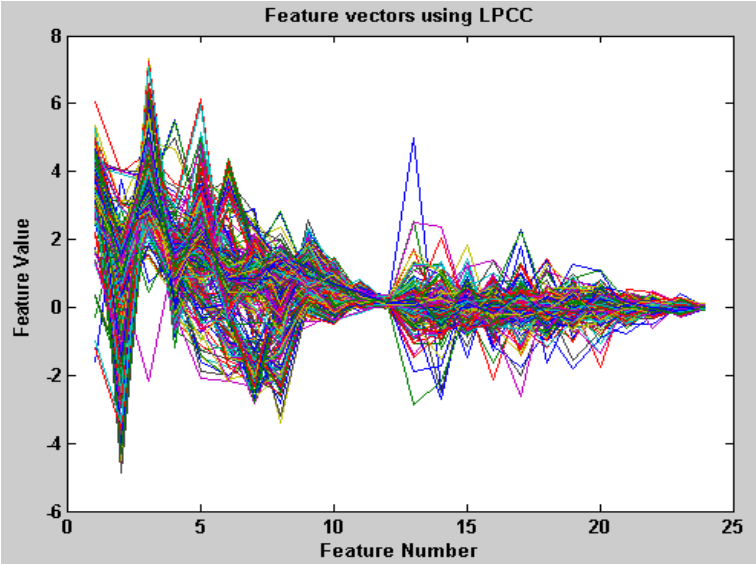

Figure 4. LPCC Feature Vectors

The extracted 24 feature vectors are as follows.

Table 1. Sample Feature vectors of LPCC

No of Samples	24coefficients values of LPCC		
1	-1.599283742528188	3.766782961316008	0.745680586677735
	1.451395497027584	-0.263688342171344	2.247363287802652
	0.660030181678723	0.662417737480927	1.006378867319193
	0.260969375699623	-0.112279563573272	0.094232552444403
	5.005168336049138	-2.512648104615616	0.694054785660278
	-1.154910652385622	1.801365312711012	-1.64807471244783
	0.039985085383287	-0.548979143128312	-.790184036156872
	-0.230596601198459	0.396263326453119	0.008909620441753
2	3.264622679760679	1.602718733848364	2.237545972763804
	0.853247394542453	1.795880729779993	0.751691154682471
	1.2034488427438	0.422973740340685	0.310603526502768
	0.146402181260902	0.123949028340086	0.052376800140136
	-0.098649936819902	-0.010343505498683	-0.065149122298549
	-0.360198492676313	0.035612700053262	0.006617754381784
	0.161380616441981	-0.149707984814701	-0.047296736347172
	0.165822627902026	0.108051536217042	0.026345007072651
3	3.067398261744015	1.619255636284996	1.98051504515753
	0.950523115463605	1.715450558191198	0.580528011576593
	1.140387208789291	0.467484164846532	0.588995581038866
	0.124612830969022	0.162811357054451	0.080506723269004
	2.337349397960358	-0.149765932273949	1.383038001631312
	-0.125795386842127	0.896894504403429	0.076347055613712
	0.713768417563733	0.107955836955762	0.316313965670924
	-0.090411826387144	-0.031085737420752	0.019094593582235
4	2.699797836725626	0.879136987008197	2.563162021198413
	1.652253821895698	1.115606022141024	0.166967031141082
	1.354941295738535	0.499889348121276	0.810920044405672
	-0.37721217969717	0.413258451662497	0.098010774926258
	1.881791794501285	-0.092643822588443	1.138279122025976
	0.091275447713394	0.566700657981445	0.252959091991107
	0.144949809580169	0.711506360406204	0.453125769121287
	-0.344486759474099	.073971015547871	-.007003362470624
5	2.657067289569832	1.146244121227948	1.695298057787456
	0.801395661500785	.312027998744104	1.972936487034075
	0.913581231108055	1.172251565759136	0.983569618398127
	-0.057398604229567	0.415496109267708	0.079404594304159
	0.39996309462644	0.286476812804803	0.535960436446698
	0.066403108978424	-0.488201393600566	-1.073225064711092
	-0.01487651651683	-0.16204145768663	0.104837012207879
	0.029225488447888	-0.158220920261808	-0.006413676988691

5. FEED FORWARD NEURAL NETWORK FOR TAMIL WORD RECOGNITION

A feed forward neural network [11] is a biologically inspired classification algorithm which falls under the category, "Networks for Classification and Prediction" and has widespread interesting applications and functions related to speech processing. It consists of a (possibly large) number of simple neuron-like processing units, organized in layers. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal and may have a different strength or weight. The weights on these connections encode the knowledge of a network. They usually consist of three to four layers in which the neurons are logically arranged. The first and last layers are the input and output layers respectively and there are usually one or more hidden layers in between the other layers. The term feed-forward means that the information is only allowed to "travel" in one direction. This means that the output of one layer becomes the input of the next layer, and so forward. Feed-forward networks are advantageous as they have the fastest models to execute.

In the Network, the input layer does have a part in calculation, rather than just receiving the inputs. The raw data is computed, and activation functions are applied in all layers. This process occurs until the data reaches the output neurons, where it is classified into a certain category. The operation of this network can be divided into two phases:

1. The learning phase
2. The classification phase

During the learning phase the weights in the FFNet will be modified. All weights are modified in such a way that when a pattern is presented, the output unit with the correct category, hopefully, will have the largest output value. In the classification phase the weights of the network are fixed. A pattern, presented as the input will be transformed from layer to layer until it reaches the output layer. Now classification can occur by selecting the category associated with the output unit that has the largest output value. In contrast to the learning phase classification is very fast. The experimental results are presented in the following section.

6. PERFORMANCE EVALUATION

Performance evaluation is done for both filtering techniques and the Neural Network separately. Mean Square Error Rate (MSE) and Peak Signal to Noise Ratio (PSNR) values are used as the performance measures to find out the best filtering signal for feature extraction. The MSE value is calculated by the equation (3)

$$MSE = \frac{\Sigma\left(y_i - \hat{y}_i \right)}{n - p} \quad \text{---------- (3)}$$

and PSNR value is calculated by the equation (4)

$$PSNR = 10\log_{10}\left[\frac{R^2}{MSE} \right] \quad \text{---------- (4)}$$

The average values of 10 speech samples were taken for the performance evaluation. The figure 5 (a) and (b) presents the performance evaluation of four filters based on MSE and PSNR values. Among these filters Bandstop filter offer least MSE value and high peak signal value.

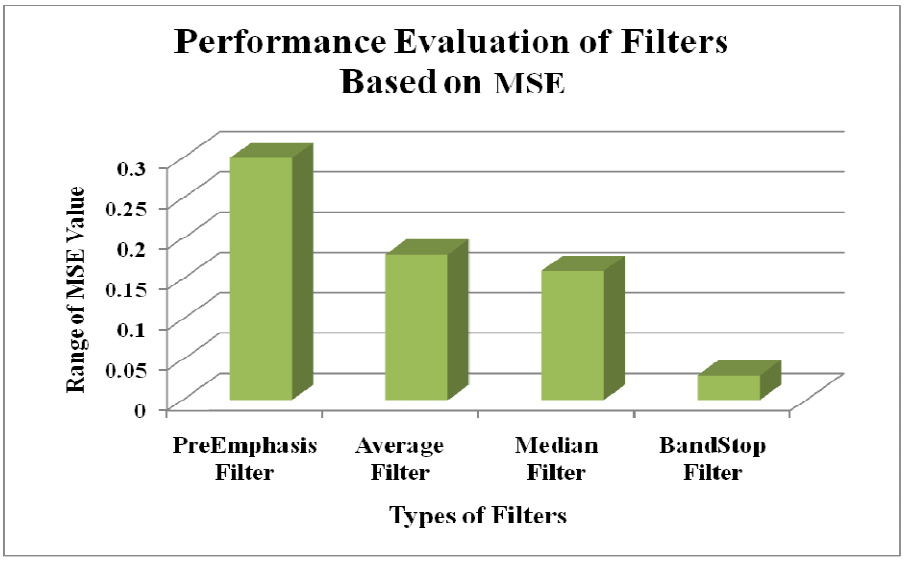

Figure 5 (a). Performance evaluation of filters based on MSE

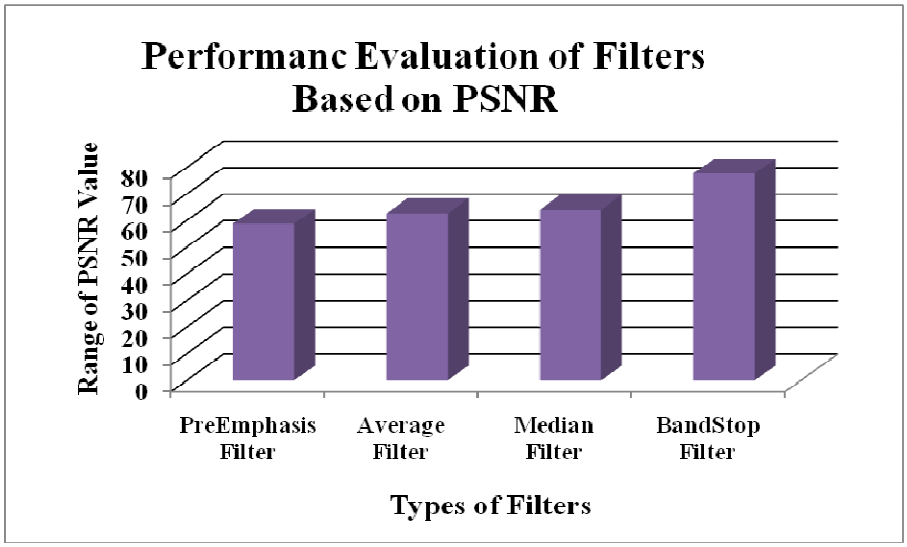

Figure 5 (b). Performance evaluation of filters based on PSNR

In this research work, the input speech signal is preprocessed with Bandstop filter. This filtered signal is divided into short frames and hamming window is applied to reduce the discontinuities between these frames. For further process of creating and training a network, the values given for the input vector are 24 coefficients of LPCC. There are 24 neurons in the first layer and 2 neurons in the second (output) layer. The transfer function used in the first layer is tan-sigmoid, and the output layer transfer function is linear. With standard back-propagation, the learning rate is held constant throughout training. The performance of the algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm may oscillate and become unstable. If the learning rate is too small, the algorithm will take too long to converge. However this sensitivity can be improved if it allows the learning rate to change during the training process. For this research work the learning rate of 0.03 is used and epoch count was

limited to 300. Different epoch was employed in the network in order to find the performance. The experiments reveal that as the number of epoch increases, the error rate was minimized and the performance was improved. The following figure 6 (a) and (b) shows the experimental results of actual and predicted value of a sample Tamil speech[11][11] signal. The performance goal is achieved with the parameters specified for the network.

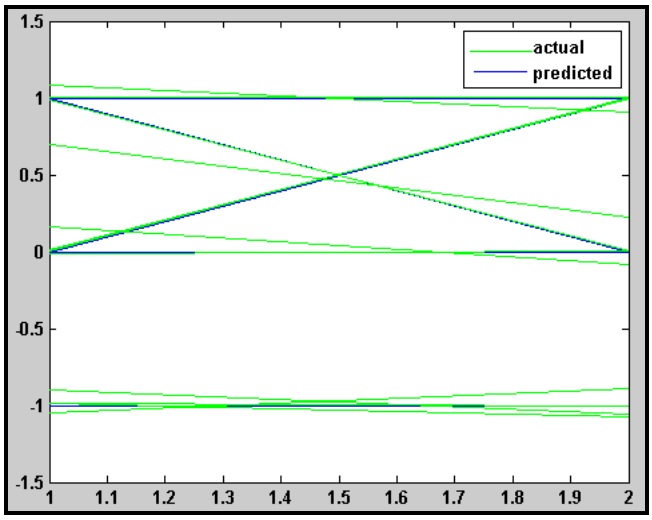

Figure 6(a). actual and predicted value

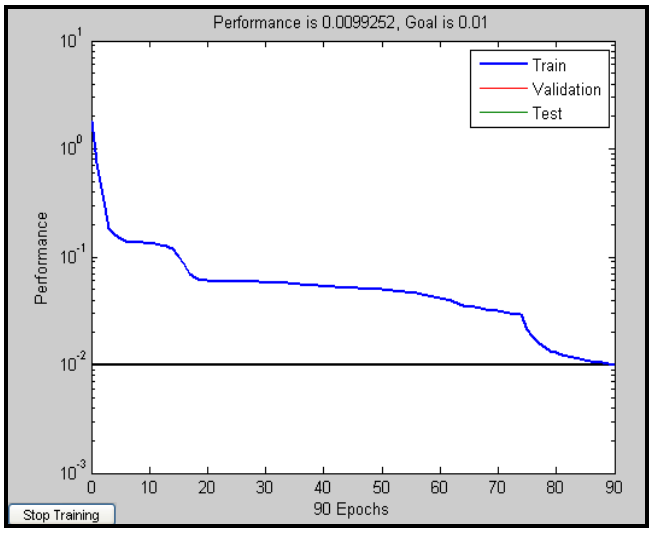

Figure 6(b). performance of the Network

The default performance evaluation measure for the adopted network is mean square error rate. It is computed by "summing the squared differences between the predicted values vs. actual value, and then divide the sum by the number of components. A threshold of 0.01 mean square error is good. Our research work achieved minimum MSE for the Tamil speech [11] signal i.e 0.00515/0.01.

7. CONCLUSION

In recent years, neural network has become an enhanced technique for tackling complex problems and tedious tasks such as speech recognition. Speech is a natural and simple communication method for human beings. However, it is an extremely complex and difficult job to make a computer respond to spoken commands. Recently there is a momentous need for ASR system to be developed in Tamil and other Indian languages. In this paper such an important effort is carried out for recognizing Tamil spoken words. To accomplish this task, feature extraction is done after employing required preprocessing techniques. The most widely used LPCC method is used to extract the significant feature vectors from the enhanced speech signal and they are given as the input to the feed forward neural network. The adopted network is trained with these input and target vectors. The results with the specified parameters were found to be satisfactory considering less number of training data. More number of isolated and continuous words to be trained and tested with this network in future. This preliminary experiment will helps to develop ASR system for Tamil language using different approaches like Hidden Markov Models or with other hybrid techniques.

ACKNOWLEDGEMENTS

The authors thank UGC Major Research Project, New Delhi, India, for funding the project on "Automatic Tamil Voice Recognition Browser using Continuous Hidden Markov Model for Visually impaired People".

REFERENCES

[1] Abdul Manan Ahmad', Saliza Ismail+, Den Fairol SamaonL(2004) "Recurrent Neural Network with Backpropagation through Time for Speech Recognition', *Intemational Symposium on Communications and Information Technologies 2004* (ISCIT2004), Sapporo, Japan, October 26- 29.

[2] Amin Ashouri Saheli, Gholam Ali Abdali, Amir Abolfazl suratgar,(2009) "Speech Recognition from PSD using Neural Network", *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009,* Vol I,IMECS 2009, March 18 - 20, 2009, Hong Kong.

[3] S K Hasnain, Azam Beg(2008), "A Speech Recognition System for Urdu Language", in *International Multi-Topic Conference (IMTIC'08)*, Jamshoro, Pakistan, 2008, pp. 74-78.

[4] S K Hasnain, (2008) "Recognizing Spoken Urdu Numbers Using Fourier Descriptor and Neural Networks with Matlab", *Second International Conference on Electrical Engineering,*25-26 March 2008,University of Engineering and Technology, Lahore (Pakistan) (2008)

[5] Patricia Melin, Jerica Urias, Daniel Solano, Miguel Soto, Miguel Lopez, and Oscar Castillo, (2006) "Voice Recognition with Neural Networks", *Type-2 Fuzzy Logic and Genetic Algorithms, Engineering Letters*, 13:2, EL_13_2_9 (Advance online publication: 4 August).

[6] Muthanantha murugavel, (2007) "Speech Recognition Model for Tamil Stops, Proceedings of the World Congress on Engineering 2007 Vol I, WCE 2007, July 2 - 4, 2007, London, U.K.

[7] Dr.V.Radha,Ms.Vimala.C and Ms.M.Krishanveni, (2010) "Reconstruction Methodology for Tamil Speech Recognition System", *Proceedings of International conference on computing (ICC2010)*, Advanced Computing Research Society, Institute for Defense Studies and Analysis in New Delhi, pp 152-157,ISBN:978-81-920305-1-7.

[8] S. Saraswathi1, T.V. Geetha, (2010) "Design of language models at various phases of Tamil speech recognition system", *International Journal of Engineering, Science and Technology,* Vol. 2, No. 5, 2010, pp. 244-257.

[9] S. Saraswathi and T.V. Geetha, (2004) "Implementation of Tamil Speech Recognition System Using Neural Networks", *Lecture Notes in Computer Science*, 2004, Volume 3285/2004, 169-176, DOI: 10.1007/978-3-540-30176-9_22.

[10] R. Thangarajan, (2008) "Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language", *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, Issue 3, Volume 4, March 2008.

[11] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, (2010) "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", *International Journal of Computer Theory and Engineering*, Vol.2, No.6, December, 2010,1793-8201.

Authors

Dr.V.Radha, is the Associate Professor of Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, TamilNadu, India. She has more than 21 years of teaching experience and 7 years of Research Experience. Her Area of Specialization includes Image Processing, Optimization Techniques, Voice Recognition and Synthesis, Speech and signal processing and RDBMS. She has more than 45 Publications at national and International level journals and conferences. She has one Major Research Project funded by UGC. Email id: radhasrimail@gmail.com

Ms.Vimala.C, currently doing Ph.D and working as a Project Fellow for the UGC Major Research Project in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women. She has more than 2 years of teaching experience and 1 year of research experience. Her area of specialization includes Speech Recognition and Synthesis. She has 9 publications at National and International level conferences.
Email id: vimalac.au@gmail.com

Ms. M. Krishnaveni, 5 Years of Research Experience. Working as Research Assistant in Naval Research Board project, Area of Specialization: Image Processing, Pattern recognition, Neural Networks. She has 40 publications at national and International level journals and conferences. She has one Major Research Project funded by UGC and one under DST. Email id:krishnaveni.rd@gmail.com.