Modeling of Speech Recognition Using Artificial Neural Network

Nidhi Srivastava

Abstract*: Automatic speech recognition has attained a lot of significance as it can act as easy communication link between machines and humans. This mode of communication is easy for man to use as it is effortless and easy. Many approaches for extraction of the features of the speech and classification of speech have been considered. This paper unveils the importance of neutral network and the way it can be used for recognition of speech. Mel Frequency Cepstrum Coefficients is made use of for extraction of the features from the voice. For pattern matching neural network has been used. MATLAB has been used to show how the speech is recognized. In this paper the speech recognition has been done firstly by multilayer feed forward neural network using Back propagation algorithm. Then the process of speech recognition is shown by using Radial basis function neural network. The paper then analyzes the performance of both the algorithms and experimental result shows that BPNN outperforms the RBFNN.

***Index Terms*: BPNN, Feature Extraction, Neural Network, RBFNN, Speech Classification.**

I. INTRODUCTION

Humans always prefer a natural way of communication. Human to human communication is done using more than one mode of communication like speech, eye gaze, hand gestures, facial expressions and body gestures. But among all these communications, speech is most preferred. The humans prefer using the same effortless and natural mode of communication and so speech has gained a lot of importance as significant mode of interface with machines. Speech is very subjective in nature and analysis of speech is very difficult. Speech of different people differs with respect to different changeable factors like background noise, emotional state of a person, the age of a person, position of holding the mike, pronunciations of the words, accent of a person etc. [1].

II. PROCESS OF RECOGNITION OF SPEECH

The process of recognition of speech involves four steps. These are speech input, preprocessing of speech, extraction of features and classification of speech.

A. *Speech*

In this the voice of the speaker is accepted and stored in a waveform. In today's world the speech can be recorded using any of the software among the several available in the market.

B. *Speech Pre-processing*

This step is necessary to remove the noise from the input speech. It involves noise filtering, framing, smoothing, windowing, etc. [2, 3].

C. *Feature Extraction*

Speech has different individual characteristics which separate or identify one speech from another. Feature extraction is the method through which the features of the speech for identification can be extracted. Feature extraction is a very critical and elementary step in speech recognition [4]. There are different techniques which exist for feature extraction. Among these are LDA, LPC, PCA, HMM, etc. Each of these techniques is explained briefly:

LPC (Linear Predictive Coding) – It is the most commonly used method to extract the feature vectors. It computes the feature extraction using linear prediction coefficients like amplitude, pitch, filter coefficient, etc. These coefficients are then changed to cepstral coefficients which are then normalized in the range from -1 to 1. At a low bit rate the LPC can encode good quality speech and analyze it. In LPC all the earlier samples of the signal are combined linearly to give the next sample. LPC is very sensitive to quantization noise [4, 5, 6, 7].

LDA (Linear Discriminant Analysis) - Linear Discriminant Analysis helps in classification of the data and also reduces the dimensionality of the data. So as to maximize the class separability, high dimensional data is mapped onto a low dimensional space in this technique. The cases where within class frequencies are unequal, LDA are a good choice [8].

PCA (Principal Component Analysis) – PCA is also a well-known technique for extraction of the feature and also for reduction of dimensionality of data. It can easily extract structure from possibly high-dimensional data sets. PCA identifies patterns in data and emphasizes and depicts the similarities in data and dissimilarities in data. It is a dominant and prevailing tool for analyzing data. This technique is more useful for linear structure rather than non-linear structure [9].

D. *Speech Classification*

Once the features have been extracted these are matched with the new voice to ascertain the speech or recognize the speaker. There are many different methods for speech classification like HMM, Vector Quantization, DTW, etc. All these methods involve complex mathematical calculations. Some of these methods are described below:

HMM (Hidden Markov Model) – HMM uses statistical methods to analyze the variations in the features. It is used in signal processing, pattern recognition, speech recognition, etc. HMM can classify unknown feature vector sequences, as it has the capability of dealing with

Revised Manuscript Received on July 04, 2019.

Dr. Nidhi Srivastava, Asst. Professor, Amity Institute of Information Technology, Amity University, Uttar Pradesh, Lucknow Campus, Lucknow, India.

time sequential data, time scale invariability and learning capability. HMM treat observations as a sequence in time or space. HMM uses a stochastic approach [9, 10].

VQ (Vector Quantization) – This is the easiest algorithm of feature matching. It is an unsupervised learning algorithm. The clusters of the speakers are constructed in VQ and used for matching. Once the feature vectors are gathered, the clusters are created which represents the models of the speakers. This model is then compared to the new feature vector collected from the speaker [11, 12].

DTW (Dynamic Time Warping) – Speech of a person is dependent on time. The utterance of the similar word can differ with respect to duration of the word. Further, the same word spoken in equal duration can be different in the central part of the speech. The two different speech arrangements are aligned in time in DTW to get the global distance between them. DTW is dynamic programming techniques which follows the principle of divide and conquer. DTW basically measures the similarity in the two dynamic patterns by calculating a minimum distance between them [13, 14].

MFCC (Mel-Frequency Cepstral Coefficients) - This was given by Davis and Mermelstein. MFCC deals with frequencies. MFCC can be divided into different components like frame blocking, windowing, FFT, etc. The different components of the MFCC-

i. Frame blocking: Blocks of frames are made from the continuous speech by dividing them into N blocks. Each of these blocks are then separated by M, where M<N.

ii. Windowing: The start of the frame and the end of the frame has signal discontinuities. So, each of the frames are windowed. If N is the number of samples in each frame and window is defined as w(n), 0<= n <=N then the result of windowing is the following signal

$$y1(n) = x1(n) \,.\, w(n) \qquad 0 \le n \le N -1$$

In this the Hamming window is used. Its form is:

$$W(n) = 0.54 – 0.46\cos (2\pi n \;/ N-1)$$

iii. Fast Fourier Transform: Each of the above N samples is converted from time domain to frequency domain.

iv. Mel-frequency Wrapping: Mel scale is used so as to measure the pitch of every tone having a frequency f.

v. Cepstrum: This is the last step in which MFCC (mel frequency cepstrum coefficients) is calculated. For this we use DCT so that the log mel spectrum is reconverted into time. [15, 16]

III. LITERATURE REVIEW

Different researchers have used different methods for speech recognition. Some of the works done by such researchers are described in this section. [17] in their work have used MFCC and HMM respectively for feature extraction and classification of speech. The experiment has been done for speech recognition on English digits. The authors have achieved a relatively successful result in both clean and noisy environment. [18] in their paper designed a system where connected Hindi digits are recognized using HMM. This is a speaker-independent system. In this the authors have compared the performance of various feature extraction methods like Perceptual Features (PLP, MF-PLP), MFCC and ΔMFCC, Bark Frequency Cepstral Coefficient and Revised Perceptual Linear Prediction (BFCC, RPLP). Results show that MF-PLP outperforms the various other feature extraction methods. [19] in their paper have used LPC and ANN for recognition. The technique has been applied on vowels and a fair accurate result has been achieved by the authors. [9] in their work have performed various experiments with different feature extraction techniques. Using only single technique reduces the accuracy or increases the time taken in feature recognition. A hybrid approach for feature extraction like MFPCA, MFLDA, MFDWT, MFPDWT and MFLDWT performs better and gives good result.

IV. SPEECH RECOGNITION USING MFCC & NEURAL NETWORK

In this paper, for recognition of speech, Neural Network with MFCC has been used. MFCC has been explained in detail in the above section. It is used for extraction of the features of the speech. The concept of neuron is taken from the neurons present in the human body, which are biological cells. These cells using electrical or chemical change pass information from one neuron to another. Neural networks in computer are general-purpose programs that have the capacity to learn difficult input-output relationships which are non-linear in nature. These networks are trained and they adapt to the data fed to them [20]. Some of the neural networks used are the feed-forward network, Self-organizing Map (SOM), Radial Basis function network, etc. Multilayer Feedforward Network using Back Propagation algorithm is used to train the samples for classification of the speech. In our experiment we have used BPNN and Radial basis network and shown a performance evaluation of the two. In this we have the actual input which gives the desired output and this BPNN [21] algorithm tries to reduce the mean square error between the two. In Radial Basis Function Network (RBFN) the function are symmetric around their centers and so called radial. Basis functions means that the function can generate an arbitrary function from the linear combination of their functions.

V. IMPLEMENTATION & RESULT

MATLAB R2015a has been used in the conducted Experiment. Neural network toolbox has been used. 5 people, 3 male and 2 female, each emitting six words has been taken in the study. Each one of these people recorded each word twenty times. Thus, in total 600 voice samples were used. Mic was used to record the samples. 44100 Hz sampling frequency was used for recording the voices in MATLAB. Noise from each of these samples was removed. Finally, out of 600 voices, for training 300 samples were used and for testing 300 samples were used. So as to differentiate between notes, extraction of the features of the speech plays a significant part in speech recognition system. For extraction of the features MFCC technique has been used. To calculate MFCC, VOICEBOX speech processing toolbox [22] has been used in the experiment. MFCC coefficients are calculated using melcepst function from this toolbox. 12 coefficients of MFCC were calculated and passed to

NN as input. To create, train and simulate the networks, Neural Network toolbox of MATLABR2015a was used. The performance was evaluated by using mean square error as the parameter. One of the important steps in this experiment is to train a neural network. To get a specified output, from the given input it becomes necessary to train the neural network. The train algorithm has been used for training the data. The training algorithm, in particular trainscg which is a conjugate gradient algorithm, seems to achieve more over a wide diversity of complications. The SCG algorithm practically works the same way as the LM algorithm. It differs from trainrp in that when the error is reduced its performance does not diminish as rapidly as compared to trainrp performance. The memory requirements in conjugate gradient algorithms are moderately modest. So, it is the best choice in our case. Keeping the above observations in mind, we have used the train function, trainscg, in our case [23]. In the experiment first the training and testing is done using BPNN and the result is shown followed by training and testing using RBFNN. Neural networks are made up of neurons. To produce the desired output, the neurons use any differentiable transfer function f. For BPNN the tan-sigmoid transfer function is used for hidden layers. The output layer for BPNN is linear. The data is trained using neural network toolbox. Then, simulation is done on the remaining 300 test samples alongside the trained network. The NN used in our experiment can be viewed. The graphical diagram given below (figure 1) shows the neural network of BPNN.

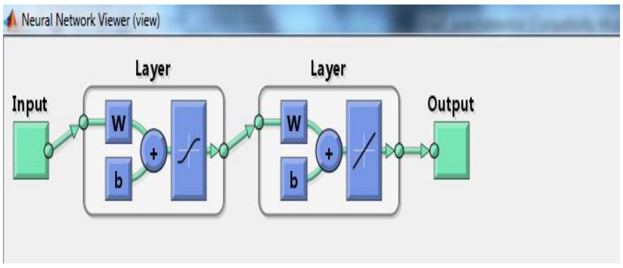

Figure 1: Neural Network Viewer of BPNN

The performance function of the network is measured using mean squared error (MSE). The plot (figure 2) plots the performance of the training data. It clearly depicts that the network is learning. The mean squared error of the network reduces to a very small value in the end as compared to the starting value which is very large in size. During simulation if same data is tested against training data then 100% data is correctly classified. This is shown in the table 1. In our case it shows that if speech samples have been tested against the trained data then none of the data has been wrongly misinterpreted or misclassified. For each and every output class, we plot the ROC which gives the Receiver Operating Characteristic for class. If the curve touches the left side and is near the top edges then it is supposed to be better. The corresponding ROC for the above data is given below in figure 3. From the graph it can be seen that it is touching the left and top of the edges.

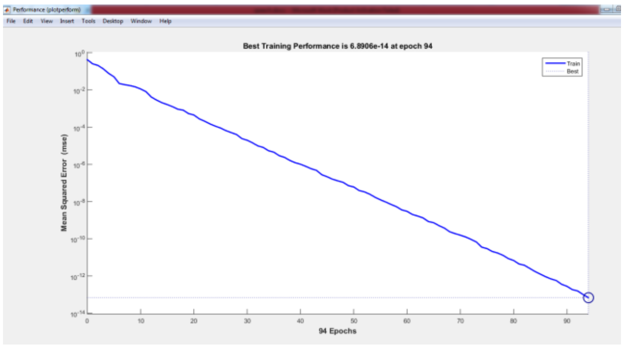

Figure 2: Training the BPNN

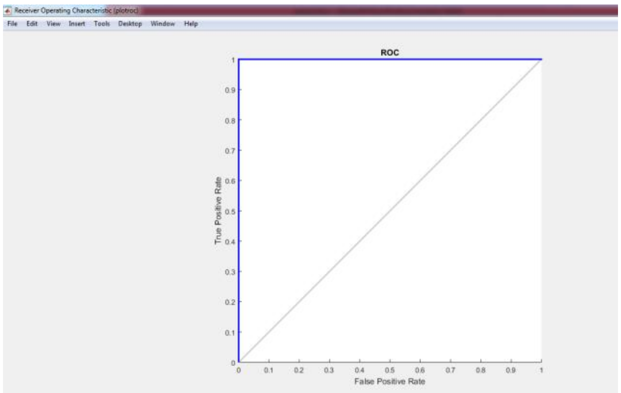

Figure 3: Receiver Operating Characteristic Graph when training and testing data is same in BPNN

Now we simulate the result against a new set of data. The testing data is entirely different from the training data. Here, the performance when the testing and training data is different is not as accurate as when the same data was tested. But still 90% of the data is correctly identified and only 10% of the data is misclassified as shown in table 1. This is a good result.

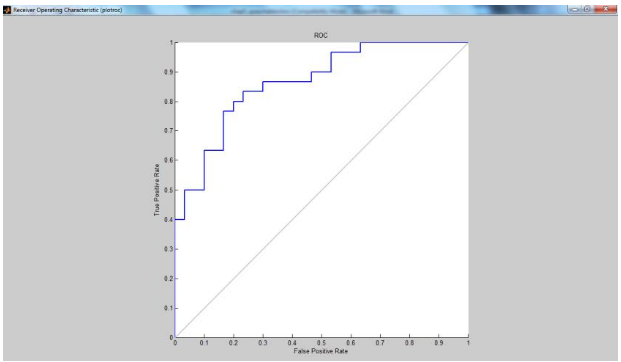

Figure 4: Receiver Operating Characteristic when training and testing data is different in BPNN

The ROC for the data is plotted in Figure 4 and shows that it is very much towards the left and top of the edges but it does not totally touch the edges. Still it shows good performance. For the Radial basis neural network we have Gaussian function for hidden layers and for the output layer it is linear. The neural network for RBFNN is given in Figure 5.

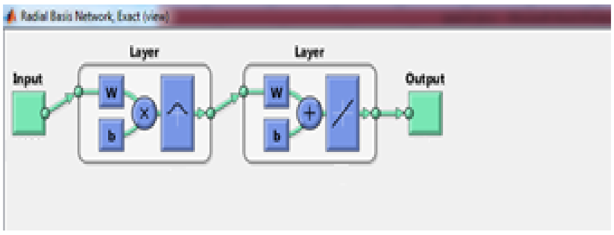

Figure 5: Neural Network Viewer of RBFNN

In the radial basis also it is seen that if same data is simulated against test data then we get 100% results. This is presented in the table 1. In our case it shows that no data is wrongly misinterpreted or misclassified and 100% data is correctly classified. The performance graph for the above data is displayed in figure 6 and the equivalent ROC graph is shown in figure 7. From the graph it can be seen that it is touching the left and top of the edges.

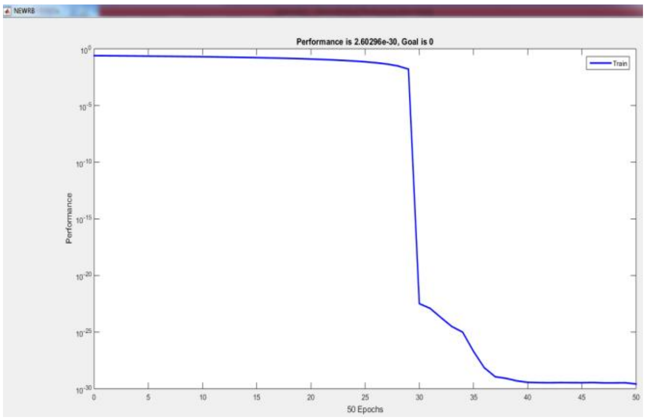

Figure 6: Training the Radial Basis NN

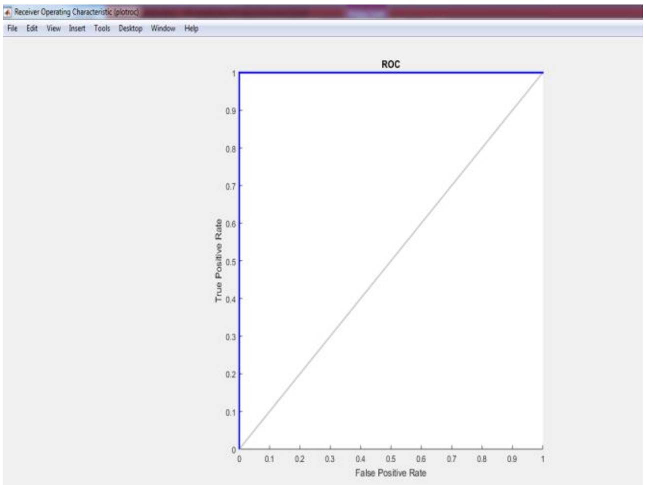

Figure 7: Receiver Operating Characteristic Graph when training and testing data is same in RBFNN

Similarly, we see that once the training and testing data is different, only 50% of the data is correctly identified and rest 50% of the data is misclassified. This is shown in figure 8. This is not a good result.

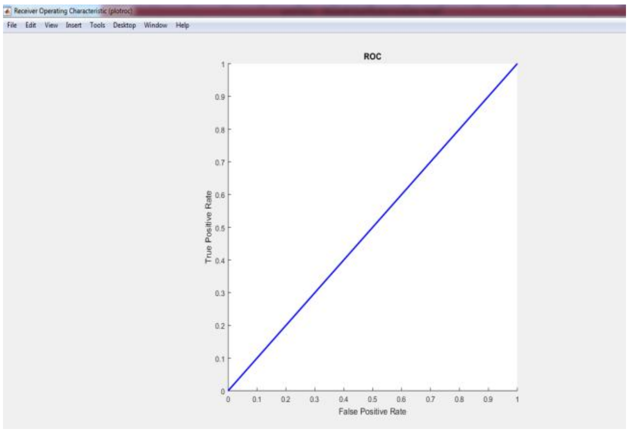

Figure 8: Receiver Operating Characteristic Graph when training and testing data is different in RBFNN

The below table 1 clearly shows the performance of the RBFNN and BPNN and that BPNN outperforms the RBFNN in our case as number of correctly identified cases are more in BPNN when training and testing data is different.

Table 1: Performance analysis of BPNN & RBFNN

Neural Network	Data Trained /Tested	Correctly Classified cases (%)	Wrongly Classified Cases (%)
BPNN	Training data is same as testing data	100	00
	Training data is different from testing data	90	10
RBFNN	Training data is same as testing data	100	00
	Training data is different from testing data	50	50

VI. CONCLUSION

Speech recognition is really very important in many cases. Different methods have been used by different researchers for speech recognition, which we have described in the paper. Some of these methods yield good result. The experimental result in the paper shows successful recognition of speech. For this MFCC has been used along with Neural Network. In this Neural Network toolbox of MATLAB R2015a has been used. The two algorithms of neural network, Feedforward with backpropagation and the Radial basis functions are executed in MATLAB and both successfully recognizes speech, but results shows that BPNN works better than RBFNN. The simulation shows high accuracy in result. However, if still larger training data is used, the result will further improve and we will yield more correct and exact result.

REFERENCES

1. Tanjin Taher Toma, Abu Hasnat Md. Rubaiyat, A.H.M Asadul Huq,"Recognition of English Vowels in Isolated Speech using Characteristics of Bengali Accent", 2nd International Conference on Advances in Electrical Engineering (ICAEE 2013)19-21 December, 2013, Dhaka, Bangladesh, pp. 405-410.
2. L. Singh and S. Sridharan, "Speech Enhancement using Pre-processing" IEEE TENCON - Speech and Image Technologies for Computing and Telecommunications, 1997 pp. 755- 758.
3. Saha. G., Chakroborty. S., Senapati. S, "A new Silence Removal and End Point Detection Algorithm for Speech and Speaker Recognition Applications", in Proc. of Eleventh National Conference on Communications (NCC), IITKharagpur, India, January 28-30, 2005, pp. 291-295.
4. R.Vijayalakshmi, S.Priya, "An Interactive Speech Therapy Session using Linear Predictive Coding in Matlab and Arduino", International Conference on Advanced Communication Control and Computing Technologies, 2016, pp. 217 – 220.
5. Anand H. Unnibhavi, D S Jangamshetti, "LPC Based Speech Recognition for Kannada Vowels", International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques, 2017, pp. 642-645.
6. Akshay Chamoli ; Ashish Semwal ; Nomita Saikia , "Detection of emotion in analysis of speech using linear predictive coding techniques (L.P.C), International Conference on Inventive Systems and Control, 2017, pp. 1-4.
7. Satyam P. Todkar , Snehal S. Babar , Rudrendra U. Ambike , Prasad B. Suryakar, Dr. J. R. Prasad, "Speaker Recognition Techniques: A review", 3rd International Conference for Convergence in Technology (I2CT) Pune, India, 6–7 April, 2018
8. Olivier Siohan, "On The Robustness Of Linear Discriminant Analysis as a Preprocessing Step for Noisy Speech Recognition", International Conference on Acoustics, Speech, and Signal Processing, 1995 Volume: 1 , pp. 125-128
9. Santosh Gaikwad, Bharti Gawali, S.C.Mehrotra, "Novel Approach Based Feature Extraction for Marathi Continuous Speech Recognition", Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM, 2012, pp. 795-804.
10. Xuefeng Jiang, "A facial expression recognition model based on HMM", Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, 12-14 Aug. 2011
11. Ali Zulfiqar, Aslam Muhammad, Martinez Enriquez A. M., "A Speaker Identification System using MFCC Features with VQ Technique" Third International Symposium on Intelligent Information Technology Application, IEEE 2009.
12. Balwant A. Sonkamble, D. D. Doye "Speech Recognition Using Vector Quantization through Modified K-means LBG Algorithm", Computer Engineering and Intelligent Systems, Vol 3, No.7, 2012, pp.137-144.
13. Talal Bin Amin, Iftekhar Mahmood, "Speech Recognition using Dynamic Time Warping", 2nd International Conference on Advances in Space Technologies, IEEE, Islamabad, 29th -30th November 2008, pp. 74-79.
14. Lindasalwa Muda, Mumtaj Begam, I. Elamvazuthi "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" Journal Of Computing, Vol. 2, Issue 3, March 2010
15. Abdul Syafiq B Abdull Sukor, "Speaker Identification System using MFCC Procedure and Noise Reduction Method", University Tun Hussein Onn Malaysia, January 2012.
16. Hamdy K. Elminir, Mohamed Abu ElSoud, L. M. Abou El-Maged, "Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition", International Journal of Science and Technology, Vol. 2 No.10, pp. 689-695, October 2012
17. Ahmad A. M. Abushariah, Teddy S. Gunawan, Othman O. Khalifa, Mohammad A. M. Abushariah, "English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer and Communication Engineering, May 2010.
18. N. Mishra, Mahesh Chandra, Astik Biswas, S. N. Sharan, "Robust Features for Connected Hindi Digits Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 2, June, 2011
19. R. B. Shinde , Dr. V. P. Pawar, "Vowel Classification based on LPC and ANN", International Journal of Computer Applications, Vol. 50 No.6, July 2012, pp. 27-31.
20. Firoz Shah.A, Raji Sukumar.A, Babu Anto.P. "Automatic Emotion Recognition from Speech, Using Artificial Neural Networks With Gender-Dependent Databases", International Conference on Advances in Computing, Control, and Telecommunication Technologies, IEEE, 2009.
21. Georgi T. Tsenov, and Valeri M. Mladenov, "Speech Recognition Using Neural Networks", 10th symposium on Neural network applications in electrical engineering, IEEE, September 23-25, 2010, pp181-186.
22. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.
23. Srivastava Nidhi, "Speech Recognition using Artificial Neural Network", IJESIT, vol.3, 2014.

AUTHORS PROFILE

Dr. Nidhi Srivastava is currently working as Assistant professor in Amity University. She has done her Ph.D. from UPRTOU, Allahabad in the area of Human Computer Interaction. Her other research area is e-governance and Cloud Computing. She has about 14 years of teaching experience. She has published many research papers in various reputed national and international journals and conferences. She is in the editorial and reviewer board of many prestigious journals and conferences and holds life membership of Computer Society of India (CSI) and International Association of Engineers (IAENG).