© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Comparative Study of Multilayer Feed-forward Neural Network and Radial Basis Function Neural Network Models for Speech Recognition

Priyanka Tyagi[1], Dr. Jayant Shekhar[2]

[1]M. Tech Student, Department of Computer Science, Subharti University, Meerut, Uttar Pradesh, India
[2]Professor (Director, SITE), Department of Computer Science, Subharti University, Meerut, Uttar Pradesh, India

ABSTRACT

The most common way of human-to-human communication is speech. As speech provides the easiest and most natural way of interaction, it becomes the need of human-to-machine communication as well. Automatic speech recognition (ASR) is the technology to enable machines to understand process and recognize speech. Due to its applicability in various application domains, ASR becomes one of the most fascinating areas of pattern recognition. In this paper, we are analyzing the performances of multilayer feed-forward neural network and Radial basis function neural network models for the recognition of speech signals. The work is conducted in four stages: speech signal acquisition & pre-processing, feature pattern vector creation, implementation & training of selected neural network models and comparative analysis of the performances of selected neural networks.

Proposed work is conducted with 10 speech samples of English alphabets .Digital signal processing operations are applied on signals to convert them and make them appropriate for further processing. Five feature pattern vectors are created to be used for training and testing of the network models. Performance of selected neutral network models is measured and analyzed for the created feature pattern vectors. Results indicate that feed-forward neural network model performs better than the Radial basis function neural network for all the test pattern vectors.

Keywords : Automatic Speech Recognition, Digital Signal Processing, Sampling, Quantization, Feed-Forward Neural Network, Radial Basis Network.

I. INTRODUCTION

Speech is the most essential and effective medium of human-to-human communication. It provides human beings a way is to express their feelings, thoughts and views etc. Speech also conveys information about the personality, identity and linguistic information of a human being. Now-a-days, speech also becomes an important medium of communication between man and machine [1]. Due to the increasing popularity and acceptance of speech, automatic speech recognition has become one of the most interesting areas of pattern recognition. A number of speech recognition techniques have been used in variety of applications of engineering and scientific fields [2].

Automatic speech recognition is a technology to make machine learn to recognize the speech spoken by a person and convert it into the text form. The features of the speech may depend on characteristics of the person or the environment like speaking speed, noise elements, etc. These features may affect the functioning of the speech recognition system. Other factors which may influence the working of a speech recognition system are the varying speaking styles or accents, age, emotional state of speaker/s, etc. [3]. Hence to handle these problems, the modular structure of speech recognition system can be considered to be similar as the human mechanism to speech perception.

An automatic speech recognition system consists of following five modules like signal acquisition, signal pre-processing, feature extraction, speech classification and speech recognition. Accuracy of a speech recognitions system depends on a number of parameters such as environmental noise, acoustical distortions, various ways of speaking like shouting&whispering, size of the vocabulary and so on. Based on the types of speeches, the most common types of speech recognition systems applied in the area of automatic speech recognition are given below:

1. ***ISOLATED WORD RECOGNITION SYSTEM-*** The isolated word recognition system accepts either a single word or single sound at a particular instance of time. These systems have only two states- listen and not-listen[4].
2. ***CONNECTED WORD RECOGNITION SYSTEM*** –These systems works by processing two or more sounds or words together. Words or sounds should have short silence between them[5].
3. ***CONTINUOUS SPEECH RECOGNITION SYSTEM*** –These systems work by accepting continuous speech between two sounds and there is no time boundary limitation on the input speech [6]. The most common problem with these systems is that it is difficult to determine the boundary of the utterance.

Till date, a number of sincere efforts have been done to enable machines to recognize speech. Although, earlier attempts were focused on to design machines which can speak instead of understand or recognize speech. In 1879, Charles Wheatstone built a speaking machine using resonators made of leather to produce various speech –like sounds[7]. Since then, a lot of work has been done in this area. A review of efforts performed in the last decade is presented next.

Davis et al. [8] built a system for isolated digit recognition for a single speaker. The system worked by measuring the spectral resonances during the vowel region of each digit. Atal&Rabinder [9] assigned the input speech signal to one of the three predefined output classes – voiced, unvoiced and silence by applying the minimum distance rule. Five measurements such as zero-crossing rate, the speech energy, the correlation between adjacent speech samples, the first predictor coefficient from a 12-pole LPC (linear predictive coding) analysis and the energy were taken to determine the error between the actual and desired output. Rabiner et al. [10] performed a series of experiments to design speaker independent speech recognition system by applying a wide range of clustering algorithms. Experiments were done by representing all variations of different words across a wide user population. Erell et al. [11] derived a spectral estimation algorithm to improve the robustness of the speech recognition system with additive noise such as environmental noise.

Hidden Markov Models based phonemes recognition techniques were proposed by Cui et.al. [12]. Feature pattern vector was created by extracting features from a single phoneme. A phoneme is defined as the smallest unit of speech that distinguishes a meaning. An interactive voice response based back-propagation neural network system for the classification of audio signals with 90% recognition accuracy was presented by Shah et al.[13]. In this system, feature extraction was performed by applying the Mel Frequency Cepstral Coeffiecient (MFCC) method. An investigation of the performance of Feed-forward neural network and Radial basis function neural network for speech recognition is also performed in [1].

Despite of a huge amount of research work done in the area of automatic speech recognition, still there is some space left for systems with good recognition accuracy.

In this paper, we are investigating the performance of Multilayer feed-forward back-propagation neural network and Radial basis function neural network models for the speech recognition of alphabets of English language. A comparative analysis of performances of both neural network models for noiseless and noisy input speech samples is also done.

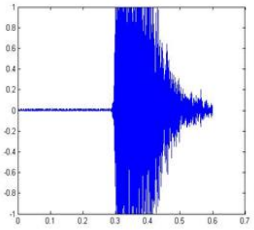
This paper is further organized in five sections. In section 2, feature extraction method for the input speech samples presented for the experiment is discussed. In section 3, implementation details of selected neural network models used for the work are given. Section 4 presents the simulation results, comparative study of recognition accuracy &performances of the selected neural network models and a complete discussion of the results. Section 5 considers the conclusion followed by references.

II. METHODS AND MATERIAL

Feature Extraction

Feature extraction phase is equally important as the classification and recognition phase of a recognition process. Proper functioning and performance of a recognition system depends on the successful fulfillment of this phase of the system. Feature extraction is defined as the *"problem of extracting from the raw data the information which is most relevant for classification purposes, in the sense of minimizing the with-in class pattern variability while enhancing the between-class pattern variability"* [14].During the feature extraction phase, significant and meaningful features are extracted from a number of available features. These features should be independent of the size, orientation and location of the pattern for the smooth functioning of the recognition system. Thus, the goal of feature extraction process is to create an optimal feature vector to support classification process and maximize the efficiency of the recognition process. Commonly used feature extraction techniques used for speech signal processing are Linear predictive cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCCs) and Perceptual linear predictive coefficients (PLP) [15].

Data set used in the present experiment consists of 10 speech signal samples English letters 'A', 'B', 'C', 'D' and 'E'. All input speech signals are collected as audio files and each audio file is stored with the extension *.wav*. A set of 5input signals is presented in figure 1.



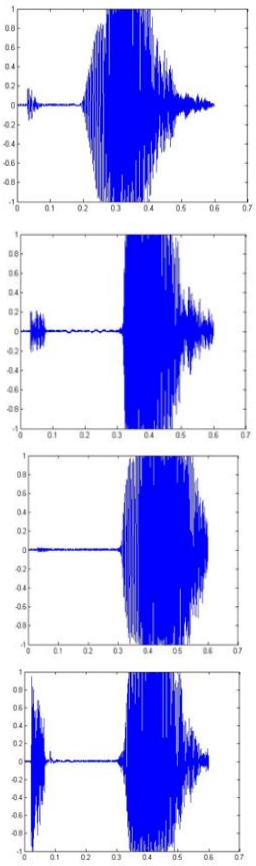


Figure1 : A set of five input signal samples

Training pattern vectors are created by applying sampling, quantization and coding operations on collected input signal samples. Sampling is the process of obtaining signal values from continuous signal at regular time intervals. It is performed to convert the continuous-time analog input signal samples to the discrete-time signals as [16]:

$$x(n) = x_a(nT) \qquad (1)$$

Where x[n]is the discrete time sequence signal
n is the sample index
T is the sampling interval

T between successive samples is called the sampling period.

The amplitude of x[n] obtained in this equation is known with infinite precision [17].To represent each value an infinite number of digits are required instead of a finite number of digits. Quantization converts a continuous–amplitude signal into a

discrete-amplitude signal. Thus, the amplitude of discrete-amplitude signal is known with finite precision. Let $x_q(n)$ denotes sequence of quantized samples after the quantization process as:

$$x_q(n) = Q[x(n)] \quad (2)$$

where x(n) represents the samples.

Result of the quantization process is the digital signal. Difference between the discrete-time signal and digital signal is called the quantization error (e_q) [17]. Following assumptions are considered about the quantization error:

- The error sequence [$e_q(n)$] is a stationary white noise sequence.
- The error sequence [$e_q(n)$] is uncorrelated with the signal sequence x(n)
- The signal sequence is of zero mean and stationary.

After quantization, coding process is used to for the digital representation of the signal. Coding assigns a unique binary number to each quantization level. So, for L levels, at least L different binary numbers are required.

Training feature pattern vectors are created by applying these operations on input speech signals. The test pattern vectors by introducing the 10%, 20%,30%and 40% noise respectively to training pattern vectors of speech signals.

Implementation of Neural Network Models

An artificial neural network or ANN is a computational model which is designed to perform the complex pattern recognition tasks such as pattern classification, pattern mapping, pattern association, etc. [18] ANN can be defined as a massively parallel distributed processor, contains a large number of highly interconnected processing units, called neurons which process the information and works in combination to solve a specific problem and accomplish a particular task. Neural networks store information in the strengths of the interconnections. Neurons are arranged in layers and each derives its input from one or more other neurons and/or external sources. Number of nodes in input layer depend on the number of elements in input feature vector provided to the network, number of nodes in output layer depends upon the number of classes in which data is to be classified and number of nodes in hidden layer are decided by the number of nodes in input layer and complexity of the problem to be solved. Each neuron consists of a summing part, a threshold value and an output part. Directed communication links exists between neurons and each link is associated by a weight, whose value represents the strength of the connection between the units. The basic model of a neuron is presented in figure 2.

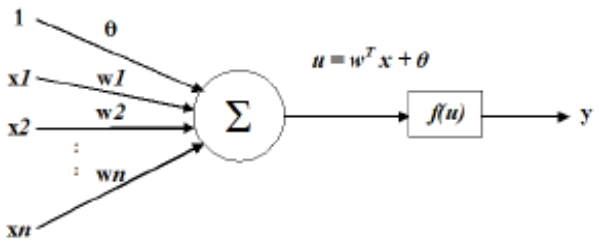


Figure 2 : Basic model of a neuron

The summing part of each neuron receives input values from n inputs x_1, x_2, x_3,…..,x_n, weights each input value and calculates the weighted sum, called the activation value of the neuron as shown in the following equation:

$$O = f(net) = f\left(\sum_{i=1}^{n} w_i x_i\right) \quad (3)$$

where O is the output signal,

w_i is the weight vector and
$f(net)$ is the activation function

The activation value of the neuron is compared with predefined threshold value. Based on the result of the comparison, output value is obtained -

$$O = f(net) = \begin{cases} 1, & net \geq \theta \\ 0, & otherwise \end{cases} \quad (4)$$

where $\boldsymbol{\Theta}$ is the threshold value

This output may be given as input to other units as well as the unit which produces it.

In this paper, we used multilayer feed-forward back-propagation neural network and Radial basis function neural network. Back-propagation is a supervised learning algorithm and belongs to a class of "learning

with the teacher"[19]. Back-propagation is a systematic method of training multilayer artificial neural networks in which a predefined desired target output (t) for each input patter is prepared. This target output is compared with the actual output (o) and difference is termed as error (E). The value of the error term is propagated backward from the output layer to hidden layer/s to update the weights in the hidden layer/s.

The first neural network model we used is Multilayer feed-forward neural network. Network is trained with Levenberg-Marquardt learning. In this experiment, size of input pattern vector P and target output vector T is same. The parameters used for the architecture and training of the network are presented in table 1.

Table 1 : Parameters used for creating the Multilayer Feed-forward network

Parameter	Value
Number of hidden layers	3
Number of neurons in first hidden layer	7
Number of neurons in second hidden layer	11
Number of neurons in third hidden layer	15
Number of neurons in output layer	10
Transfer function for first layer	Hyperbolic Tangent Sigmoid
Transfer function for second layer	Hyperbolic Tangent Sigmoid
Transfer function for third layer	Hyperbolic Tangent Sigmoid
Training function	Levenberg-Marquardt
Maximum number of epochs	1000
Performance function	Mean squared error
Error goal	0.00001
Adaption rate	1
Back-propagation learning rate	0.1
Initial weights and biased term values	Values generated randomly between 0 and 1

Second neural network model used in this work is Radial basis function network (RBF). RBF network is a three layer feed-forward neural network and consists of a single hidden layer in its structure (as shown in figure 3), where hidden layer is non-linear and output layer is linear [20]. Due to the non-linear characteristics, RBF is able to model the complex pattern mapping problems. In this network, the number of neurons in the first layer is less than the number of samples and each unit implements a radial basis function such as Gaussian radial function, Quadratic function, Inverse quadratic function, Thin plate spline, etc. Activation function of the hidden layer computes the Euclidean distance between the input vector and center of that unit and the value of the function increases or decreases monotonically with the distance from a center point.

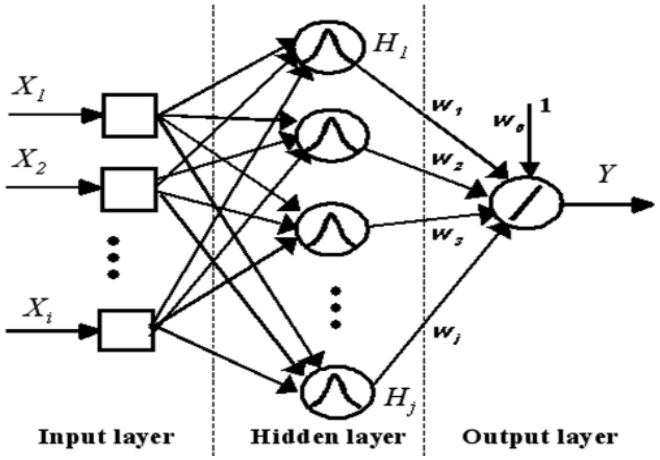


Figure 3 : Radial basis function neural network

Gaussian function is widely used to compute the activation value for the unit of the middle layer. Gaussian activation function for the RBF network can expressed as [21]:

$$z_j(X_l) = \exp\left[-\frac{\| X_l - \mu_j \|^2}{2\sigma_j^2}\right]$$

for j = 1,2,3,……, M

where ||.|| : denotes Euclidean norm

x : N-dimensional input vector,

σ_j : width of the neuron, and

μ_j : mean of the j^{th} Gaussian function.

Two-fold learning is performed by updating the position and spread of centers as:

$$c_{ij}(t+1) = c_{ij}(t) - \eta \frac{\partial E}{\partial c_{ij}} \qquad (6)$$

for i=1 to size of the input pattern vector, j=1 to h and updating weights w to produce the desired output related to the input pattern vectors as:

$$w_i(t+1) = w_i(t) - \eta \frac{\partial E}{\partial w_i} \qquad (7)$$

where η : learning parameter

The parameters used for the architecture and training of the Radial basis function network are presented in table 2.

Table 2 : Parameters used for creating the Radial basis function network

Parameter	***Value***
Performance function	Mean squared error
Spread of Radial basis function	1.0
Number of neurons in layer 1	10
Number of neurons in layer 2	4800
Transfer function in layer 1	Radial basis transfer function
Transfer function in layer 2	Linear transfer function
Back-propagation learning rate	0.1

III. RESULTS AND DISCUSSION

In the proposed simulation, we are analyzing the performance of feed-forward neural network and Radial basis function network models for created training pattern vectors and test pattern vectors of speech signals. The results presented in the simulation are considered from both selected feed-forward multilayer neural network models. Performance of these neural network models for the training patterns presented in table 3.

Table 3: Regression value of training pattern vectors for Multilayer feed-forward network and Radial basis function network

Network	Signal									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Multilayer Feed-forward Network	.6862	.8699	.9717	.975	.8364	.7183	.9495	.9718	.9032	.9697
Radial Basis Function Network	1	1	1	1	1	1	1	1	1	1

Graphical presentation of regression values of training pattern vectors presented to multilayer feed-forward network and Radial basis function network is shown in figure 4.

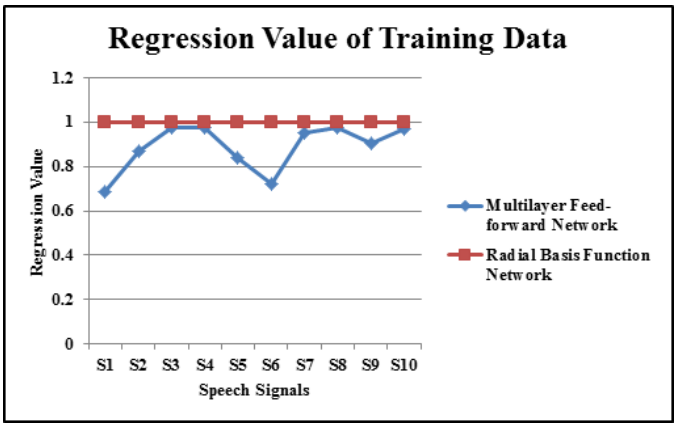


Figure 4: Comparison of regression value of training pattern vectors for Multilayer feed-forward network and Radial basis function network

Testing pattern vectors are created by. Performances of selected networks for testing patterns created by introducing 10%, 20%, 30% and 40% error respectively in training pattern vectors of all signals are presented in table 4,5,6 and 7 respectively.

Table 4: Regression value of testing pattern vectors with 10% error for Multilayer feed-forward network and Radial basis function network

Network	Signals with 10% noise									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Multilayer Feed-forward Network	.7038	.7558	.7688	.7715	.7656	.6402	.6342	.6959	.6311	.6722
Radial Basis Function Network	.9904	.0898	.0111	.9865	.0072	1	.0045	.0008	.0198	.0017

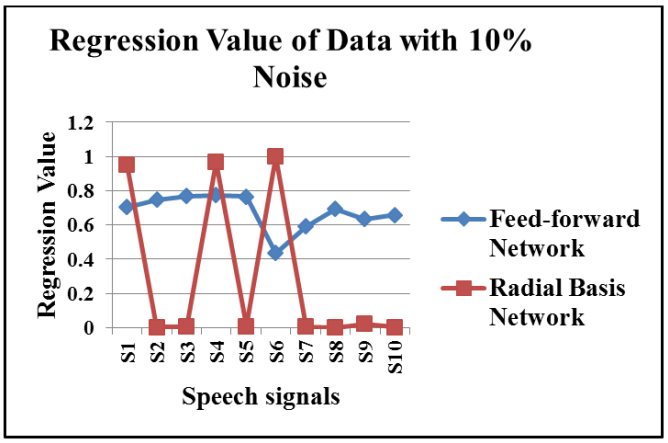


Figure 5: Comparison of regression value of testing pattern vectors with 10% error for Multilayer feed-forward network and Radial basis function network

Table 5: Regression value of testing pattern vectors with 20% error for Multilayer feed-forward network and Radial basis function network

Network	Signals with 20% noise									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Multilayer Feed-forward	.7038	.7558	.7688	.7715	.7656	.5849	.6334	.6947	.6287	.6579
Radial Basis Function Network	.9766	.0829	.00002	.9795	.0072	1	.0045	.0008	.0198	.0017

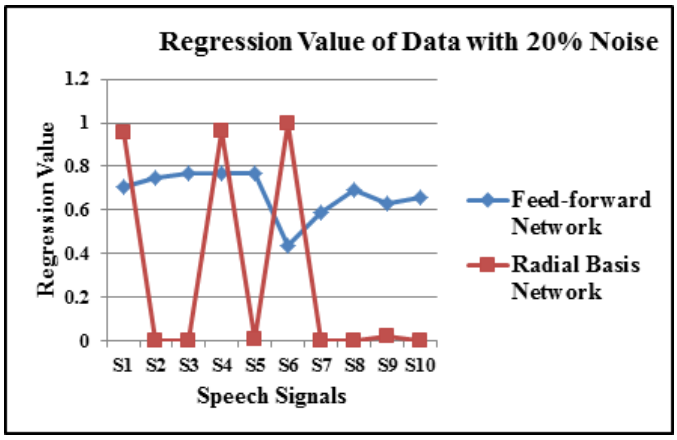


Figure 6: Comparison of regression value of testing pattern vectors with 20% error for Multilayer feed-forward network and Radial basis function network

Table 6: Regression value of testing pattern vectors with 30% error for Multilayer feed-forward network and Radial basis function network

Network	Signals with 30% noise									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	10
Multilayer Feed-forward Network	.7038	.7558	.7688	.7715	.7656	.4881	.6377	.6947	.6315	.6579
Radial Basis Function Network	.9651	.0807	.0010	.9735	.0071	1	.0045	.0008	.0198	.0017

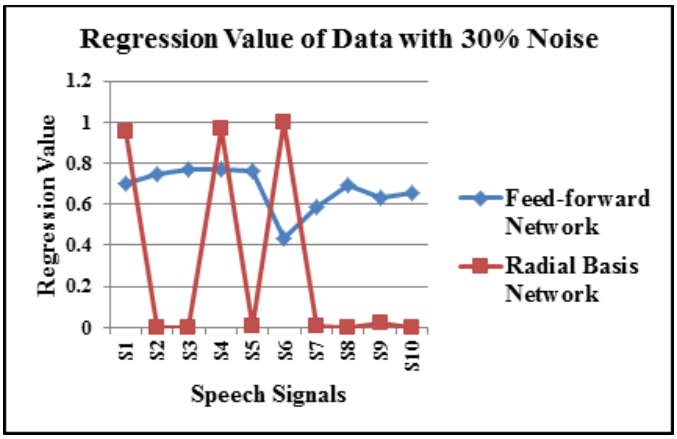


Figure 7: Comparison of regression value of testing pattern vectors with 30% error for Multilayer feed-forward network and Radial basis function network

Table 7: Regression value of testing pattern vectors with 40% error for Multilayer feed-forward network and Radial basis function network

Network	Signals with 40% noise									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Multilayer Feed-forward Network	.7038	.7486	.7688	.7715	.7656	.4349	.5901	.6905	.6319	.6579
Radial Basis Function Network	.9626	.0027	.003	.9654	.0071	1	.0045	.0008	.0198	.0017

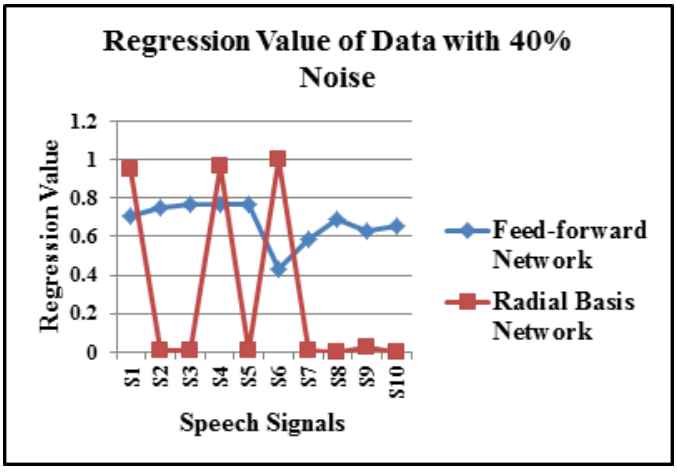


Figure 8: Comparison of regression value of testing pattern vectors with 40% error for Multilayer feed-forward network and Radial basis function network

Tables-3 is showing the comparison of performance of the multilayer feed-forward network and Radial basis function network models for the training pattern signals, while tables 4, 5, 6 and 7 are showing the comparison of performance of these networks for test pattern vectors with 10%, 20%, 30% and 40% error.

Results given in table 3displayed that the multilayer feed-forward neural network performs with 68% to 97% recognition accuracy, while Radial basis function network model performs with 100% accuracy for all the training signal pattern vectors.

Results also show that multilayer feed-forward neural network show 60% - 75% recognition accuracy for most of the test signal pattern vectors, while Radial basis function network shows poor classification accuracy for all test patterns if noise or error is more than 10%. Hence, results are indicating that the Radial basis function network does not perform well for signal classification due to its large sample size.

IV. CONCLUSION

In this paper we analyzed the performance of Multilayer feed-forward network and, Radial basis function network for the classification of speech signals of first five alphabets of the English language. Training pattern vectors are created by applying digital signal processing operations like sampling, quantization and coding to the input speech signals respectively. Test pattern vectors are created by adding 10%, 20%, 30% and 40% error respectively in the input signals used for training. Simulated results of the performance evaluation of the selected networks are presented and discussed. The following observations have been drawn from the simulated performance evaluation.

(i) The simulated results are also indicating that the Multilayer feed-forward neural network model shows above 68% recognition accuracy for training signal patterns. The networks shows similar recognition accuracy for test pattern vectors S1, S2, S3, S4 and S5. The network shows lowest recognition accuracy for signal S6, when the percentage of error increases.

(ii) The highest and lowest recognition accuracy presented by the network is 77% for the signal S4 and 43% for the signal S6 respectively.

(iii) The simulated results are also exhibiting that the Radial basis function neural network model shows 100% recognition accuracy for training pattern vectors and all test pattern vectors created for signal S6.

(iv) Results indicate that Radial basis function network shows similar recognition accuracy for test pattern signals S5, S6, S7, S8, S9 and S10. Thus, the network is showing poor behavior of generalization for large data samples.

(v) Simulation results are indicating that multilayer feed-forward neural network exhibit good approximation and generalization.

Simulation results are showing that Radial basis function network exhibit a good approximation but poor generalization.

V. REFERENCES

[1] WouterGevaert, GeorgiTsenov and ValeriMladenov, “Neural Networks used for SpeechRecognition”, Journal of Automatic Control, pg. 1-7, Vol. 20, 2010.

[2] Gerasimos Potamianos, ChalapathyNeti, Guillaume Gravier, AshutoshGarg and Andrew W. Senior, “Recent Advances in the Automatic Recogntion of Audio-Visual speech’, Proceedings of the IEEE, pp. 1-18, Vol. 91, No. 9, 2003.

[3] Nidhi Srivastava, “Speech Recognition Using Artificial Neural Network “, International Journal of Engineering Science and Innovative Technology (IJESIT), pg. 406-412, Vol. 3, Issue 3, 2014.

[4] M.A.Anusuya and S. K. Katti, “Speech Recognition by Machine: A Review”, Int. Journal of Computer Science and Information Security,pg. 181-205, Vol. 6, No. 3, 2009.

[5] Santosh K. Gaikwad, Bharti W. Gawali and PravinYannawar, “A Review on Speech

[6] Recognition Technique”, International Journal of Computer Applications, pg. 16-24, Vol. 10, No. 3, 2010.

[7] T. Landauer, C. Kamm, and S. Singhal, "Learning a Minimally Structured Backpropagation Network to Recognize Speech," In Proceedings of Ninth Annual Conference of Cogn. Sc.Soc., pp. 531–536, 1987.

[8] Sir Charles Wheatstone, The Scientific Papers of Sir Charles Wheatstone, London: Taylorand Francis, 1879.

[9] K.H.Davis, R.Biddulph, and S.Balashek, "AutomaticRecognition of spoken Digits", Acoust.Soc.Am.,24(6):637-642,1952.

[10] Bishnu S. Atal and Lawrence R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, pp 201-212, Vol. ASSP-24, No. 3, 1976.

[11] L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, andJ.G.Wilpon, "Speaker Independent Recognition ofIsolated Words Using Clustering Techniques", IEEETrans. Acoustics, Speech, Signal Proc., ASSP-27:336-349, 1979.

[12] AdoramErellet.al.,"Energy Conditioned SpectralEstimation for Recognition of Noisy Speed" , IEEETransactions on Audio, Speech and Language processing,Vol.1,No.1, Jan 1993.

[13] Xiaodong Cui et.al., "A Study of Variable-ParameterGaussian Mixture Hidden Markov Modeling for NoisySpeech Recognition", IEEE Transactions On Audio,Speech, And Language Processing, Vol. 15, No. 4, 2007.

[14] Syed Ayaz Ali Shah, Azzam ul Asar and S.F. Shukat, " Neural Network Solution for Secure Interactive Voice Response", World Applied Sciences Journal, pg. 1264-1269, Vol. 6, No. 9, 2009.

[15] Shih F.Y., "Image Processing and Pattern Recognition - Fundamentals and Techniques", Wiley Pub. 2010.

[16] Sandrine Revaz, "Statistical Models in Automatic Speech Recognition", Master's Thesis, Department of Mathematics, University of Fribourg Idiap, 2015.

[17] John G. Proakis and Dimitris G. MAnolakis, "Digital Signal Processing – Principles, Algorithms and Applications", Prentice Hall Publication, Third Edition, 2005.

[18] DimitrisManolakis and Vinay Ingle, "Applied Digital Signal Processing – Theory and Practice", Cambridge University Press, First Edition, 2011.

[19] Yagnanarayana B., "Artificial Intelligence", Prentice Hall Pub., Ninth Edition, 2004.

[20] Jesus O. D. and Hagan M. T., "Backpropagation Algorithms for a Broad Class of Dynamic Networks", IEEE Transactions on Neural Networks, pp. 14-27, Vol. 18, no. 1, 2007.

[21] Powell M.J.D., "Radial Basis Functions for Multivariate Interpolation: A Review", In Algorithms for the Approximation of Functions and Data, J.C. Mason and M.G. Cox, eds., Clarendon Press, pp. 143-167, 1987.

[22] Chen S., Cowan C.F.N. and Grant P. M. "Orthogonal Least Square Learning Algorithm for Radial Basis Function Networks", IEEE Transactions on Neural Networks, pg. 302-309, Vol.2, No. 2, 1991.