International Journal of Science, Environment and Technology, Vol. 3, No 1, 2014, 291 – 302
ISSN 2278-3687 (O)

ARTIFICIAL NEURAL NETWORK BASED PATHOLOGICAL VOICE CLASSIFICATION USING MFCC FEATURES

V. Srinivasan[1], V. Ramalingam[1] and P. Arulmozhi[2]
[1]Professor, Department of Computer Science and Engineering,
Annamalai University, Tamilnadu – 608002
[2]M.E. (CS & E.) IV Sem, Annamalai University, Tamilnadu -608002

Abstract: The analysis of pathological voice is a challenging and an important area of research in speech processing. Acoustic voice analysis can be used to characterize the pathological voices with the aid of the speech signals recorded from the patients. This paper presents a method for the identification and classification of pathological voice using Artificial Neural Network. Multilayer Perceptron Neural Network (MLPNN), Generalised Regression Neural Network (GRNN) and Probabilistic Neural Network (PNN) are used for classifying the pathological voices. Mel-Frequency Cepstral Coefficients (MFCC) features extracted from audio recordings are used for this purpose.
Keywords: Pathological voice, MFCC, MLPNN, GRNN, PNN.

1. INTRODUCTION

Speech refers to the processes associated with the production of sounds used in spoken language. Speech signal is produced as a result of time varying excitation of the time varying vocal tract system. Speech pathology is a field of the health science which deals with the evaluation of speech, language, and voice disorders. Over the past few years, a considerable number of studies have focused on the extraction of acoustic parameters for this objective and automatic features have been utilized in the time and frequency domains. Among acoustic parameters, the most important ones are pitch, jitter, and shimmer. These parameters are based on the fundamental frequency. In recent years, Mel Frequency Cepstral Coefficient (MFCC) has been reported as a very successful parameter for pathological voice detection. Diagnosis of pathological voice is one of the most important issues in biomedical applications of speech technology.

1.1 Related works

In the recent works of speech pathology discrimination, researchers are mostly concentrating in the implementation of feature extraction techniques and pattern classification techniques. The ability of acoustic parameters like pitch, Formants, Jitter and Shimmer [1] in the

*Received Dec 31, 2013 * Published February 2, 2014 * www.ijset.net*

discrimination of normal voices from pathological voices has been discussed. The classification of pathological voice from normal voice is implemented using support vector machine (SVM). A Genetic Algorithm (GA) based feature selection is utilized to select best set of features which improves the classification accuracy. A classification technique is proposed [2] which focus on the acoustic features of the speech using Mel frequency cepstral coefficient and Gaussian Mixture Model. This paper presents the detection of vocal fold pathology with the aid of the speech signal recorded from the patients. An approach [3] investigates the adaptation of Mel-frequency Cepstral Coefficients (MFCC) and Support Vector Machine (SVM) for the diagnosis of neurological disorders in the voice signal. A method [4] based on hidden markov model is used to classify speeches into two classes: the normal and the pathological. Two hidden markov models are trained based on these two classes of speech and then the trained models are used to classify the dataset.

The vocal quality of a group of men and women with and without voice disorders, are analysed [5] based on evaluations of a group of judges experienced in the field of vocal rehabilitation. The use of Modulation Spectra [6] in the voice pathology detection and classification has been discussed. The effect of voice source parameters in the analysis of the LF model parameters from the database of pathological voice has been presented [7]. Pathological voice consists of voice signals from the patients who have disease on their vocal folds.

2. SCOPE OF THE WORK

This paper focuses on the classification of pathological voices from the normal voices using MFCC features and Artificial Neural Network. Mel-frequency cepstral coefficients are the widely used features to characterize voice signals. The cepstral representation of the signal allows us to characterize the vocal tract as a source-filter model and the Mel frequency characterizes the human auditory system. The MFCC feature vectors are used by Artificial Neural Network to identify and classify the pathological voices.

3. PROPOSED WORK

The speech samples for extracting the MFFCs were obtained from audio recordings of people with normal voice and pathological voice. In our implementation, two requirements are imposed. First, the features have to be efficient in terms of measurement and time. Second, both the vocal tract and excitation source information have to be included. The MFCC features are obtained by a standard short-term speech analysis, along with frame-level pitch, to form the feature vectors. Artificial Neural network classifiers (MLPNN, GRNN and

PNN) are considered for the assessment of feature vectors. The architecture of the proposed system is given in Figure 3.1

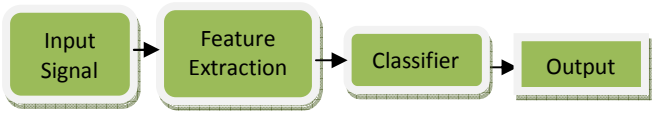

Figure 3.1: System Architecture of the proposed system

3.1. Feature Extraction

Features play a major role in identifying the voice and making a decision that highly depends on how much we are successful in extracting useful information from the voice in a way that enables our system to differentiate between voices and identify them according to their feature. This is accomplished by extracting Mel frequency Cepstral Coefficients (MFCC) features. Figure 3.2 shows the block diagram of extraction of Mel Frequency Cepstral Co-efficients (MFCC).

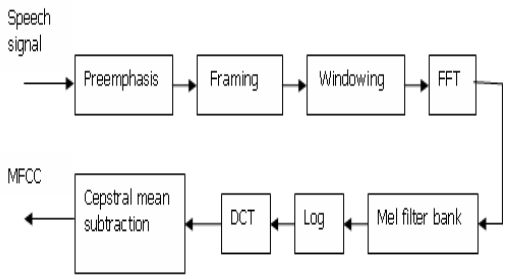

Figure 3.2: MFCC feature extraction

Pre-emphasis - To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. The aim of this process is to boost the amount of energy in the high frequencies. The drop in energy across frequencies (which is called spectral tilt) is caused by the nature of the glottal pulse. Boosting the high frequency energy makes information from these higher formants available to the acoustic model. The pre-emphasis filter is applied on the input signal before windowing.

Framing - It is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behavior within the short time period of 20-40 ms. It is shown in Figure 3.3.

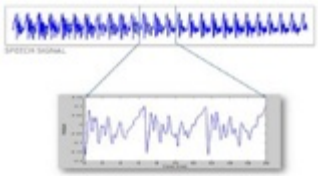

Figure 3.3: Framing the speech signal

Windowing - Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame. Each frame is multiplied by an N sample window W (n). Here we use a hamming window. This hamming window is used to minimize the adverse effects of chopping an N sample section out of the running speech signal. While creating the frames the chopping of N sample from the running signal may have a bad effect on the signal parameters. To minimize this effect windowing is done. Figure 3.4(a) shows the widely used Hamming window and a single frame is multiplied by hamming window and the resulting signal is shown in Figure 3.4(b).

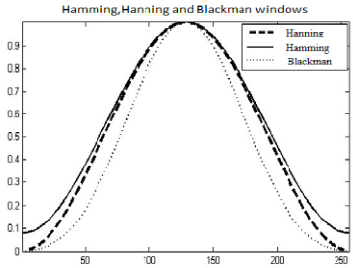

Figure 3.4(a): Hamming window for speech signal

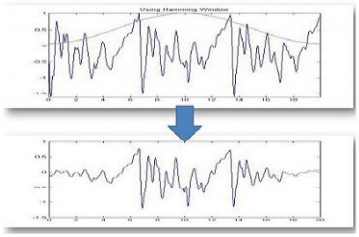

Figure 3.4(b): Windowing for speech signal

Fast Fourier Transform - The basis of performing Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain. When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but this is likely to introduce undesirable effects in the frequency response.

Mel Scaled Filter Bank - The Mel-frequency scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. We can use the Equation 1 to compute the Mel for a given frequency f in Hz:-

$$Mel(f) = 2595\ \log_{10}(1 + \frac{f}{700}) \qquad \ldots\ldots \quad (1)$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel-frequency interval shown in Figure. 3.5.

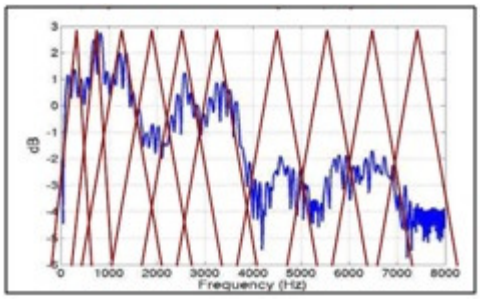

Figure 3.5: Mel Filter banks

Logarithm - The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition

Discrete Cosine Transform - The signal is real (we took the magnitude) with mirror symmetry. The IFFT needs complex arithmetic, the DCT does not. The DCT implements the same function as the FFT more efficiently by taking advantage of the redundancy in a real signal. The DCT is more efficient computationally. The MFCCs may be calculated using this Equation 2

$$X(k) = \propto (k) \sum_{n=0}^{N-1} x(n)\cos((2n+1)k\,\pi/\ 2N); \qquad 0 \le k \le N-1 \ldots\ldots \quad (2)$$

Since the log power spectrum is real and symmetric, inverse FFT reduces to a Discrete Cosine Transform (DCT). By applying the procedure described above, for each speech frame of about 30 ms with overlap, a set of Mel-frequency cepstrum coefficients is computed. This set of coefficients is called an acoustic vector.

3.2 CLASSIFICATION

Multilayer Perceptron Neural Network (MLPNN)

MLPNN is composed of three layers consisting of an input layer, one or more hidden layers and an output layer. The input layer distributes the inputs to subsequent layers. Input nodes have linear activation functions and no thresholds. Each hidden unit node and each output

node have thresholds associated with them in addition to the weights. The hidden unit nodes have nonlinear activation functions and the outputs have linear activation functions. The number of neurons in the hidden layer is dependent on the size of the input vector. The output layer has one neuron. MFCC features from the normal and pathological speech samples are input to the ANN for training. This is used for classifying the pathological voice from the normal voice.

The input samples are propagated in a forward direction on a layer-by-layer basis. The network computes its output pattern, and if there is an error − or in other words a difference between actual and desired output patterns − the weights are adjusted to reduce this error. In a back-propagation neural network, the learning algorithm has two phases. First, a training input pattern is presented to the network input layer. The network propagates the input pattern from layer to layer until the output pattern is generated by the output layer. If this pattern is different from the desired output, an error is calculated and then propagated backwards through the network from the output layer to the input layer. The weights are modified as the error is propagated. This process is explained below

Step 1: Initialisation - Set all the weights and threshold levels of the network.

Step 2: Activation - Activate the back-propagation neural network by applying inputs and desired outputs. Calculate the actual outputs of the neurons in the hidden layer and the output layer.

Step 3: Weight training - Update the weights in the back-propagation network propagating backward the errors associated with output neurons. Calculate the error gradient for the neurons in the output layer and the hidden layer.

Step 4: Iteration - Increase iteration one by one, go back to *Step 2* and repeat the process until the selected error criterion is satisfied.

Figure 3.6. represents the actual activity of the neuron cell. All inputs are summed together and modified by the weights. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output.

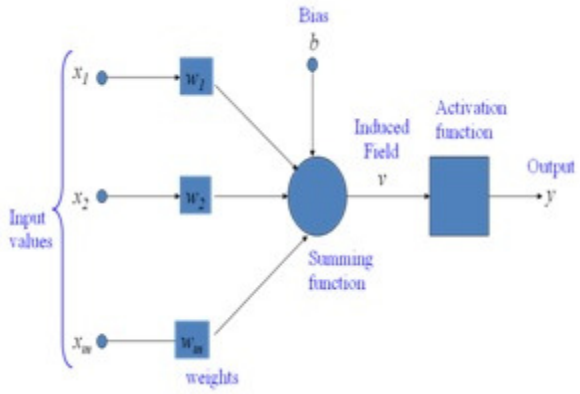

Figure 3.6: MLPNN Model

General Regression Neural Network (GRNN)

A GRNN network is a three-layer network that contains one hidden neuron for each training pattern. There is a smoothing factor that is used when the network is applied to new data. The smoothing factor determines how tightly the network matches its predictions to the data in the training patterns. Figure 3.7 shows the Structure of the GRNN.

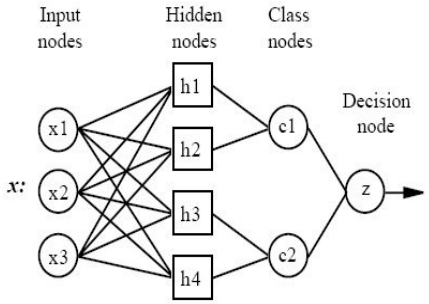

Figure 3.7: General Structure of GRNN

Input layer: There is one neuron in the input layer for each predictor variable. In the case of categorical variables, *N*-1 neurons are used where *N* is the number of categories. The input neurons then feed the values to each of the neurons in the hidden layer.

Hidden layer: This layer has one neuron for each case in the training data set. The neuron stores the values of the predictor variables for the case along with the target value. When presented with the *x* vector of input values from the input layer, a hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function using the sigma value(s).

Decision layer: The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron

that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent.

Probabilistic Neural Network (PNN)

A Probabilistic Neural Network **(PNN)** is a feed forward neural network used in classification problems. When an input is present, the first layer computes the distance from the input vector to the training input vectors. This produces a vector where its elements indicate how close the input is to the training input. The second layer sums the contribution for each class of inputs and produces its net output as a vector of probabilities. Finally, a complete transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 (positive identification) for that class and a 0 (negative identification) for non-targeted classes. Figure 3.8 shows the Structure of the PNN.

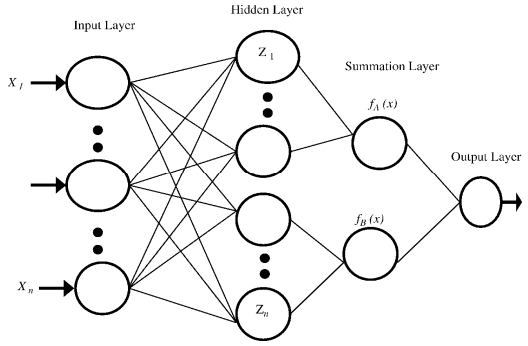

Figure 3.8: Probabilistic Neural Networks

Input layer

Each neuron in the input layer represents a predictor variable. In categorical variables, *N-1* neurons are used when there are *N* numbers of categories. It standardizes the range of the values by subtracting the median and dividing by the interquartile range. Then the input neurons feed the values to each of the neurons in the hidden layer.

Pattern layer

This layer contains one neuron for each case in the training data set. It stores the values of the predictor variables for the case along with the target value. A hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function using the sigma values.

Summation layer

For PNN networks there is one pattern neuron for each category of the target variable. The actual target category of each training case is stored with each hidden neuron; the weighted

value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent.

Output layer

The output layer compares the weighted votes for each target category accumulated in the pattern layer and uses the largest vote to predict the target category.

4. EXPERIMENTAL RESULTS

4.1 Feature Extraction

The speech samples were obtained from audio recordings of people with normal voice and pathological voice. 20 samples are taken for analysis (10 for training and 10 for testing). MFCC features are extracted from the samples after applying the techniques such as pre-emphasis, framing, windowing, FFT, Mel Filter banks, log, and DCT, . The speech signals and the extracted features are shown in Figure 4.1 - Figure 4.4.

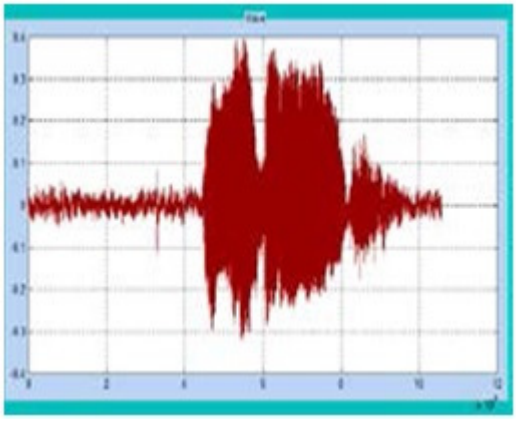

Figure 4.1: Normal voice

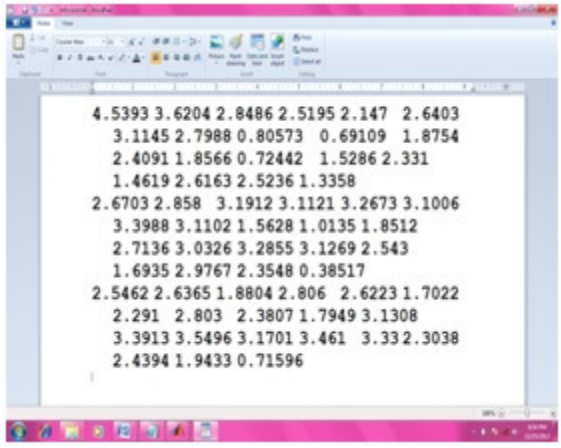

Figure 4.2 Feature Extraction for normal Voice

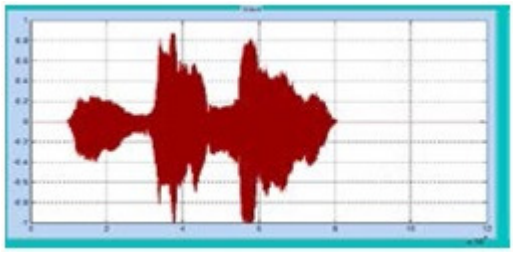

Figure 4.3: Pathological voice

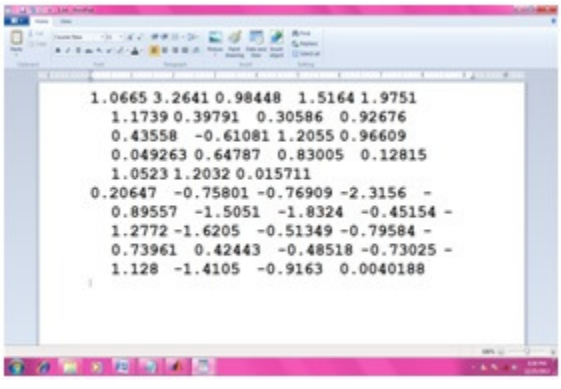

Figure 4.4: Feature extractions for pathological voice

4.2 Classification

The MFCC feature vectors derived from the speech samples (both normal and pathological voices) are used as input for the Neural Network for training and testing. The Neural Network Model is shown in Figure 4.6.

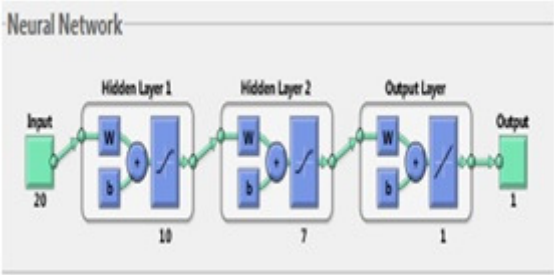

Figure 4.6: Multilayer Perceptron Neural Network

There are two hidden layers in the network. Number of outputs for the first hidden layer is 10 and that of the second hidden layer is 7. The output produces either 1 (for pathological voice) or 0 (for normal voice).

Performance of MLPNN, GRNN and PNN

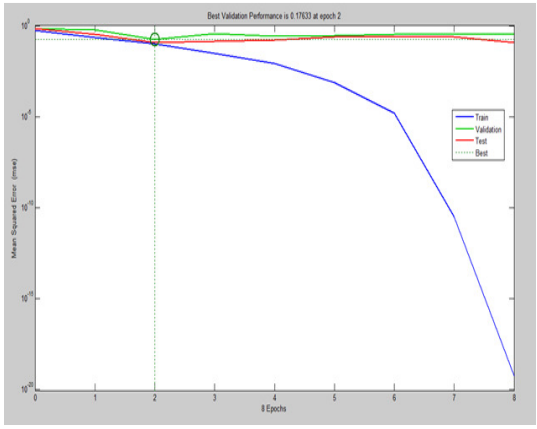

Figure 4.7: Mean Square Error Vs Number of Epochs for training, validation, testing and best fit for Multilayer Perceptron Neural Network (MLPNN).

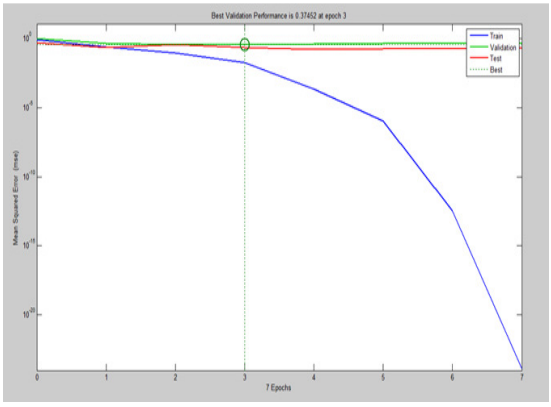

Figure 4.8: Mean Square Error Vs Number of Epochs for training, validation, testing and best fit for General Regression Neural Network (GRNN).

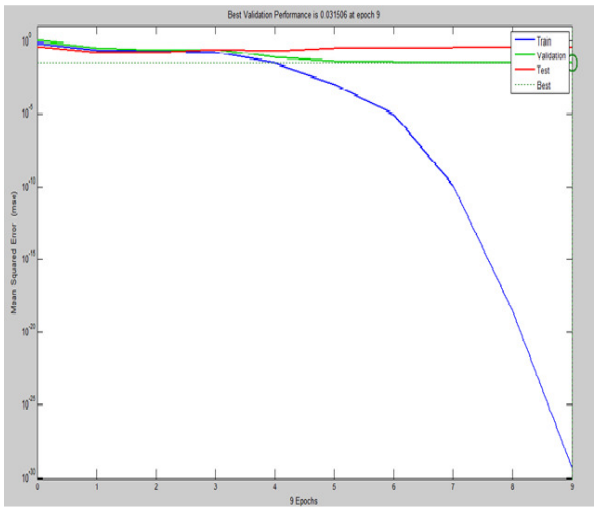

Figure 4.9: Mean Square Error Vs Number of Epochs for training, validation, testing and best fit for Probabilistic Neural Network (PNN).

The Mean Square Error Vs Number of Epochs for training, validation, testing and best fit for MLPNN, GRNN and PNN are shown in Figure 4.7, Figure 4.8 and Figure 4.9. It is observed that GRNN and PNN behave in similar manner. The MLPNN shows a better result than GRNN and PNN.

The classification accuracy of MLPNN is nearly 100% in identifying pathological voice compared to GRNN and PNN which are lesser. This is shown in Figure 4.10.

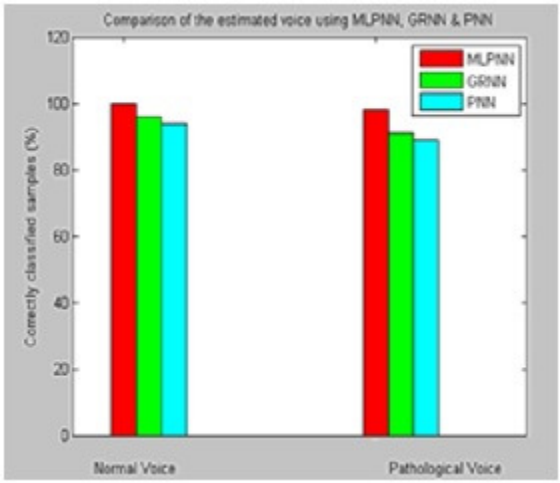

Figure 4.10: Comparsion of MLPNN,GRNN,PNN

5. CONCLUSION

This paper aims at developing an automated pathological voice recognition system. This is an isolated speech recognition system. Mel-frequency cepstral coefficients are extracted from the audio recordings. These MFCC features are used as input for various Neural Network

models for classification. The behavior of Multilayer Perceptron Neural Network (MLPNN), Generalized Regression Neural Network (GRNN) and Probabilistic Neural Network (PNN) in the classification of pathological voices from normal voices has been analyzed. The Probabilistic Neural Network (PNN) behaves in a similar way to Generalized Regression Neural Network (GRNN).It is found that Multilayer Perceptron Neural Network (MLPNN) with MFCC features performs better than PNN and GRNN in the classification of pathological voices. Further numbers of features are selected for training the sample values are justified by the expected result of 100% accuracy in the classification of voices in Multilayer Perceptron Neural Network (MLPNN).

REFERENCES

[1] V. Srinivasan , V. Ramalingam, V. Sellam, "Classification of Normal and Pathological Voice using GA and SVM", International Journal of Computer Applications,Volume 60 – No.3, December 2012.

[2] D. Pravena, S. Dhivya, A. Durga Devi, "Pathological Voice Recognition for Vocal Fold Disease", International Journal of Computer Applications, Vol-47, No.13, 2012.

[3] C.M. Vikram and K. Umarani, "Pathological Voice Analysis To Detect Neurological Disorders Using MFCC & SVM", International Journal of Advanced Electrical and Electronics Engineering, Volume-2, Issue-4, 2013.

[4] Vahid Majidnezhad, Igor Kheidorov, "A HMM-Based Method for Vocal Fold Pathology Diagnosis", International Journal of Computer Science, Vol. 9, Issue 6, No 2, November 2012.

[5] Cheolwoo Jo, " Source Analysis of Pathological Voice", Proceedings of the international Multi conference of Engineers and Computer Scientists, Vol II, IMECS March 2010.

[6] Alireza A, Dibazar, Theodore W. Berger, and Shrikanth S. Narayanan,"Pathological Voice Assessment", IEEE EMBS 2006,New York.

[7] Mandar Gilke, Pramod Kachare, Rohit Kothalikar,Varun Pius Rodrigues, Madhavi Pednekar, "MFCC-Based Vocal Emotion Recognition Using ANN", International Conference on Electronics Engineering and Informatics (ICEEI), vol 49, 2012, Singapore.

[8] Vahid Majidnezhad, Igor Kheidorov, "An ANN-Based Method for Detecting Vocal Fold Pathology", International Journal of Computer Applications, (0975 – 8887) Volume 62– No.7, January 2013.