Processing Mel Speech Power Spectrum for Speech Restoration

Basil Sh. Mahmood[1], Nada N. Ibrahim[2]

[1,2]Computer Engineering Department, University of Mosul, Mosul, IRAQ

ABSTRACT

Speech restoration is one of the most important processes application in speech processing domain. The speech signal may be distorted for many reasons, three stages of processing operations were used as a new way to handle the distorted speech signal (distortion may be taken place due to shrinking in time or in wave length). These stages are : features extraction using Mel frequency cepstral coefficients (MFCCs), processing the extracted features using back propagation neural network and finally restoringthe speech wave forms from the reconstructed MFCCout coming from the neural network. The results of training the neural network as a new way reveals that this approach is a very good way according to the mean square error and the average peak signal to noise ratio for this system where 0.054 is the value for MSE and 26.12 is the value for PSNR.

Keywords: Mel frequency cepstral coefficients, neural network, speech distortion, speech restoration, time shrinking.

1. INTRODUCTION

Nowadays speech is the best way that is used in the world for communication among members of human beings [1]. Usually, the speaker generates ideas in his mind and then translates these ideas into words and sentences clips by using certain linguistic rules.

Speech processing means studying and analysing speech signals, speech processing fieldattract an attention of many researchers for conducting researches in this area since many decades ago [2]. These efforts leads to get big progresses and good results are earned in that field [3], especially after the development of speech processing techniques such as Neural Network, Fuzzy Logic, Hidden Markov Model, and Vector Quantization and these techniques are used in various speech processing applications for example speech recognition, speaker recognition, speech filtering, speech to text, speech restoration, ... etc.

Generally, the term of restoration is aterm intended to remove defects, impurities, and unwanted noise from speech or sound signal [4]. The distortion of speech occurs for many reasons and there are different techniques for retrieving the original speech from the distorted speech [5]. Speech restoration is constituted a major challenge for the researchers because of losing some parts of the speech information due to the distortion.

This paper aimed to fined a new approach for speech restoration, which starts with recording speech signal using digital recording program followed by using Mel Frequency Cepstral Coefficient (MFCC) technique to extract the features as the first step in any system for speech processing, that means the identification of using the speech signal components to distinguish the linguistic component and neglecting all other things that contain some information like background noise. Finally these features are fed to a neural network to restore the distorted speech signal which happens due to the time shrinking of the speech signal.

2. PREVIOUS WORKS

Arundhoti et al.,(2011) [6] found that MFCC is the best way for speaker identification during his study by using 8000 Hz to sample the speech signal followed by pitch sound and MFCC to identify the speaker identity.

Based on using self organization map neural network (SOM) Teswarlu et al.,(2011) [7] submit a new method for speech recognition by using four different ways to extract features: MFCC, linear predictive cepstral coefficients (LPCC), pitch, and intensity. Good results was obtained in both training and testing phases for all four types of features and the best result of median-SOM performance was recorded (98.17%) when intensity was used for extracting features.

Comparative study was carried out by Cutajar et al.,(2013) [8] to compare between the common techniques that used in automatic speech recognition system (ASRS), whichsplits the recognition system into to three stages: feature extracting stage, recognition stage, and system modelling stage. A good and acceptable results were found in all tested techniques for feature extracting with the exceed of MFCC and mostly used in this area. Although there are some techniques gave good results like neural network and hidden Markov model regarding to recognition stage but they are still far away from the ideal precision and human recognition ability, while the integration between more than one technique gave best results in their experiments. Back-propagation was used for training the neural network by Joshi et al,. (2014) [9] for speech recognition.

The speech signal was recorded upon Audacity program in 8000 Hz sampling frequency in wave form. Followed by extracting the features via MFCC. The obtained results showed that neural network is efficient way to recognition the speech signal which reported 80% accuracy during testing phase. During the design system for isolated speech recognition for Indian assamese language through using neural network, Medhi et al,.(2015) [10] which was consisting of three phases(training, testing, and recognition) the data base collected from 20 persons (50% male and 50% female) through repeating the most famous 100 word in assamese language for 20 times for each. Feature extracted for obtained data using Zero crossing rate and MFCC, followed by introducing it to the neural network which consist of three layers: input, output, and one hidden layer tuned for 35, 45, 64 and 80 neurons which represents the best one. Both speaker dependant and independent gave a good result in speech recognition.

According to previous studies, it is obvious that neural network is an efficient way to handle different speech processing applications task because it is the closest way to simulate human mind and to overcome the difficulties in other ways.

3. SPEECH RESTORATION USING NEURAL NETWORK

Speech restoration is one of the most important application of speech processing, the purpose of restoration process is to remove defects, distortion, and unwanted noise from the speech signal. The speech signal is distorted in many ways for different reasons such as distort speech signals by back-ground noise, or error occurred in transforming channel during transforming signal ...etc, which lead to loss of some important parts from speech signals.

This study focus on restoring the speech signal that was distorted as a result of time shrinking. Speech restoration process can be divided into three stages as shown in Fig. 1. The first stage represents extracting the important features from speech signals through MFCC. The second stage is managing the process of extracting features in order to edit and recover the correct MFCC values via neural network. The final stage is represented by reconstructing the speech waveforms from the modified MFCC in order to re-listen to the speech signal and make sure that the signal is restored in a correct way, upon inverse MFCC as will be explained in details in the followingsections.

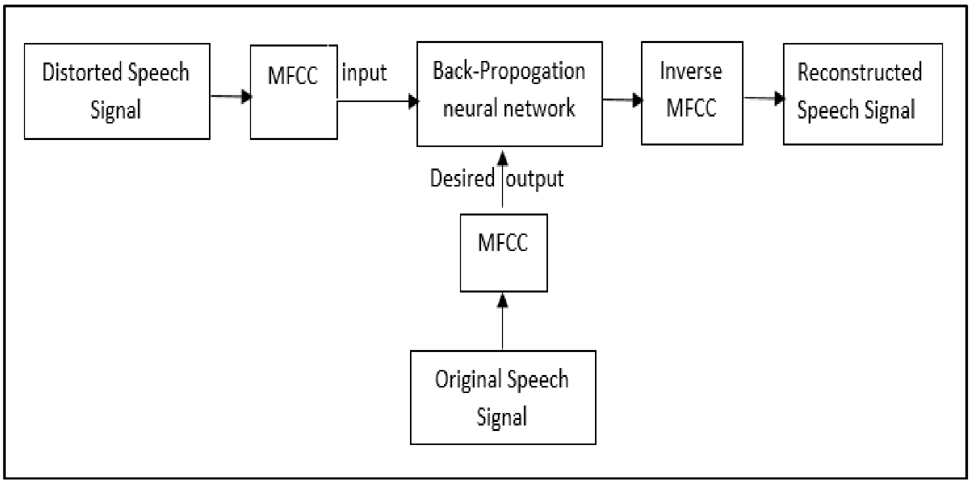

Figure 1: Main stages of speech restoration system.

3-1: Feature extraction for speech signals:

Feature extraction is the first step in any speech processing system which refers to identifying the useful linguistic components of speech signal and neglecting the other components like noise, MFCC is used for this perpuse as follows:

3-1-1: Mel Frequency Cepstral Coefficient (MFCC)

MFCC is the famous and the most widely used technique in speech processing area [11], because of its ability to mimic the human response to different frequencies. Since the human ear can not recognize between two little different frequencies. MFCC is calculated by executing the following steps, as shown in Fig. 2.

- Framing: This step is performed because the speech signal is changed constantly and very quickly, therefore, it is difficult to handle it, so we assume that it is statically fixed for a very short time period. Therefore, the signal is divided to small overlapped segments called frames, the frame length is between 20-40 Milliseconds [11]. In this research the frame length is 32 Milliseconds with overlapping ratio about 50%.
- Windowing: In this step the framed signal is convoluted with window function as in equation (1) to reduce the differences at the beginning and at the end of each frame, and also to maintain the continuity of the first and last point in each frame.

$$y(n) = x(n) * w(n)(1)$$

Were : x(n): is the speech signal samples, w(n): is the window function.

- Fast Fourier Transform (FFT): in this step the signal is converted from time to frequency domain as in equation (2).

$$Y(n) = FFT[x(n) * w(n)] = X(n).W(n) \qquad (2)$$

Were : x(n): is the speech signal samples, w(n): is the window function, * denote convolution operation.

- Power Spectrum: The human ear cochlea is vibrating in different area depending on the heard sound frequency, depending on cochlea location the nerves inform the brain about existing sound frequency. Power spectrum calculation process is similar to cochlea task. In the other word power spectrum tells us the existing frequency in the frame.
- Mel filter: The result of the previous step is multiplied by Mel filters to compute the power spectrum for each filter. Mel scale is linear for frequencies bellow 1000Hz, and it is logarithmic for frequencies above 1000 Hz. So, Mel scale tells us appropriate width for each filter to simulate the human ear ability for frequency recognition. The basic equation to convert the frequency from hertz to mel scale as in [13].

$$F_{mel} = 2595 \log_{10}(1 + F_{HZ}/700) \qquad (3)$$

F_{mel} : is the frequency in mel scale, F_{HZ} : is the frequency in hertz scale.

- Logarithm: Mel-scale filter bank output is passed to logarithmic operation. This is done for normalization purpose [19].
- Discrete Cosine Transform (DCT): Discrete cosine transform is used to convert the log. values of mel cepstral magnitude into time domain to produce the mel frequency cepstral coefficients.

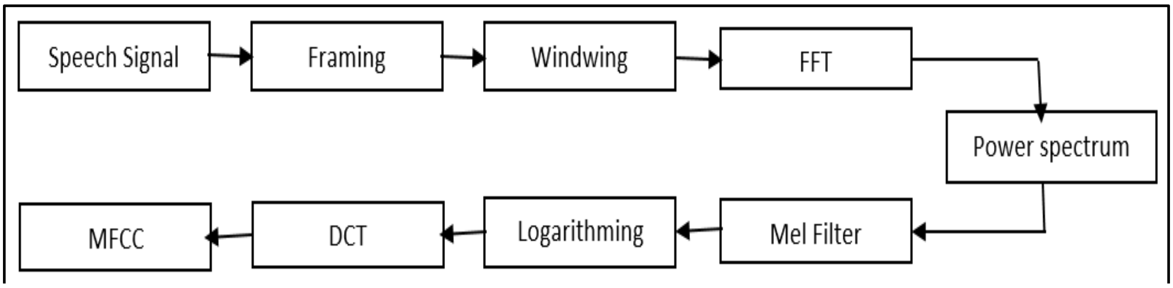

Figure 2: Block digram for MFCC calculation steps.

3-2: Neural network for speech restoration

Neural network is one of the most important field of artificial intelligence, it is an associative system for processing data in a simulation way to the natural neural network operation [14]. The neural network consists of a set of neurons worke in parallel, each neuron represents a small processing unit [15]. The function of neural network here is to modify the MFCC values of the distorted speech signal to restore the speech signal followed by processed the data into two phases: training with back-propagation algorithm and testing phase [16].

3-2-1: Training the neural network

There is four main steps to train and test neural network by using back propagation algorithm [17], these steps include: gathering the training data, create the network object, train the network, and simulate the network response to new inputs for testing the network. Fifty speech signals have been recorded by using Audacity software with 8000 Hz sampling frequency for assembling data, Then the speech signals are distorted by shrinking the signal period time to the half of its value, and MFCCs is calculated for both recorded and distorted words to create the dataset that are introduced to the neural network. Because of the large size of the dataset, it has been divided into seven groups. So, seven neural network are created, also the activation function for each layer and the training algorithm that used to train the network are determined by the following matlab instruction:

```
net=newff (minmax(Indata),[50 110],{'logsig','purelin'},'trainlm');
```

A very good results was gained during training the networks in both training and testing phase. This is followed by using IMFCC to get the speech waveform from the restored MFCCs valuesto re-listen to the restored speech words.

3-3: Inverse Melfrequency cepstral coefficients (IMFCC)

As mentioned above, MFCCis one of the most important technique in signal processing domain. On the other hand, the big challenge is in inverse process i.e. reconstruct the speech waveforms from MFCCs. Generally, there are two main complicated troubles in reconstructing the speech waveforms from MFCCs. First problem, is the sparse of power spectrum that means some information are lost through calculation. Second, Mel scale is logarithmic for frequencies above 1000 Hz, which means that it is impossible to recalculate the inverse of Mel matrix. Several ways have been proposed to overcome these two problems, but the best result was obtained uponusing L2 norm criteria [2]. Fig. 3.is ablock diagram illuminating the main steps to retrieve the speech wave forms from MFCCs. From the diagram below, it is obvious that the first and second steps can be achieved easily as in equation (5).

$$\mathrm{Z} = \exp(\mathrm{IDCT}\{\mathrm{MFCC}\})(5)$$

Iteratively Reweighted L2 Minimization (IRLM) algorithm is used for exceeding the mentioned problems to calculate the power spectrum. Also, more details are explained in [18]. Then inverse FFT is used to calculate the signal phase. Finally, the speech waveforms was reconstructed by overlap add procedure.

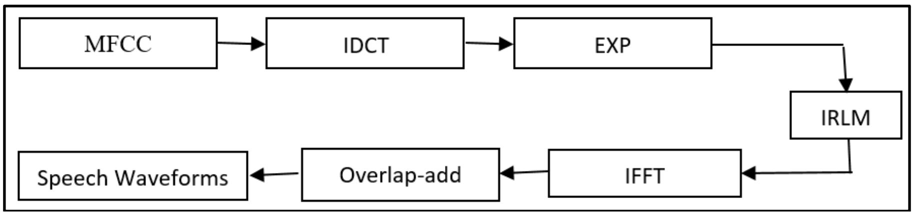

Figure 3: Block diagram shows IMFCC steps.

4. RESULTS AND DISCUSSION

MATLAB R2015a was used to perform the suggested technique. The original and the distorted speech signal are plotted as shown in the Fig. 4.The figure declares that the distorted speech signal is shrinked to its half period time (just like when the speed of a play back recorder increases to its double normal speed).

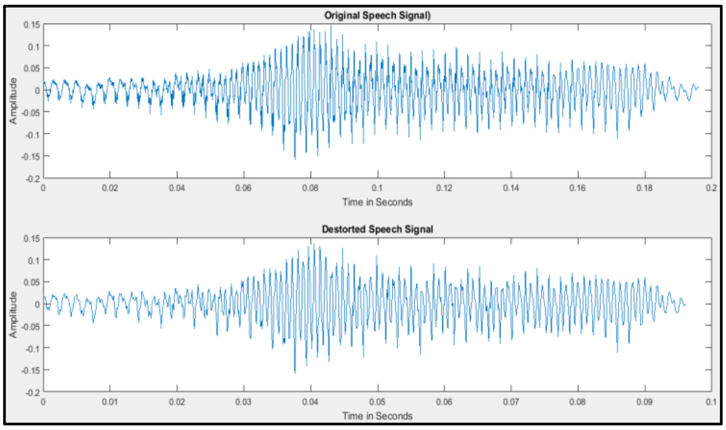

Figure 4: Original and distorted speech signal.

Seventy coefficients of MFCC were computed for each frame signal for both original (consisting of 11 frames) and the distorted speech signal(consisting of 5 frames)as shown in Fig. 5.

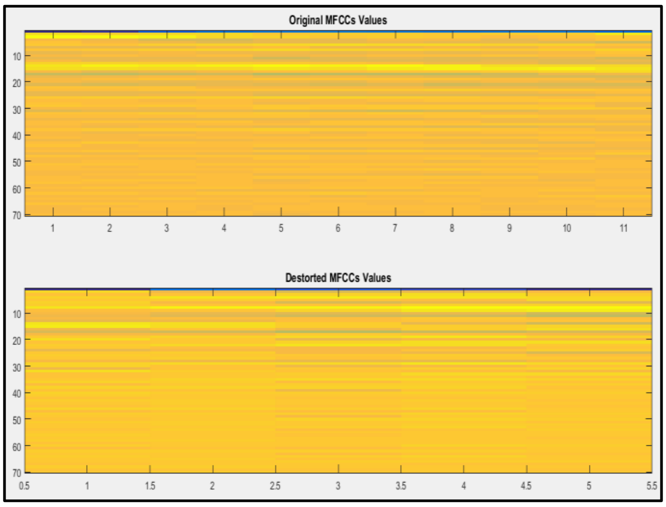

Figure 5: Mel frequency spectrum for original and distorted speech signal.

Seven neural networks were trained by feeding 35 recorded speech signals, then they were tested by other 15 speech signals. Training stopping criterion is subjected to10^{-15}Which is the limiting training parameter goal for stopping the training operation. Table (1) shows the epoch number and training time that the neural network reaches its best performance.

High stability and a very good performance gained in a suitable time in training each network. Fig. 6.shows the performance of one of the neural networks.

Table 1: Network performance .

Network number	Best performance (MSE)	Iteration Number	Training time
Network 1	$2.04 * e^{-23}$	12	0:05:11
Network 2	$2.65 * e^{-18}$	12	0:04:12
Network 3	$4.05 * e^{-16}$	10	0:03:59
Network 4	$4.96 * e^{-26}$	11	0:04:22
Network 5	$6.58 * e^{-17}$	11	0:04:33
Network 6	$2.04 * e^{-20}$	11	0:04:38
Network 7	$4.26 * e^{-15}$	10	0:03:51

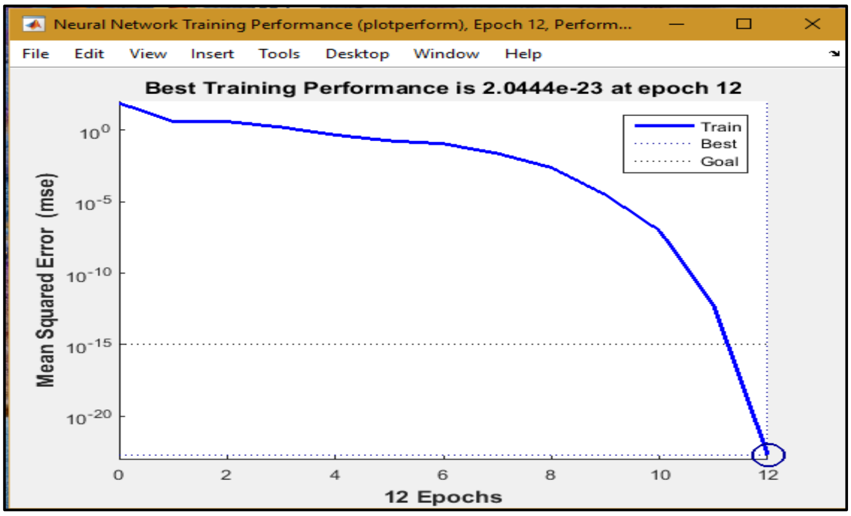
Neural Network Training Performance (plotperform), Epoch 12, Perform...
File
Edit
View
Insert
Tools
Desktop
Window
Help
Best Training Performance is 2.0444e-23 at epoch 12
Train
Best
Goal
Mean Squared Error (mse)
10^0
10^-5
10^-10
10^-15
10^-20
0
2
4
6
8
10
12
12 Epochs

Figure 6: performance of the first neural network.

The peak signal to noise ratio (PSNR) and the mean square error (MSE) were calculated according to the recorded and the restored speech waveforms after retrieving the speech wave forms from the restored MFCCs values,they are plotted as shown in the Fig.7, 8. respectively. Meanwhile the data of table (2) clarifies the mean of MSE and the PSNR values for the 35 trained, 15 tested and the whole 50 speech signals.

According to the results of table (2) , PSNR for thetrained samples records higher values than that of the tested samples , this difference decreases as the network trained on extra samples..

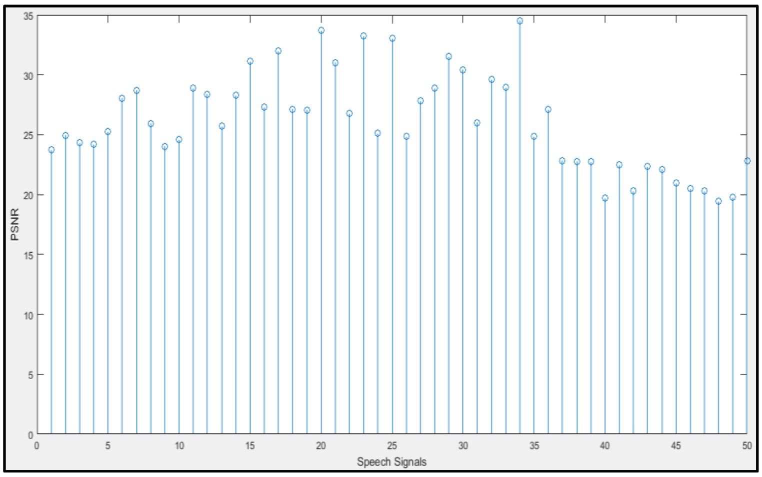

Figure 7: Pick Signal to noise ratio to restored speech signals.

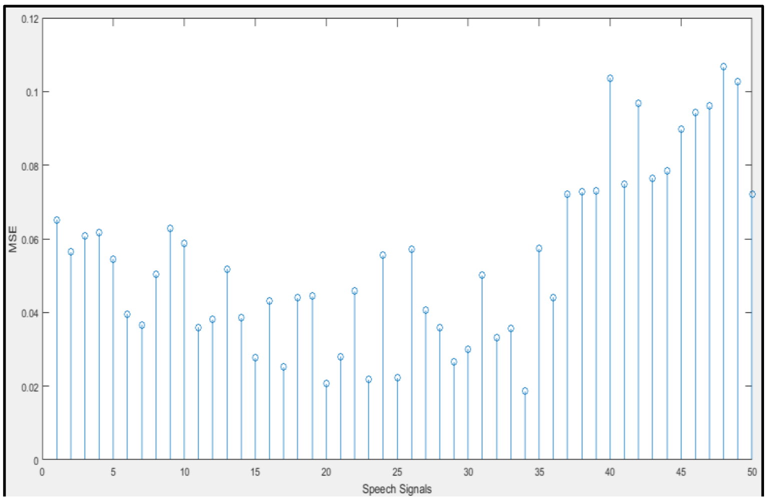

Figure 8: Mean square error to restored speech signals.

Table 2: average values for PSNR and MSE.

S. No.	**Sets**	**Mean PSNR**	**Mean MSE**
1	Training set	28.00213	0.042147
2	Testing set	21.746292	0.083595
3	Whole set	26.12557	0.0546

CONCLUSION

Speech restoration is one of the most important processing application, which enables the listener to understand the distorted speech. It supports many applications such as restoring the signal from damaged recorders and restoring the speech signal due to transmitting the speech signal between different physical media. A new approachfor restoring the distorted speech signal subjected to time shrinking was achieved . Mel frequency cepstral coefficient techniqueoffer valuable results overother techniques according to its abilityto simulate the human response for different frequencies.
For restoration process back-propagation algorithm is used for training the neural network. The results show that the neural network is an efficient way to restore the distorted speech signal.

REFERENCES

[1]. Meena, K.; Subramaniam, K.and Gomathy, M. (2013).‘Gender classification in speech recognition using fuzzy logic and neural network’. International Arab Journal of Information Technology.Vol.10, no.5, pp. 477-485.

[2]. Hossain, A.; Rahman, M.; Prodhan, U. K. and Khan, F.(2013). ‘Implementation Of Back-Propagation Neural Network For Isolated Bangla Speech Recognition’. International Journal of Information Sciences and Techniques (IJIST). Vol.3, no.4, pp. 1-9.

[3]. Paliwal, K.K. (1990). ‘Speech processing techniques’. Advance in speech, hearing and language processing. Vol.1, pp. 1-78.

[4]. Godsill, S. J. and Rayner, P. J. W. (1995). ‘A Bayesian Approach to the Restoration of Degraded Audio Signals’. IEEE Transactions on Speech and Audio Processing. Vol. 3, no. 4, pp. 267-278.

[5]. Dahimene, A.; Noureddine, M. and Azrar, A. (2008). ‘A simple algorithm for the restoration of clipped speech signal’.Informatica (Ljubljana). Vol. 32, no.2, pp.183-188.

[6]. Mehendale, A.; and Dixit, M. (2011). ‘Speaker Identification’. Signal & Image Processing: An International Journal (SIPIJ). Vol. 2, no. 2, pp. 62-69.

[7]. Venkateswarlu, R.L.K.; Kumari, R. V. and Nagayya A.K.V. (2011). ‘Novelapproach for speech recognition by using self - organized maps’.International Journal of Computer Science & Information Technology. Vol. 3, no. 4, pp.199-210.

[8]. Cutajar, M.; Gatt, E.; Grech, I.; Casha, O. and Micallef, J. (2013). ‘Comparative study of automatic speech recognition techniques’. IET Signal Processing. Vol.7, no.1, pp.25 – 46.

[9]. Joshi, S.C. and Dr. Cheeran A.N. (2014). ‘MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition’. International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering. Vol. 3, No. 7, pp. 10498-10504.

[10]. Medhi, B. and Prof. Talukdar, P. H. (2015). ‘Isolated Assamese Speech Recognition Using Artificial Neural Network’. International Symposiwn on Advanced Computing and Communication (ISACC).ISBN:978-1-4673-6707-3, pp. 141 - 148.

[11]. Rachna; Singh, D. and Vikas. (2014). ‘Feature extraction using using mel-frequency cepstral cofficients’. International Journal of Research in Engineering and Technology (IJRET). Vol.3, no.6, pp. 273-276.

[12]. Gupta, S.; Jaafar, J.; Ahmad, W. F.and Bansal, A. (2013). ‘Feature extraction using MFCC’. Signal & Image Processing : An International Journal (SIPIJ). Vol.4, no.4, pp. 101-108.

[13]. Tiwari,V. (2010). ‘MFCC and its applications in speaker recognition’. International Journal on Emerging Technologies. Vol.1, no.1, pp. 19-22.

[14]. Alsaif, Kh. I. and Alnuaimy, M. Kh. (2010). ‘Neural NetworkBased for Gender Recognition’Alrafidain Journal of Computer Science and Mathematics, proceeding of the 3rd Scientific Conference on Information Technology. pp. 29-39.

[15]. Albakry, A. M. and Ismael, G. (2010). ‘Using Neural Network to recognize geometrical shapes’. Journal of Babylon University/ Pure and Applied Science. Vol.18, no.5, pp. 1889-1898.

[16]. Bose, N. K. and Liang P. (1996). ‘Neural network fundamentals with graphs, algorithms, and applications’. McGraw-Hill.

[17]. Beale, M. H.; Hagan, M. T. and Demuth, H. B. ‘Neural Network Toolbox User's Guide’. COPYRIGHT 1992–2016 by The MathWorks, Inc. https://www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf.

[18]. Min, G.; Zhang, X.; Tan,W.; Zou,Xia. and Yang J. ‘A simple speech recovery algorithm using iteratively reweighted l2 minimization for DSR’. http://www.mathworks.com/matlabcentral/fileexchange/53186-invmfccs.

[19]. Tychtl, Z. and Psutka, J. ‘Speech Production Based on the Mel-Frequency Cepstral Coefficients’. http://mirlab. org/conference _ papers/International_Conference/Eurospeech%201999/PAPERS/S11P2/P024.PDF.