Speaker and Speech Recognition using Deep Neural Network

Article · June 2018

DOI: 10.23956/ijermt.v6i8.126

3 authors:

Gurpreet Kaur
Rajendra Institute of Technology and Sciences
30 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Mohit Srivastava
Indian Institute of Technology Roorkee
20 PUBLICATIONS **62** CITATIONS
SEE PROFILE

Amod Kumar
National Institute of Technical Teachers Training and Research
63 PUBLICATIONS **458** CITATIONS
SEE PROFILE

International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-8)

Research Article | **August 2017**

Speaker and Speech Recognition using Deep Neural Network

Gurpreet Kaur[1,2], Mohit Srivastava[3], Amod Kumar[4]

[1] Research Scholar, I.K Gujral Punjab Technical University, Kapurthala, India

[2] Assistant Professor, University Institute of Engineering & Technology, Panjab University, Chandigarh, India

[3] Professor, Chandigarh Engineering College, Landran, Mohali, India

[4] Scientist, Central Scientific Instruments Organisation, Chandigarh, India

Abstract—

In command and control applications, feature extraction process is very important for good accuracy and less learning time. In order to deal with these metrics, we have proposed an automated combined speaker and speech recognition technique. In this paper five isolated words are recorded with four speakers, two males and two females. We have used the Mel Frequency Cepstral Coefficient (MFCC) feature extraction method with Genetic Algorithm to optimize the extracted features and generate an appropriate feature set. In first phase, feature extraction using MFCC is executed following the feature optimization using Genetic Algorithm and in last & third phase, training is conducted using the Deep Neural Network. In the end, evaluation and validation of the proposed work model is done by setting real environment. To check the efficiency of the proposed work, we have calculated the parameters like accuracy, precision rate, recall rate, sensitivity and specificity..

Keywords— Speech Recognition, Speaker Recognition, Mel Frequency Cepstral Coefficient, Genetic Algorithm, Deep Neural Network

I. INTRODUCTION

Study of speech and its processing methods is known as speech processing. Speech signal is having linguistic as well as speaker information. Although speech and speaker recognition are two different fields but these two fields are overlapped [1-3]. The research is going in this field since last 40 years to address the issues of recognition accuracy, degradation in accuracy with noise, channel variability, vocabulary size, poor pronunciation etc.[4-6]. Speech signal consists of different attributes like loudness, pitch, fundamental frequency, spectral envelope, formants etc. These attributes help to find out the speaker as well as speech features. There are different existing methods to extract these features from the speech signal like linear Predictive Coding Coefficient (LPCC), Perceptual linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC) and Relative Spectra Filtering (RASTA)[7-12]. These methods can be implemented in speech recognition as well as in speaker recognition. Out of these, MFCC is the best known and most prominent method for feature extraction.

Speaker and speech recognition systems contain three main modules viz. feature extraction, feature optimization and feature matching[13-15]. The work reported in this paper is for combined speech and speaker recognition using MFCC feature extraction method with genetic algorithm as feature optimization and deep neural network for feature matching[16-18].

..

II. PROCESS OF SPEECH RECOGNITION

Figure 1 shows the process undertaken in the proposed methodology

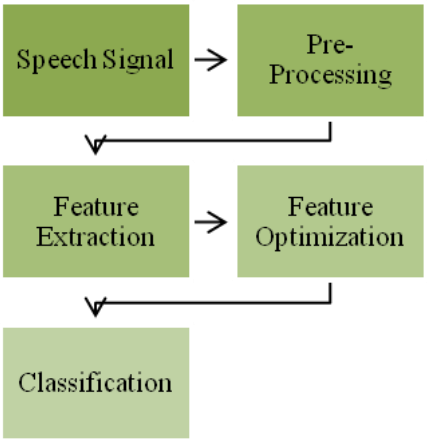

Fig.1: Methodology of combined speaker and speech recognition

The speech database is recorded in sound recorder with the help of headphone at 16000 Hz frequency at room environment in mono format. In our dataset, two hundred words are recorded by four speakers of age 27-34, two females (F1,F2) and two males (M1,M2). The recorded words are forward, backward, left, right and stop. After the acquisition of speech signal, preprocessing of speech signal is done which includes silence removal of the speech signal, pre-emphasis, framing and windowing. MFCC technique is used for feature extraction. The procedure involves Fast Fourier Transform (FFT) applied on the frame and then power spectrum converted into a Mel frequency spectrum. Logarithm of the spectrum and then its inverse Fourier transform gives the fundamental frequency and spectral envelope for each speaker with spoken words.

A Deep Neural Network (DNN) is trained by optimized features[19-23]. Genetic algorithm is used for determining the weights and biases of the neural network. The fitness function of genetic algorithm can be defined according to the requirement. In proposed work, fs is the current selected feature and ft is the threshold value of feature points. On the basis of given condition, we check the fit value which can exist in new feature set.

$$f(fit) = \begin{cases} 1, & fs < ft \\ 0, & fs \geq ft \end{cases} \qquad (1)$$

Where f(fit) is fit value according to the fitness function.

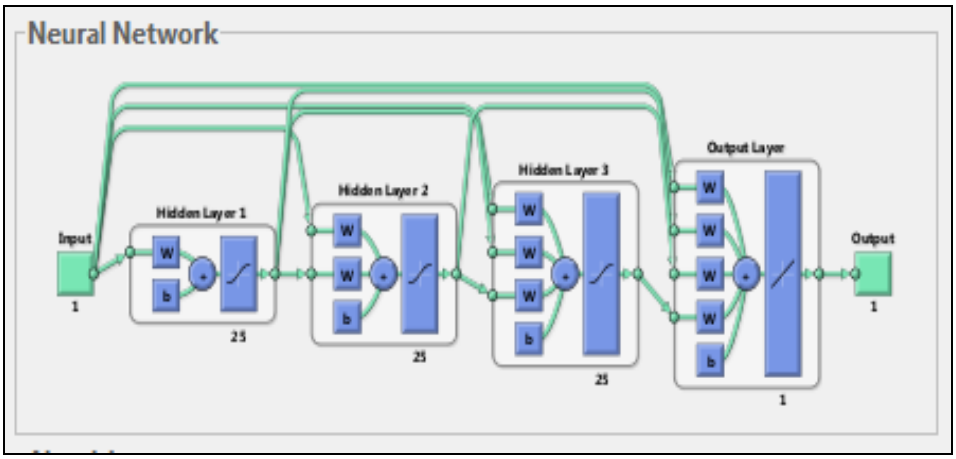

Fig.2: Deep Neural Network Training Panel

Neural Network Toolbox supports a variety of supervised and unsupervised architectures for the training in pattern recognition. Figure 2 shows the Deep Neural Network training panel. Through the toolbox's modular approach of building networks, the user can develop custom network architectures for the specific problem.

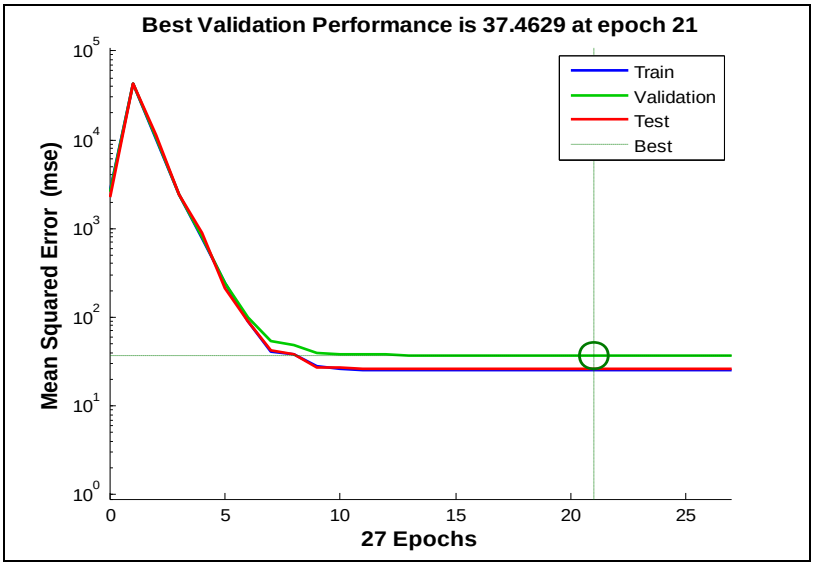

Fig.3: Performance of DNN

Figure 3 shows the graph of performance parameters of training function. The circle denotes the best performance in terms of least mean square error value of 37.4629 at iteration number 21. This figure does not point to any specific problems with the training. The validation and test curves are very similar. If the test curve increases drastically before the validation curve increases, then it is possible that over fitting has occurred.

The next step is validating the network for which a decay plot is generated to show the association between the outputs of the network and the targets. If the training is ideal, the network outputs and the targets would be accurately equal, but the connection is rarely perfect in practice.

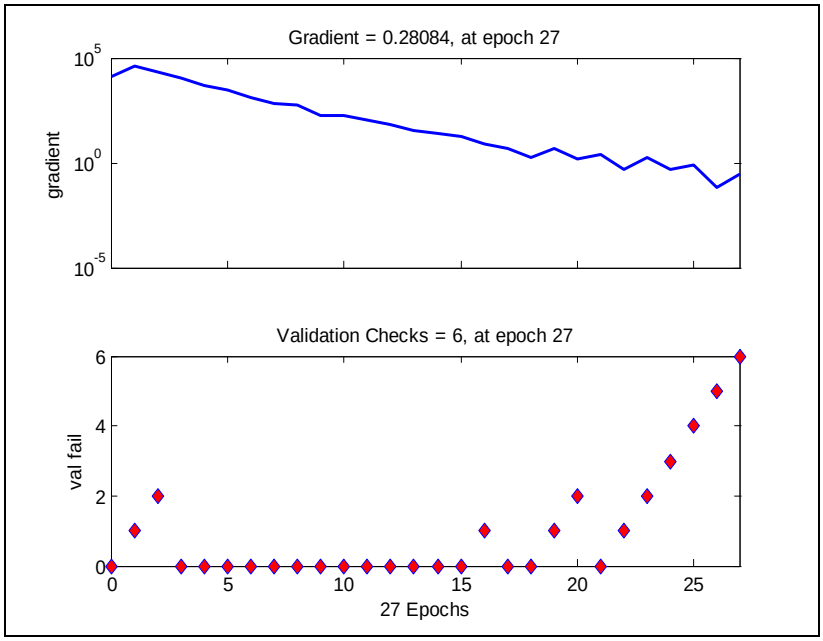

Fig.4: Parameters of DNN

Figure 4 shows the graph of different types of parameters which are generated during training of dataset.

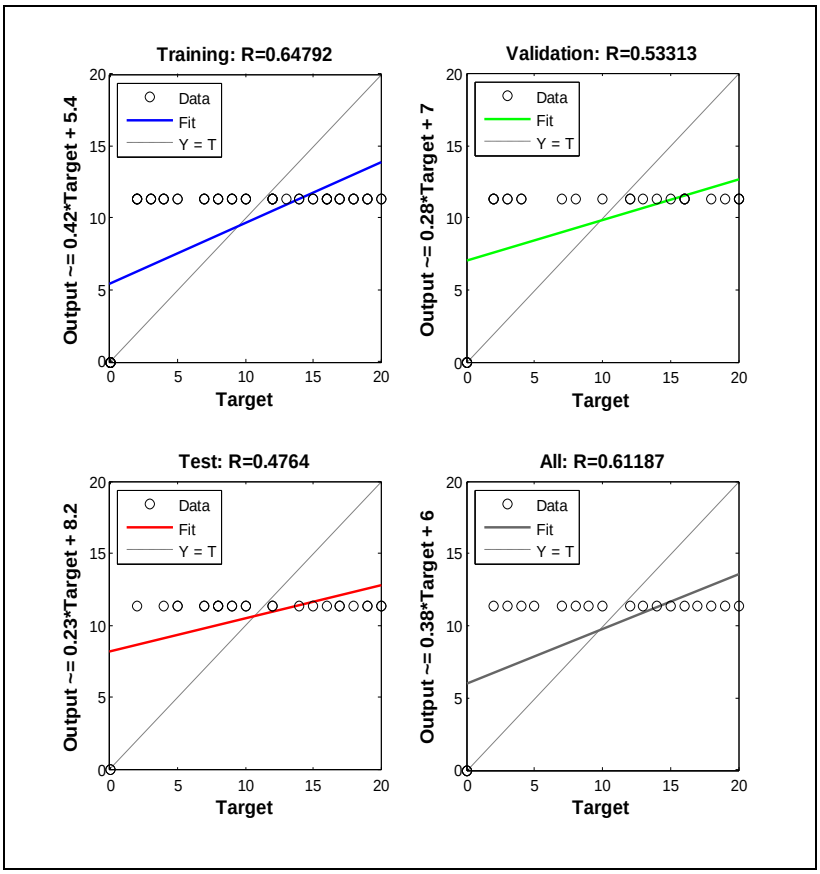

Fig.5: Dataset of DNN

Figure 5 shows the description of datasets which are used for the training purpose. There are total four graphs: first for training data, second for validation, third for test data which are automatically taken from the training dataset and last for ouput of training. In the graph, two lines are present - one is a solid line and second is dotted which represents the accuacy of training. The dashed line in every plot represents the perfect result – outputs = targets. The solid line shows the finest fit linear decay line between outputs and targets. The R value is a sign of the bond between the outputs and targets. R = 1 indicates that there is an exact direct relationship between outputs and targets. If R is close to zero, then there is no direct relationship between outputs and targets.

The process of the execution in form of algorithms is shown below with Genetic algorithm, with MFCC algorithm and Deep Neural Network.

III. RESULTS AND ANAYLSIS

In this work, the power of utilization of MFCC, Genetic Algorithm and deep neural network for combined speaker and speech recognition was established. It is quite difficult to recognize speech in the presence of noise. Proposed work is tested on various types of noise like White Gaussian Noise (WGN), Additive White Gaussian Noise (AWGN) etc. Due to noise, speech recognition becomes difficult, so, we use Genetic Algorithm for feature optimization. The experimental results have confirmed our expectations by giving high values in terms of measuring metrics like precision rate, recall rate, accuracy, sensitivity and specificity. This section also shows the results of comparison of performance parameters of two different methods – MFCC and MFCC with GA.

Graphic User Interface was also developed in this work. It is divided into two panels - training and testing panel. Figure 6 shows the uploaded speech sample and its feature set.

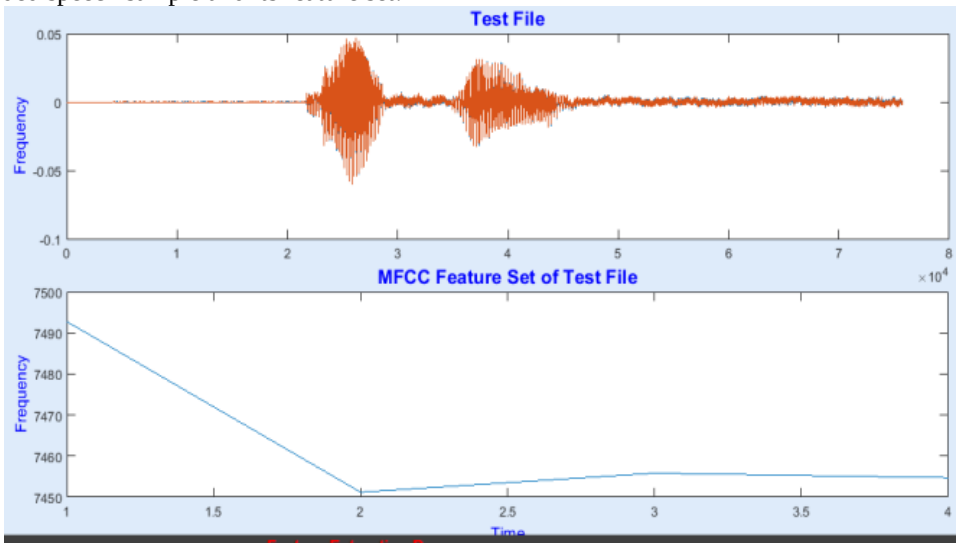

Fig. 6: Feature Extraction Process

Figure 7 shows the ROC (Receiver Operating Characteristics). It is a graphical method for comparing two empirical distributions. In this research work, true positive and false negative parameters have been taken.

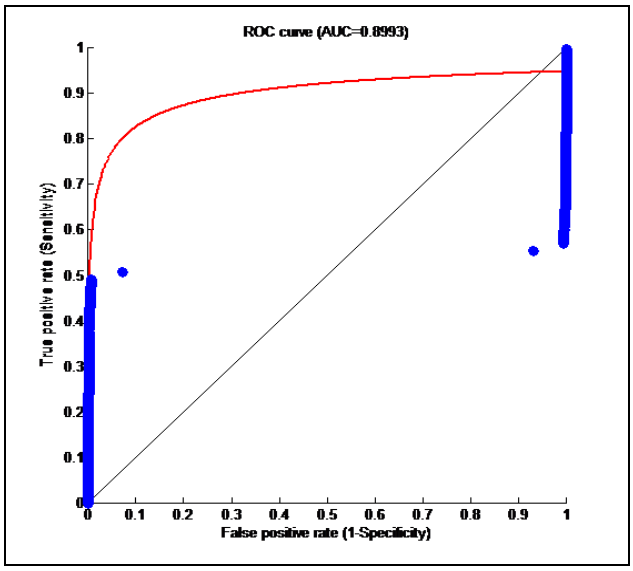

Fig. 7: ROC Curve for Proposed Work

Figure 8 and Table 1 represent the calculated parameters of proposed work for two speakers. N1 bars represent the graph for Speaker 1 and N2 for Speaker 2. In the figure, all calculated parameters are shown like precision rate, recall rate, sensitivity and specificity.

Table I Proposed Metrices Result Analysis

Parameters	N=1	N=2
True positive	0.939	0.955
False positive	0.446	0.448
True negative	0.448	0.446
False negative	0.446	0.445

Precision rate	0.669	0.678
Recall rate	0.911	0.921
Accuracy	97.15	97.12
Sensitivity	0.605	0.645
Specificity	0.502	0.503

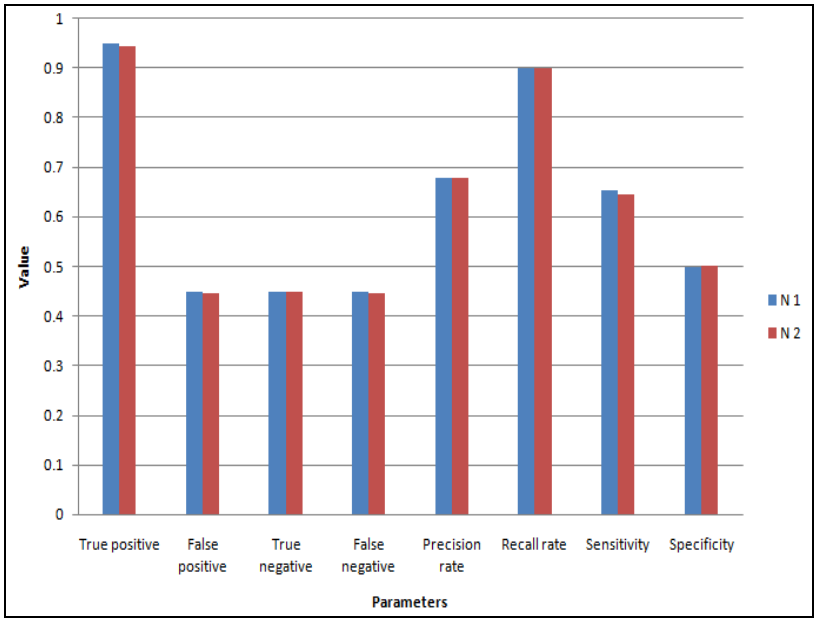

Fig.8: Evaluation results for two speakers

Table 2 Comparison of % Accuracy between Proposed Technique (MFCC with GA) and the Existing Technique (MFCC)

No. of Iterations	MFCC with GA	MFCC
1	97.44	89.11
2	98.28	88.69
3	97.11	91.43
4	97.15	90.37
5	96.46	89.42

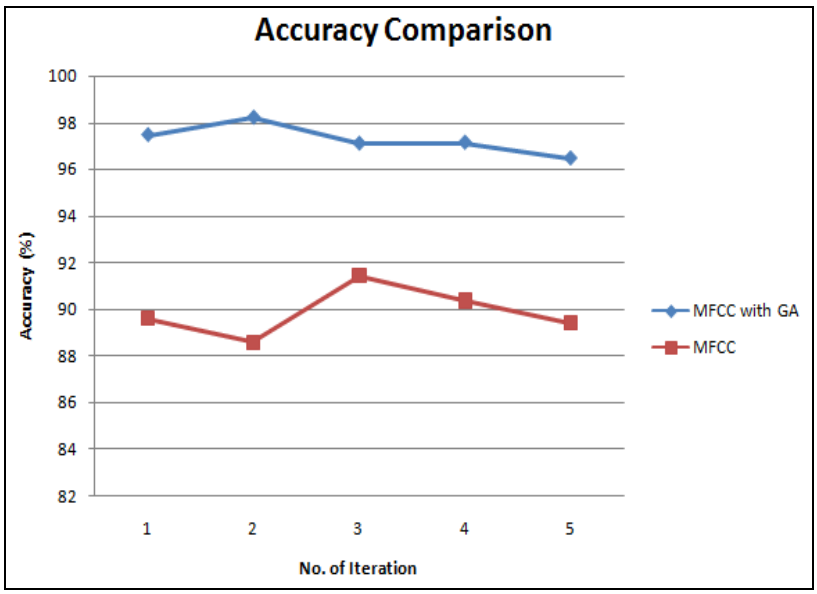

Fig.9: Accuracy Comparison

Figure 9 and Table 2 show the comparison for proposed and existing technique. The average accuracy for proposed technique was obtained as 97.18 whereas for existing technique, it was 89.08. So, it is concluded that the proposed system has more accuracy and reliability.

IV. CONCLUSIONS

We have shown that speech as well as speaker recognition system with MFCC feature extraction technique is helpful in achieving more accuracy for the proposed recognition system. To be specific, we found that optimization and feature extraction are very important as well as difficult steps in any pattern recognition system. In the proposed work, we extracted more useful feature set from signal using genetic algorithm and for the training and classification of data, we used deep neural network for suitable training in case of optimized as well as noisy signal. The experimental results indicate that proposed method has provided good results having values of true positive as 0.949, false positive 0.448, True negative 0.449, false negative 0.448, Precision rate 0.679, and Recall rate 0.901, Accuracy 97.05, Sensitivity 0.655 and Specificity 0.500. All these values are an improvement over the existing methods.

REFERENCES

[1] Herbig, T., Gerl, F., Minker, W., (2012). Self-learning speaker identification for enhanced speech recognition. Computer Speech Lang. 26, 210–227

[2] Huang, X., Lee, K.-F., (1993). On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. Speech audio Process. IEEE Trans. 1, 150–157.

[3] Kinnunen, T., Li, H., (2010). An overview of text-independent speaker recognition: From features to super vectors. Speech Communication 52, 12–40

[4] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., (2007). Automatic speech recognition and speech variability: A review. Speech Communication. 49, 763–786.

[5] Oshaughnessy, D., (2008). Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition 41, 2965–2979.

[6] Wang, L., Wang, J., Li, L., Zheng, T.F., Soong, F.K., (2016). Improving speaker verification performance against long-term speaker variability. Speech Communication 79, 14–29.

[7] Ali, H., Ahmad, N., Zhou, X., Iqbal, K., Ali, S.M., (2014). DWT features performance analysis for automatic speech recognition of Urdu 1–10.

[8] Bharti, R., (2015). Real Time Speaker Recognition System using MFCC and Vector Quantization Technique 117, 25–31.

[9] Kinoshita, K., Delcroix, M., Nakatani, T., (2009). Suppression of Late Reverberation Effect on Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction.

[10] Nehe, N.S., Holambe, R.S., (2012). DWT and LPC based feature extraction methods for isolated word recognition. EURASIP J. Audio, Speech, Music Process.

[11] Yujin, Y., (2010. Research of Speaker Recognition Based on Combination of LPCC and MFCC 765–767.

[12] Kepuska, V.Z., Elharati, H.A., (2015). Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC , PLP , RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions 1–9.

[13] Kaur G., Srivastava M., Kumar A., (2017), "Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition," International Journal of Engineering and Technology Innovation, vol. 7, No. 2, , pp. 78 - 88", ISSN 2223-5329(P), 2226-809X(O).

[14] Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plchot, O., Gonzalez-Rodriguez, J., Moreno, P.J., (2016). On the use of deep feed forward neural networks for automatic language identification. Computer Speech Language 40, 46–59.

[15] Mimura, M., Sakai, S., Kawahara, T., (2015). Reverberant speech recognition combining deep neural networks and deep auto encoders augmented with a phone-class feature. EURASIP J. Adv. Signal Process.

[16] Kaur G., khanna R., Kumar A., (2015), "Review of speech and speech recognition system using feature extraction algorithm and optimization algorithms," International Journal of advanced Trends in Computer Applications, Vol. 2, Number 3, August, ISSN 2395-351

[17] Guojiang, F., (2011). A Novel Isolated Speech Recognition Method based on Neural Network 17, 264–269

[18] Kaur G., Srivastava M., Kumar A., (2017), " Designing and Modeling of Speech and Speaker Recognition System to Control the Wheelchair Paper published in " Advances in Computational Sciences and Technology ISSN 0973-6107, Volume 10, No. 3, pp. 445-460".

[19] Dey, N.S., Mohanty, R., Chugh, K.L., (2012). Speech and Speaker Recognition System Using Artificial Neural Networks and Hidden Markov Model. 2012 Int. Conf. Communication. System Netw. Technol. 311–315. doi:10.1109/CSNT.2012.221

[20] Price, R., Iso, K., Shinoda, K., (2016). Wise teachers train better DNN acoustic models. EURASIP J. Audio, Speech, Music Process.

[21] Seide, F., Li, G., Yu, D., (2011). Conversational Speech Transcription Using Context-Dependent Deep Neural Networks 437–440.

[22] Seltzer, M.L., Yu, D., Wang, Y., (2013). An investigation of deep neural networks for noise robust speech recognition. IEEE International Conference on Acoustic Speech Signal Process. 7398–7402.

[23] Smadi, T. Al, Issa, H.A. Al, Trad, E., Smadi, K.A. Al, (2015). Artificial Intelligence for Speech Recognition Based on Neural Networks 66–72.

