Paper

Genetic Algorithm for Combined Speaker and Speech Recognition using Deep Neural Networks

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

[1] *I. K. Gujral Punjab Technical University, Kapurthala, Jalandhar, India*
[2] *University Institute of Engineering & Technology, Panjab University, Chandigarh, India*
[3] *Chandigarh Engineering College, Landran, Mohali, Punjab, India*
[4] *Central Scientific Instruments Organisation, Chandigarh, India*

https://doi.org/10.26636/jtit.2018.119617

Abstract— Huge growth is observed in the speech and speaker recognition field due to many artificial intelligence algorithms being applied. Speech is used to convey messages via the language being spoken, emotions, gender and speaker identity. Many real applications in healthcare are based upon speech and speaker recognition, e.g. a voice-controlled wheelchair helps control the chair. In this paper, we use a genetic algorithm (GA) for combined speaker and speech recognition, relying on optimized Mel Frequency Cepstral Coefficient (MFCC) speech features, and classification is performed using a Deep Neural Network (DNN). In the first phase, feature extraction using MFCC is executed. Then, feature optimization is performed using GA. In the second phase training is conducted using DNN. Evaluation and validation of the proposed work model is done by setting a real environment, and efficiency is calculated on the basis of such parameters as accuracy, precision rate, recall rate, sensitivity, and specificity. Also, this paper presents an evaluation of such feature extraction methods as linear predictive coding coefficient (LPCC), perceptual linear prediction (PLP), mel frequency cepstral coefficients (MFCC) and relative spectra filtering (RASTA), with all of them used for combined speaker and speech recognition systems. A comparison of different methods based on existing techniques for both clean and noisy environments is made as well.

Keywords— deep neural network, genetic algorithm, LPCC, MFCC, PLP, RASTA-PLP, speaker recognition, speech recognition.

1. Introduction

The study of speech signals and their processing methods is known as speech processing [1]. Speech processing is an immensely vast area and much research has been performed in this field over the past sixty years [2]. Important fields of speech processing are synthesis, recognition and coding of speech signals. Recognition itself is a wide topic consisting of three areas of recognition, i.e. speech, speaker and language. As the name implies, speech recognition aims to recognize the words spoken, while language recognition aims to recognize the language spoken and speaker recognition aims to recognize the speaker. Speech recognition may be speaker dependent and independent. In the speaker dependent mode, the system is trained to recognize one speaker only, but in the speaker independent mode, the system is trained to work with multiple speakers. The field of speaker recognition is also divided into two categories, i.e. text dependent and text independent. In the text dependent speaker recognition mode, the speaker is required to utter words which are known to the system, but in the text independent mode, the speaker may speak any words [3].

A speech signal consists of different attributes, such as loudness, voiced/unvoiced sounds, pitch, fundamental frequency, spectral envelope, formants etc. These attributes help identify the speaker and speech features [4]. Although speech recognition and speaker recognition are different fields, the feature extraction methods in both fields overlap [5]. These methods include predictive models based on the linear predictive coding coefficient (LPCC), perceptual linear prediction (PLP), mel frequency cepstral coefficient (MFCC) and relative spectra filtering (RASTA). These methods can be implemented in speech recognition as well as in speaker recognition [6]–[10]. Speech features can be optimized for improving recognition accuracy with the help of various optimization algorithms, like the genetic algorithm (GA), particle swarm optimization, ant colony search algorithm, etc. [12]. GA can be used, in deep neural networks, for improvement in recognition accuracy [13]–[17]. In past studies, many researchers have implemented GA with an artificial neural network (ANN), i.e. Lan *et al.* [18]. They have implemented GA, instead of the steepest descent method, for updating weights and achieved a 91% recognition accuracy. Balochian *et al.* [19] claimed a 96.49% accuracy level by using GA with the multi-layer perceptron (MLP) classifier.

In this paper, we first implemented some state-of-the-art feature extraction methods for combined speaker and

speech recognition. Out of these methods we have selected the best feature extraction method based upon the results obtained, i.e. MFCC for our application. Further, combined speaker and speech recognition using MFCC with the genetic algorithm and a deep neural network was performed with improved accuracy results achieved.

2. Feature Extraction Techniques

The speech production mechanism can be modeled by a linear separable equivalent circuit [20]–[22]. This model is equivalent to a sound source $G(\omega)$ inputting into the articulation filter (vocal tract) to produce speech. The sound source $G(\omega)$ can be categorized as a train of impulses (voiced) and random noises (unvoiced). Voiced sounds include /a/, /e/, /i/, /o/, /u/. On the other hand, unvoiced sounds are noise generated sounds, such as /t/, /s/. The articulation $H(\omega)$ is a transfer function which models the vocal tract of the human speech organ. The output speech wave $S(\omega)$ is the combination of the sound source multiplied with the articulation given by the equation:

$$S(\omega) = G(\omega)H(\omega). \tag{1}$$

Feature extraction techniques, like LPCC, etc., and models exploit the vocal tract articulation filter $H(\omega)$.

2.1. Linear Predictive Coding Coefficient

LPCC is one of the early algorithms that represent the spectral magnitude of speech signal and generates the vocal tract coefficients. In this method, a speech utterance at the current time can be approximated as linear combination of past speech samples [23]–[25]. The steps are as shown below in Fig. 1.

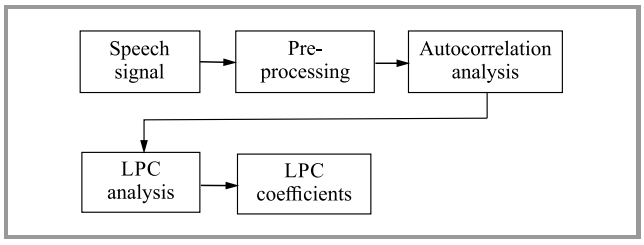


Fig. 1. LPCC technique.

Pre-processing is performed in almost every feature extraction method. The steps of pre-processing include: silence removal from the speech signal, pre-emphasis, framing and windowing. In silence removal, the digitized signal is scanned and the silence zones are removed. Pre-emphasis of the signal is done to enhance the high frequency component of the spectrum. This is performed by passing the speech signal through a digital filter, so that the energy level of the speech signal at higher frequencies is increased:

$$Y[n] = X[n] - 0.95X[n-1]. \tag{2}$$

In framing and windowing, the speech signal is divided into the analysis frames, where the signal can be assumed to be stationary. A window is applied to the emphasized speech signal. Usually the Hamming window is used.

$$W[n] = \begin{cases} 0.56 - 0.46\cos\left(\frac{2\pi m}{L-1}\right), & 0 \le m \le L-1 \\ 0, & \text{elsewhere} \end{cases}. \tag{3}$$

Linear prediction is based on the fact that the present sample $S[n]$ can be linearly predicted using the previous samples $S[n-k]$:

$$S[n] = \sum_{k=1}^{p} \alpha_k S[n-k]. \tag{4}$$

This linear prediction will introduce errors into the sequence of speech samples. This error is known as the residual error $e[n]$:

$$e[n] = s[n] - \sum_{k=1}^{p} \alpha_k S[n-k]. \tag{5}$$

Equation (5) is then transformed into z domain as:

$$E(z) = \left(1 - \sum_{k=1}^{p} \alpha_k z^{-k}\right) S(z). \tag{6}$$

The auto correlation method can be used for estimating LP coefficients. Fundamental frequency or pitch can be identified using an auto correlation analysis. It is based upon determining the correlation between the signal and a delayed version thereof. The next processing step involves a linear prediction coding (LPC) analysis, which converts the auto correlation coefficients into the LPC parameters. The Levinson-Durbin recursive algorithm can be used to identify the coefficients.

2.2. Perceptual Linear Prediction

Perceptual Linear Prediction (PLP) is a method used to obtain more auditory like spectrum based on linear LP analysis of speech. This is a combination of discrete Fourier transform (DFT) and LP techniques and this method is more suitable for the speaker independent mode [26]–[28] (Fig. 2).

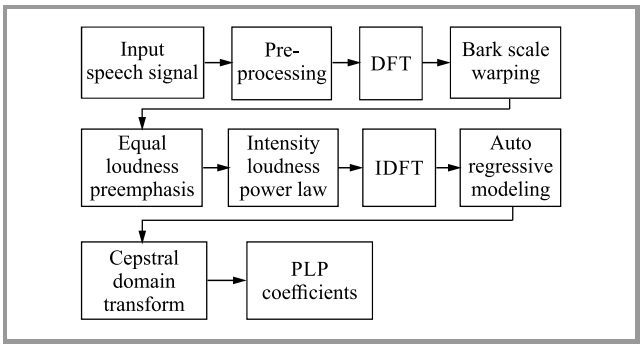


Fig. 2. Perceptual linear prediction technique.

2.3. Rasta Perceptual Linear Prediction (RASTA-PLP)

A band pass filter is added to the PLP algorithm to remove short term noise variations. The individual steps are shown in Fig. 3.

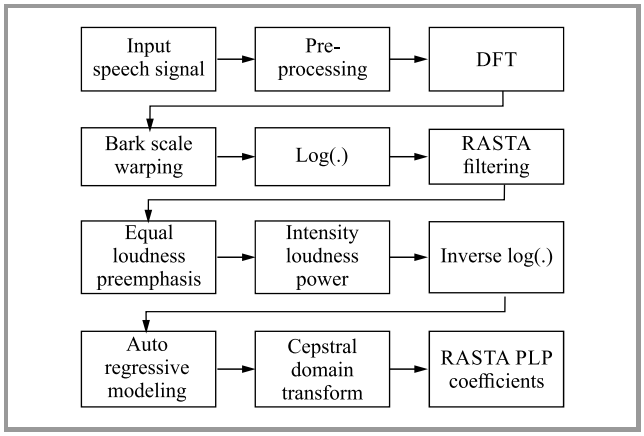


Fig. 3. Relative spectra filtering PLP technique.

2.4. Mel Frequency Cepstral Coefficient (MFCC)

It is the most popular method used for feature extraction [29], [30]. The steps involved are: fast Fourier transform (FFT) is applied first on the frame, and then power spectrum is converted into a mel frequency spectrum. Then, the logarithm of that spectrum is taken and its inverse Fourier transform is taken as shown in Fig. 4.

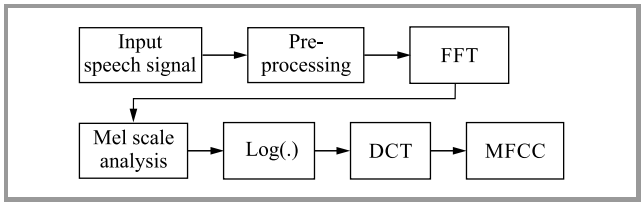


Fig. 4. MFCC extraction.

3. Experiment and Implementation

The speech database is recorded on the sound recorder with the use of headphones, in a room environment, in the mono format. The dataset contains a thousand of words recorded by four speakers aged 27–34, two females (F1, F2) and two males (M1, M2). The recorded words are: *forward*, *backward*, *left*, *right* and *stop*. For each word, fifty samples are taken. All samples are stored in .wav files (16 bps bitrate). All methods, i.e. LPCC, PLP, RASTA PLP and MFCC, are implemented to extract speaker- and speech-specific. Accuracy is calculated in terms of clean signals, as well as of those affected by adding white Gaussian noise (WGN). Further, the MFCC technique is used for feature extraction with GA, and DNN is trained using the optimized features. GA is used for determining the weights and biases of DNN. The fitness function of GA can be defined according to specific requirements.

In the proposed work, f_s is the current selected feature and f_t is the threshold value of feature points. On the basis of a given condition, the fit value is checked which can exist in a new feature set:

$$f(fit) = \begin{cases} 1, & f_s < f_t \\ 0, & f_s \geq f_t \end{cases}, \tag{7}$$

where $f(fit)$ is the fit value according to the fitness function. If the condition is true (1), then GA creates an optimal feature set. The genetic parameters and operators used are: population size, crossover function, mutation function and selection function. To organize the feature sets according to the requirements, selection of individual features is performed by means of the selection function. The selection of individual features is done according to their fitness value represented as f_s and is given by:

$$f_s = \sum_{i=1}^{popsize} f(i), \tag{8}$$

where $f(i)$ describes the individually selected features and $popsize$ denotes the population size of GA. The fitness

Algorithm 1: Optimization technique for DNN training

1: Load speech feature sets
2: Calculate the length of feature $[r,c]$
3: Define genetic parameters and operators to initialized the genetic algorithm
4: Set population size $popsize = 50$ (when number of variables is lower than 50 then value of 50 is still sufficient for optimization)
5: Selection function = handle to the function that selects parents of crossover from feature sets
6: Crossover function = handle to the function that the genetic algorithm uses to create the optimal solution
7: Mutation function = handle to the function that produces mutation children which are called optimized features
8: Define fitness function using Eq. (8)
9: **For all components of feature according to rows**
For all components of feature according to columns

$$fs = \sum_{i=1}^{popsize} f(i)$$

$$ft = \frac{\sum_{i=1}^{popsize} f(i)}{Length\ of\ feature}$$

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

$$No.\ of\ variables = 1$$

$$O_{value} = GA[f(fit), no.\ of\ variables,\ initialized\ parameters]$$

End
End
Training data = O
For each set of $Training\ data$
$Group = Training\ data(i)$
End
10: Initialize the DNN using Training data and Group
11: Train and save the DNN and create a trained structure for classification

function is defined in terms of the distance measured between the selected value and threshold values of features based on the crossover function. Crossover and mutation function are the operators used to establish the relationship between the selected feature f_s and the threshold feature value f_t. A crossover function is based on an individual feature (parents) and a new individual feature (children), while mutation changes the genes of one individual to produce a new feature (mutant), according to the fitness function [35]–[36]. New optimized feature sets are transferred to DNN as input or a training set, to create a trained DNN structure for classification. The methodology of the proposed genetic algorithm with DNN is described as Algorithm 1.

We have used the *trainlm* training function because it is the fastest back propagation algorithm. It is based on the Levenberg-Marquardt optimization algorithm [31]–[34]. During the training phase, we have used a set of 5 hidden layers and weights, and bias values were updated according to the Levenberg Marquardt optimization algorithm. After training, we have performed a simulation with a test speech signal and the process was repeated for training and testing

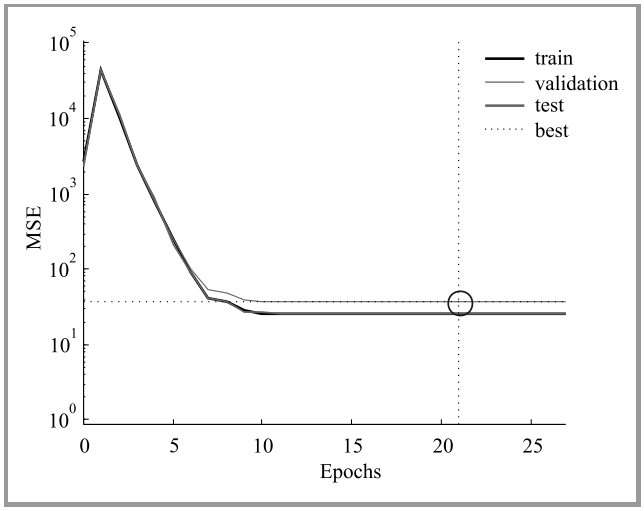


Fig. 5. MSE curve.

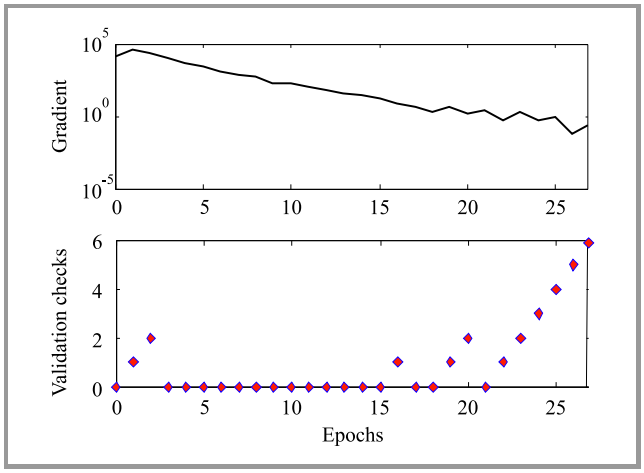


Fig. 6. DNN parameters.

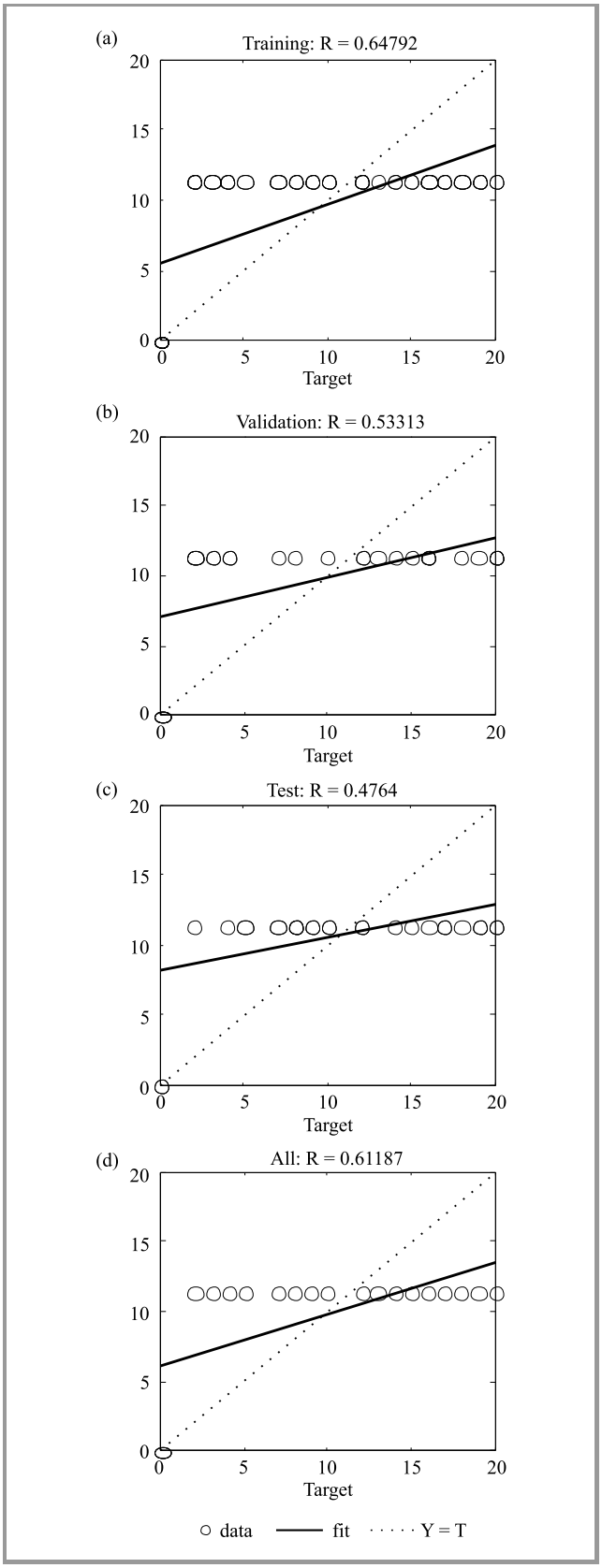


Fig. 7. DNN dataset: (a) training, (b) validation, (c) test, and (d) training output.

phases. We have checked the performance on the basis of mean of square errors (MSE). The MSE graph of the pro-

posed work is given in Fig. 5 with respect to the epochs. The epochs denote the number of iterations which is used by DNN during the speech feature training phase.
The circle shows the best performance in terms of MSE (37.4629 at iteration number 21). Validation and test curves are very similar. If the test curve increases drastically before the validation curve increases, then it is possible that overfitting might have occurred. The next step is to validate the network for which a decay plot is generated to show the association between the outputs of the network and the targets. If training is ideal, network outputs and targets would be equal, but the connection is rarely perfect in practice.
Figure 6 shows a graph presenting different types of parameters, such as gradient value and validation check, with respect to the epochs which are generated during the dataset training phase, using the DNN as a classifier.
Figure 7 shows a description of datasets which are used for the purpose of training. The solid line shows the finest fit linear decay line between outputs and targets. The R value is a signal of the bond between outputs and targets. If $R = 1$, 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

4. Results and Discussions

All feature extraction methods discussed, i.e. LPCC, PLP, RASTA PLP and MFCC, are used with the recorded database to extract the speaker- and speech-specific features, and results are evaluated in Matlab. Accuracy is calculated in the clean version, as well as in one with WGN added to the speech samples, as shown in Tables 1–4.
Table 1 shows that the average recognition rate related to speaker and words, achieved by using LPCC, equals, for a clean environment, 93.12%. However, by adding WGN to the speech signal, the recognition rate decreases to 83.48%. In Table 2 the feature extraction method used is PLP,

Table 1
Accuracy [%] in clean and with WGN
using LPCC technique and speech recognition
for two males (M1, M2) and two females (F1, F2)

Speaker	M1	M2	F1	F2
Backward	93.18	92.78	92.63	93.65
Backward with WGN	83.50	83.53	84.31	84.22
Forward	94.58	93.56	94.59	92.62
Forward with WGN	81.19	81.20	85.36	82.34
Left	95.50	94.37	91.14	94.41
Left with WGN	80.62	84.76	82.02	82.01
Right	91.17	91.27	95.44	91.79
Right with WGN	83.94	83.46	84.45	84.15
Stop	94.64	92.35	91.33	91.48
Stop with WGN	85.34	84.82	84.39	84.22

Table 2
Accuracy [%] in clean and with WGN
using PLP technique

Speaker	M1	M2	F1	F2
Backward	92.59	94.47	92.60	93.24
Backward with WGN	84.62	83.73	82.21	85.17
Forward	95.52	90.12	92.24	94.43
Forward with WGN	84.43	85.32	82.35	82.76
Left	90.66	94.09	92.54	91.54
Left with WGN	84.11	83.00	81.15	81.06
Right	90.90	92.57	94.59	95.37
Right with WGN	84.37	85.06	83.36	84.43
Stop	94.46	93.54	94.22	93.89
Stop with WGN	84.93	81.97	82.49	84.74

Table 3
Accuracy [%] in clean and with WGN
using RASTA-PLP technique

Speaker	M1	M2	F1	F2
Backward	91.18	92.35	95.43	91.31
Backward with WGN	84.59	85.46	81.71	84.58
Forward	91.32	93.09	94.49	92.19
Forward with WGN	82.39	82.51	80.95	82.26
Left	91.98	93.92	92.54	94.49
Left with WGN	85.47	83.99	84.96	82.83
Right	92.81	93.72	93.35	92.54
Right with WGN	81.63	82.73	82.21	84.58
Stop	93.59	94.87	92.37	93.12
Stop with WGN	82.45	81.35	81.64	83.86

Table 4
Accuracy [%] in clean and with WGN
using RASTA-PLP technique

Speaker	M1	M2	F1	F2
Backward	94.19	94.42	94.66	92.51
Backward with WGN	82.14	84.12	83.95	84.21
Forward	94.69	93.58	94.31	95.55
Forward with WGN	83.38	84.95	81.57	85.29
Left	94.73	94.61	94.22	93.66
Left with WGN	83.86	82.16	83.12	85.22
Right	94.63	94.45	94.41	93.96
Right with WGN	82.89	84.81	84.70	84.38
Stop	93.47	94.19	94.33	94.66
Stop with WGN	83.29	85.40	85.23	83.57

and the average speaker and word recognition rate equals, for a clean environment, 93.17%. However, by adding

WGN to the speech signal, the rate of recognition decreases to 83.56%. Similarly, Table 3 shows that the average speaker and word recognition rate achieved using RASTA PLP equals, for a clean environment, 93.16 and 83.10% respectively. The last Table 4 shows that the average speaker and word recognition rate achieved by using MFCC for a clean environment equals 94.25% and 83.98% with WGN. Figures 8 and 9 show the results of Tables 1–4. Based on the results, we have found that MFCC feature extraction method is best for our application in clean and noisy .

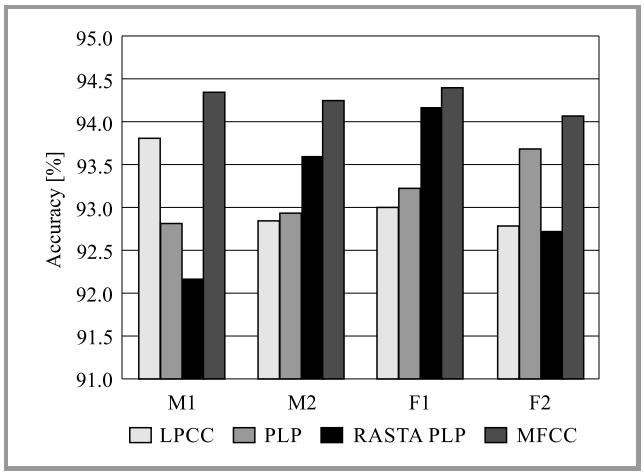


Fig. 8. Accuracy in clean environment.

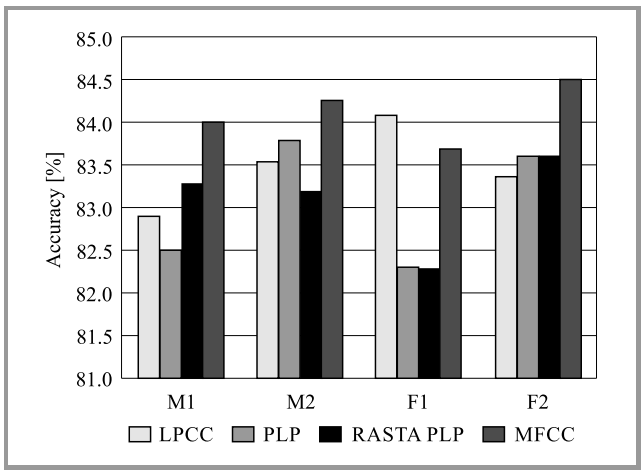


Fig. 9. Accuracy with WGN added.

5. Optimization using GA to assist in DNN Training

In this section results are shown for a system using GA and DNN. It is quite difficult to recognize speech in the presence of noise. The proposed work is tested with various types of noise, such as White Gaussian Noise (WGN), Additive White Gaussian Noise (AWGN), etc. Due to noise, recognition becomes difficult. Therefore, we have used GA for feature optimization. The experimental results have confirmed our expectations by giving good values in terms of such measurement metrics as precision rate, recall rate, accuracy, sensitivity and specificity, defined as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN},$$

$$Precision\ rate = \frac{TP}{TP+FP},$$

$$Sensitivity = \frac{TP}{TP+FN},$$

$$Specificity = \frac{TN}{FP+TN},$$

where true positive (TP) represents the truly selected feature sets using and false positive (FP) are the falsely selected feature sets during the classification of signals. True negative (TN) are all negative features which are really true and false negative (FN) are all negative features which are really false. Figure 10 shows the Receiver Operating Characteristics (ROC) curve. It is a graphical method for comparing two empirical distributions. In this work, true positive and false negative parameters have been taken.

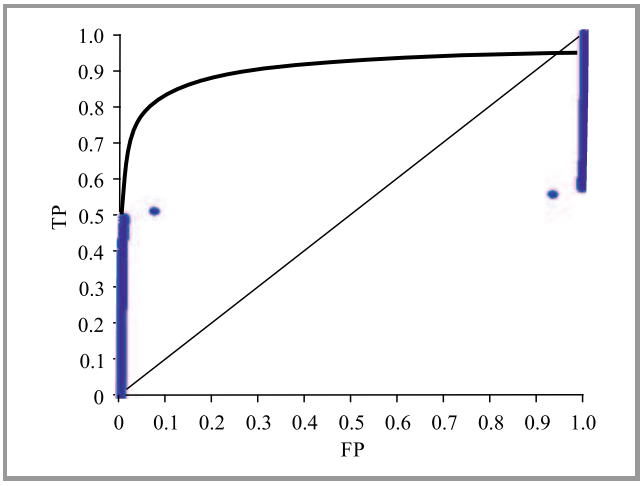


Fig. 10. ROC curve for the proposed work.

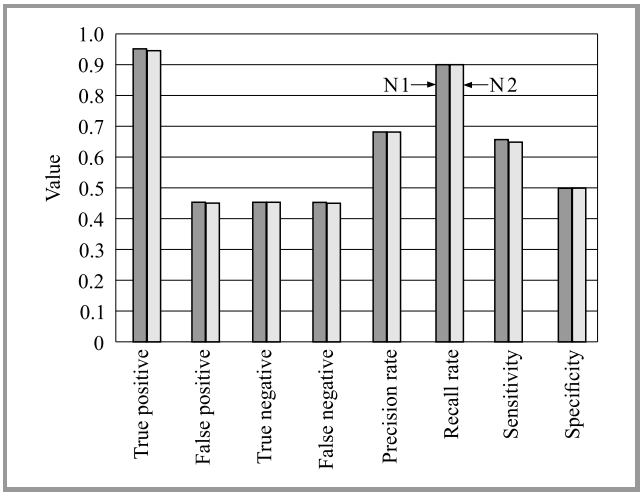


Fig. 11. Result evaluation for proposed methodology.

Table 5
Analysis of proposed metric results

Speaker	N1	N2
TP	0.949	0.945
FP	0.448	0.447
TN	0.449	0.449
FN	0.448	0.445
Precision rate	0.679	0.678
Recall rate	0.901	0.901
Accuracy	97.05	97.11
Sensitivity	0.655	0.645
Specificity	0.500	0.501

Figure 11 and Table 5 present the parameters calculated in the proposed work for two different speakers. N1 represents Speaker 1 and N2 is Speaker 2.
A comparison is also made between MFCC+DNN and MFCC+GA+DNN, as shown in Table 6 and Fig. 12. Recognition accuracy is higher when GA is used for the optimization of features.

Table 6
Comparison of accuracy between
MFCC+DNN and MFCC+GA+DNN

No. of iterations	MFCC+DNN	MFCC+GA+DNN
1	94.48%	97.19%
2	93.23%	98.73%
3	94.11%	95.57%
4	94.15%	96.45%
5	94.47%	94.57%
Average	94.08%	96.51%

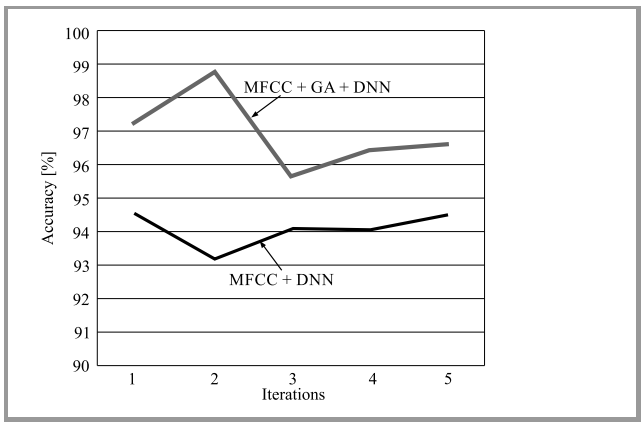


Fig. 12. Accuracy comparison.

6. Conclusions

The existing feature extraction techniques, such as LPCC, PLP, RASTA PLP and MFCC, used for combined speaker and speech recognition, are implemented for five words recorded by four persons in clean and noisy environments. The results show that out of four techniques, MFCC offers the best results in clean, as well as in noisy environments, i.e. the average percentage accuracy for combined speaker and speech recognition in a clean environment is higher than 94%, and with WGN added to the signal – it is higher than 83.5%. We have shown that speaker and speech recognition systems with MFCC and GA using DNN are helpful in achieving a higher accuracy. The experimental results indicate that the proposed method has provided good results, offering the following values: true positive 0.949, false positive 0.448, true negative 0.449, false negative 0.448, precision rate 0.679, and the following rates: recall 0.901, accuracy 96.51, sensitivity 0.655 and specificity 0.500. All these values are an improvement over the existing methods.

References

[1] D. R. Reddy, "Speech recognition by machine: A review", *Proc. of the IEEE*, vol. 64, no. 4, pp. 501–531, 1976 (doi: 10.1109/PROC.1976.10158.

[2] S. Furui, "50 Years of Progress in Speech and Speaker Recognition Research", *ECTI Transact. on Comput. and Infor. Technol.*, vol. 1, no. 2, pp. 64–74, 2005.

[3] J. Campbell, "Speaker recognition: A tutorial", *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997 (doi: 10.1109/5.628714).

[4] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition", *Speech Communic.*, vol. 50, no. 10, pp. 782–796, 2008 (doi: 10.1016/j.specom2008.04.010).

[5] I. Bhardwaj, "Speaker dependent and independent isolated Hindi word recognizer using hidden Markov model (HMM)", *Int. J. of Comp. Applic.*, vol. 52, no. 7, pp. 34–40, 2012.

[6] S. Squartini, E. Principi, R. Rotili, and F. Piazza, "Environmental robust speech and speaker recognition through multi-channel histogram equalization", *Neurocomputing*, vol. 78, no. 1, pp. 111–120, 2012 (doi: 10.1016/j.neurocom.2011.05.035).

[7] N. S. Dey, R. Mohanty, and K. L. Chugh, "Speech and speaker recognition system using artificial neural networks and hidden Markov model", in *Proc. IEEE Int. Conf. on Communic. Sys. and Network Technol. CSNT*, Bhopal, Madhya Pradesh, India, 2012, pp. 311–315 (doi: 10.1109/CSNT.2012.221).

[8] T. Gaafar, H. Bakr, and M. Abdalla, "An improved method for speech/speaker recognition", in *Int. Conf. on Infor., Electr. and Vision ICIEV*, Dhaka, Bangladesh, 2014 (doi: 10.1109/ICIEV.2014.6850693).

[9] T. A. Smadi, "An improved real-time speech signal in case of isolated word recognition", *Int. J. of Engineer. Research and Applic.*, vol. 3, no. 5, pp. 1748–1754, 2013.

[10] V. Fontaine and H. Bourlard, "Speaker dependent speech recognition based on phone–like units models application to voice dialing", in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Proces. ICASSP'97*, Munich, Bavaria, Germany, 1997, pp. 2–5 (doi: 10.1109/ICASSP.1997.596241).

[11] S. J. Wright, D. Kanevsky, and L. Deng, "Optimization algorithms and applications for speech and language processing", *IEEE Transact. on Audio, Speech and Lang. Proces.* vol. 21, no. 11, pp. 1527–1530, 2013 (doi: 10.1109/TASL.2013.2283777).

[12] M. Mitchell, "Genetic algorithms: An overview 1", *Complexity*, vol. 1, pp. 31–39, 1995 (doi: 10.1102/cplx.6130010108).

[13] M. Sarma, "Speech recognition using deep neural network – recent trends", *Int. J. of Int. Sys. Design and Computing*, vol. 1, no. 12, pp. 71–86, 2017 (doi: 10.1504/IJISDC.2017.082853).

[14] L. Deng, M. Seltzer, D. Yu, A Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto – encoder", in *Proc. 11th Int. Conf. on Speech Commun. Assoc.*, Makuhari, Chiba, Japan, 2010, pp. 1692–1695, 2010.

[15] F. Guojiang, "A novel isolated speech recognition method based on neural network", *2nd Int. Conf. on Network. and Infor. Technol.*, Singapore, 2011, vol. 17, pp. 264–269.

[16] I. Lopez-Moreno *et al.*, "On the use of deep feed forward neural networks for automatic language identification", *Computer Speech Lang.*, vol. 40, no. C, pp. 46–59, 2016 (doi: 10.1016/j.csl.2016.03.001).

[17] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep auto encoders augmented with a phone-class feature", *EURASIP J. on Advances in Signal Proces.*, vol. 62, p. 13, 2015 (doi: 10.1186/s13634-015-0246-6).

[18] M. L. Lan, S. T. Pan, and C. C. Lai, "Using genetic algorithm to improve the performance of speech recognition based on artificial neural network", in *1st Int. Conf. on Innovative Computing, Infor. and Control – Vol. I ICICIC'06*, Beijing, China, 2006, vol. 2, no. 1, pp. 6–9 (doi: 10.1109/ICICIC.2006.372).

[19] S. Balochian, E. A. Seidabad, and S. Z. Rad, "Neural network optimization by genetic algorithms for the audio classification to speech and music", *Int. J. of Signal Proces., Image Proces. and Pattern Recog.*, vol. 6, no. 3, pp. 47–54, 2013.

[20] S. King, J. Frankel, K. Livescu, and E. Mcdermott, "Speech production knowledge in automatic speech recognition", *J. of the Acoustic. Soc. of America*, vol. 121, no. 2, pp. 723–742, 2007 (doi: 10.1121/1.2404622).

[21] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech", in *Proc. Conf. on Speech Prosody*, Boston, MA, USA, 2016 pp. 84–88 (doi: 10.21437/SpeechProsody.2016-18).

[22] M. Honda, "Human speech production mechanisms", *NTT Technic. Rev.*, vol. 1, no. 2, pp. 24–29, 2003.

[23] N. S. Nehe and R. S. Holambe, "DWT and LPC based feature extraction methods for isolated word recognition", *EURASIP J. on Audio, Speech, and Music Proces.*, vol. 7 pp. 1–7, 2012 (doi: 10.1186/1687-4722-2012-7).

[24] A. Pramanik and R. Raha, "Automatic speech recognition using correlation analysis", in *Proc. World Cong. on Infor. and Commun. Technol. WICT*, Trivandnum, Kerala, India, 2012 (doi: 10.1109/WICT.2012.6409160).

[25] X. Zhang, Y. Guo, and X. Hou , "A speech recognition method of isolated words based on modified LPC cepstrum", in *IEEE Granular Computing Conf.*, San Jose, CA, USA, 2007 (doi: 10.1109/GrC.2007.96).

[26] I. Hermansky, K. Tsuga, S. Makino, and H. Wakita, "Perceptually based processing in automatic speech recognition", in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Proces. ICASSP'86*, Tokyo, Japan, 1986 (doi: 10.1109/ICASSP.1986.1168649).

[27] S. Swamy and K. V. Ramakrishnan, "An efficient speech recognition", *Int. J. of Comp. Science and Engineer.*, vol. 3, no. 4, pp. 21–27, 2013 (doi: 10.5121/cseji.2013.3403).

[28] H. Ali, N. Ahmad, X. Zhou, K. Iqbal, and S. M. Ali, "DWT features performance analysis for automatic speech recognition of Urdu", *Springer Plus*, vol. 3, pp. 1–10, 2014 (doi; 10.1186/2193-1801-3-204).

[29] G. Kaur, R. Khanna, and A. Kumar, "Automatic speech and speaker recognition using MFCC: Review", *Int. J. of Advances in Science and Technol.*, vol. 2, no. 3, 2014.

[30] G. Kaur, R. Khanna, and A. Kumar, "Implementation of Text Dependent Speaker Verification on Matlab", in *Proc. 2nd Conf. on Recent Adv. in Engineer. and Comput. Sciences RAECS*, Chandigarh, India, 2015 (doi: 10.1109/RAECS.2015.7453344).

[31] R. Price, K. Iso, and K. Shinoda, "Wise teachers train better DNN acoustic models", *EURASIP J. of Audio, Speech, Music Proces.*, vol. 10, art. no. 88, 2016 (doi: 10.1186/s13636-016-0088-7).

[32] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models", *Digital Signal Proces.*, vol. 10, pp. 19–41, 2000 (doi: 10.1006/dspr.1999.0361).

[33] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks", *Interspeech*, pp. 437–440, 2011.

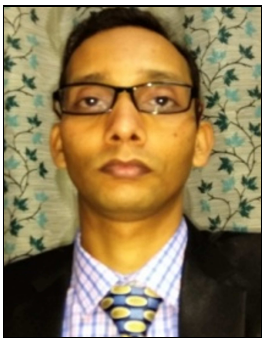
[34] M. L. Seltzer, D. Yu, Y. Wang, "An investigation of deep neural networks for noise robust speech recognition", in *IEEE Int. Conf. on Acoust. Speech Signal Proces. ICASSP'13*, Vancouver, BC, Canada, 2013 (doi: 10.1109/ICASSP.2013.6639100).

[35] S. Casale, A. Russo, and S. Serrano, "Classification of speech under stress using features selected by genetic algorithms", in *Proc. 14th European Signal Proces. Conf.*, Florence, Tuscany, Italy, 2006 pp. 1–4.

[36] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers", *Engineer. Applic. of Artif. Intel.*, vol. 51, pp. 191–201, 2016 (doi: 10.1016/j.engappai.2016.01.012).

Gurpreet Kaur is an Assistant Professor at the Department of Electronics and Communication Engineering at University Institute of Engineering and Technology, Panjab University, Chandigarh, India. She received her B.Tech. (with honors) in Electronics and Communication Engineering from Kurukshetra University, Haryana in 2004, M.Eng. (with distinction) in Electronics and Communication from the University Institute of Engineering and Technology, Panjab University, Chandigarh in 2007 and is pursuing Ph.D. in Electronics Engineering from IKG Punjab Technical University, Jalandhar. Her current research interests are speech processing and neural networks.
E-mail: regs4gurpreet@yahoo.co.in
I. K. Gujral Punjab Technical University
Kapurthala, Jalandhar, India

University Institute of Engineering & Technology
Panjab University
Chandigarh, India



Mohit Srivastava is a Professor at the Department of Electronics and Communication Engineering and R&D Dean at Chandigarh Engineering College, Landran, Mohali, Punjab, India. He received his B.Tech. in Electronics and Communication Engineering from Magadh University, Bodh Gaya, M.Tech. in Digital Electronics and Systems from K.N.I.T. Sultanpur and Ph.D. in Image processing & Remote Sensing from Indian Institute of Technology Roorkee in 2000, 2008 and 2013 respectively. He has more than 15 years of work experience in various environments, including industry, as well as educational and research centers. He has completed two IEDC (DST) funded projects. His current research interests are digital image and speech processing, remote sensing and their applications in land cover mapping, as well as communication systems.
E-mail: mohit.ece.@cgc.edu.in
Chandigarh Engineering College
Landran, Mohali, Punjab, India

Amod Kumar received his B.E. (Hons) in Electrical and Electronic Engineering from Birla Institute of Technology and Science, Pilani, M.E. in Electronics from Punjab University, Chandigarh and Ph.D. in Biomedical Signal Processing from IIT Delhi. He has about 38 years of experience in research and development of different instruments in the area of process control, environmental monitoring, biomedical engineering and prosthetics. He is currently working as Chief Scientist at Central Scientific Instruments Organisation (CSIO), Chandigarh, which is a laboratory of CSIR. He has more than 70 publications in reputed national and international journals. His areas of interest are digital signal processing, image processing and soft computing.
E-mail: csioamod@yahoo.com
Central Scientific Instruments Organisation
Chandigarh, India