Robust Hearing-Impaired Speaker Recognition from Speech using Deep Learning Networks in Native Language

Jeyalakshmi Chelliah
Department of ECE, K.Ramakrishnan College of Engineering, India
lakshmikrce.2016@gmail.com

KiranBala Benny
Department of Artificial Intelligence and Data Science, K.Ramakrishnan College of Engineering, India
kiranit2010@gmail.com

Revathi Arunachalam
School of EEE, Sastra Deemed to be University, India
revathidhanabal@rediffmail.com

Viswanathan Balasubramanian
Department of ECE, K.Ramakrishnan College of Engineering, India
viswa.ece@krce.ac.in

Abstract: *Several research works in speaker recognition have grown recently due to its tremendous applications in security, criminal investigations and in other major fields. Identification of a speaker is represented by the way they speak, and not on the spoken words. Hence the identification of hearing-impaired speakers from their speech is a challenging task since their speech is highly distorted. In this paper, a new task has been introduced in recognizing Hearing Impaired (HI) speakers using speech as a biometric in native language Tamil. Though their speech is very hard to get recognized even by their parents and teachers, our proposed system accurately identifies them by adapting enhancement of their speeches. Due to the huge variety in their utterances, instead of applying the spectrogram of raw speech, Mel Frequency Cepstral Coefficient features are derived from speech and it is applied as spectrogram to Convolutional Neural Network (CNN), which is not necessary for ordinary speakers. In the proposed system of recognizing HI speakers, is used as a modelling technique to assess the performance of the system and this deep learning network provides 80% accuracy and the system is less complex. Auto Associative Neural Network (AANN) is used as a modelling technique and performance of AANN is only 9% accurate and it is found that CNN performs better than AANN for recognizing HI speakers. Hence this system is very much useful for the biometric system and other security related applications for hearing impaired speakers.*

Keywords: *Speaker recognition, voice impaired, energy, deep learning based convolutional neural network, mel frequency cepstral coefficient, Auto associative neural network, back propagation algorithm.*

Received November 26, 2020; accepted December 26, 2021
https://doi.org/10.34028/iajit/20/1/11

1. Introduction

Speaker identification or recognition refers to the process of identifying a person from a group of people given as input. Speaker verification, is a different process which involves giving different samples of the same speaker to verify whether he is the authorized person or not. Speaker identification is probably utilized for criminal investigations and in media where people are recognized by their voice. Speaker verification, in most cases, is utilized in banks and secure confidential areas.

But the challenging thing in speaker identification is the various emotional states of the speakers considered due to the nonlinear behaviour of their speech. Speaker recognition system with respect to 3 emotional states in two different languages: Arabic and English is proposed [15]. When considering all biometrics methods, human speech biometric utilizes acoustic information [1]. Input data of dysarthric speech is preprocessed for better result and this work justifies that Empirical Mode Decomposition and Hurst-Based mode selection (EMDH-CNN) system is more efficient and effective than Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM) [21]. Recognizing the speech of dysarthria and cerebral palsy patients with limited training data has been done by the application of HMM/Artificial Neural Network hybrid networks [16]. Based on this, multilayer perceptron neural network and convolutional neural networks has been tried for recognizing hearing impaired speakers instead of state of the art HMM technique. Some of the research works using machine learning techniques are discussed here.

Now a days, many research works in speech are effectively using Convolutional Neural Network (CNNs) [23]. Using a grid like structure CNNs are employed in image processing for processing data. Later on, it is utilized for speech and language processing. Initially the local features from the input are captured by the convolutional layer which comprises of multiple filters followed by the pooling layer. It is used to condense the aspect of each feature map while holding the most important features. Then the last layer is used to attain the required calculation for regression or classification tasks which is the fully connected layer. Another novel work has Recurrent Convolution Neural Network (RCNN) for speech processing which is used to identify the temporal speech and for classification Support Vector Machine (SVM) is utilized [24]. Deep Convolution Neural Network (DCNN), which gives effective results in terms of fast convergence and better results comparable to CNN, is proposed by the author [2].

Recurrent Neural Network (RNN) proposed for Multidimensional Recurrent Neural Network (MDRNN) which increases the performance when compared to the CNN [13]. Similarly, convolution and Recurrent Neural Network are utilized for speaker identification [14], which uses Multilayer Perception for both spatial and temporal parameters to identify the visible and Non visible noise. The above literature depicts that, for normal speech recognition, various types of CNN will be the best choice and in this proposed work we have utilized this method for hearing impaired speaker identification. Even for normal speakers due to climatic changes, different ways of speaking style of various regions and physical problems, the voice may vary for the same individual. In this scenario, if we consider the hearing-impaired speaker, in spite of these variations due to the inability of hearing, they don’t know how to speak a particular phoneme or word. If the severity of the problem is more such as hard of hearing or profoundly deaf, their speech is very worse. So by adopting some preprocessing technique their speech can be improved and then speaker recognition can be done. The remainder of the paper is organized as follows. Section 2 describes about the research work related to the proposed system. Details regarding the feature extraction and data base are depicted in section 3 followed by CNN model in section 4 and Back propagation neural network in section 5. Results are discussed in section 6 and conclusion is given in section 7.

2. Related Works

Speaker recognition is a major work in which with the help of input as spectrograms, processing of the data set using convolution process is done initially and then the identification of the speaker is done after finally clustering the data with trained data [22]. The main objective of this paper [7] is, to characterize and identify the speaker information and this process can be done by Mel Frequency Cepstral Coefficient (MFCC) and vector quantization. In this paper three feature extractions made, namely Speaker Modeling, Feature Matching and Decision logic is done to identify the speaker. To identify and characterize a speaker some steps are followed, namely, silence detection, windowing, Fast Fourier Transform (FFT), cepstrum and Mel Frequency Cepstral Coefficients (MFCC). In the implementation part, two categories, namely single user and multiple user with FAR, FRR with accuracy is done for effective implementation. Automatic real time speaker identification using statistical feature and Gabor filter with Random forest algorithm produces effective results [11].

This research work regarding speaker recognition has two types of speaker identification. One is text dependent another one is text independent. Text dependent means user speaks existing sentences for evaluation process whereas text independent means user speaks different sentences for the evaluation process. Here for feature extraction Deep Neural Network (DNN) is used for the entire process and for verification purpose, GMM-Univeral Background Model (UBM) is also proposed [20]. Another interesting research work in speaker recognition was verification and identification from singing, using conventional Gaussian mixture model, dynamic time warping and DNN. For speaker identification process pitch and MFCC is considered as a feature and for verification GMM, Dynamic Time Warping (DTW) and DNN are also investigated and final results show that DTW was the most effective method for text dependent process compared to GMM [25].

Survey about robust speaker recognition by the perspectives of domain adaptation and speech enhancement is dealt in detailed manner [5]. Deep learning approach is utilized by the author to train Artificial Neural Network (ANN) and CNN in which ANN is fed on diverse extracted features, while CNN is trained on spectrograms. Then transfer learning strategy is used on both methods to get a reasonable output using limited data [8]. Feed-forward neural network with optimized particle swarm optimization for speaker recognition has been proposed and recognition accuracy of 97.83% in clean voice environments is achieved [3]. Speaker identification in a noisy environment using CNN and MFCC in a text-independent condition is proposed and 87.5% accuracy is achieved [4].

By considering the previous literature works, in this proposed article, text dependent speaker identification for hearing impaired speakers using CNN has been attempted. In this article 10 HI speakers have been considered and then their speech is applied to MFCC block to extract the features which is converted into 2D and given as input to CNN comprising of various no. of convolution, pooling and fully connected layers. Then this network is compared with the Multilayer perceptron with Back Propagation Algorithm (BPNN) which is a kind of auto associative neural network AANN.

3. Materials and Methods

Any speech or speaker recognition system needs two major functions, i.e., feature extraction and recognition employing any state-of-the-art methods. Recognition accuracy will be achieved at a higher rate, only if the features are distinct. Then only the characteristics of different speech or speakers will be unique. At present, plenty of features are proposed by many researchers and each type has its own pros and cons. In general, according to the literature, for speech or speaker recognition MFCC features are mostly used by all. As far as CNN is concerned, the raw speech or image can be applied for recognition which is the special trait of this network and there is no need to extract the features manually. For image processing applications, we can apply the 2D image directly to the CNN networks, which is not the case for speech processing applications. Since speech signal is a 1D input, the spectrogram of the speech signal is given as 2D image data to CNN. But few researchers extracted MFCC from speech and combine them with the CNN features extracted from spectrograms of speech, and using a proper feature on them, these selected features are applied to a DNN for classification [5]. In spite of MFCC and CNN combined, in most of the situationsgood accuracy couldn't beachieved for speaker recognition. Often there isalso a greater degree of audio degradationsdue to environmental noise, performance of the microphone, and distance of the speakers from the microphone which can also result in degradations in the speech of the individuals. In such circumstances, the degraded signals are unreliable for speaker recognition and in a few applications like forensic, it is very tough to identify a speaker from his voice. Hence MFCC features are not reliable for degraded speech since it captures only the perception characteristics of speech. At the same time LPC (Linear prediction coefficients) captures speech production features from the speech. So these two features are combined in a novel manner by using a 1D Triplet CNN to recognize the degraded speech of the speakers [6]. So in our proposed work, MFCC features have been utilized along with CNN to recognize the hearing impairedspeech instead of the spectrogram of raw speechwhich is the challenging scenario.

3.1. Data Base Creation

There are several data sets made use for speech processing applications like TIMIT, NIST SRE 2008 and 2010. But for HI speakers no standard database is available and hence they are asked to speak a simple Tamil word twenty times. So our database consists of 200 utterances from 10 speakers. Among this 50% is utilized for training and remaining used for testing.

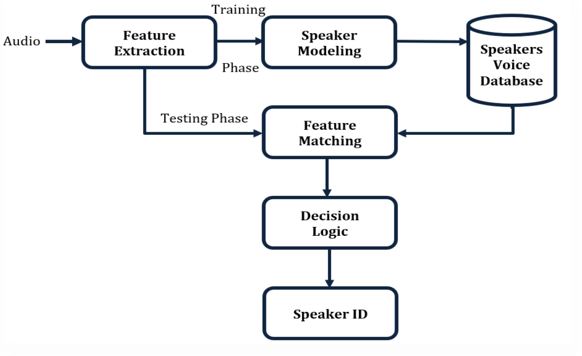


Figure 1. General procedure for speaker identification system.

Similarly, speaker recognition is often categorized into text-dependent and text-independent methods. With text-dependent speaker recognition the same sentence is used for training and testing while for text-independent speaker recognition any sentence can be spoken by the speakers for training and testing. So, their vocal characteristics play a major role. In this work we will be focusing on text-dependent speaker recognition since every HI speaking style is different according to the severity of their hearing problem and the speech therapy they have undergone [12].

Figure 1 shows the general block diagram for speaker identification or recognition system. From the speaker's speech database, initially, particular features are extracted and models are created for each speaker using any specific method and stored in speaker's voice database. Whenever an input speech is given, features are extracted from the speech and it is compared with the already trained model in the feature matching block to identify the specific speaker.

3.2. MFCC Feature Extraction

Based on the human being's perception of critical bandwidth, MFCCs are calculated. At the same time the mel filters should be linearly spaced at low frequencies since up to 1000Hz normal and mel frequencies are same. Beyond that, it is logarithmically spaced to seize the phonetic important characteristics of speech. It is explained in detail step by step in Figure 2.

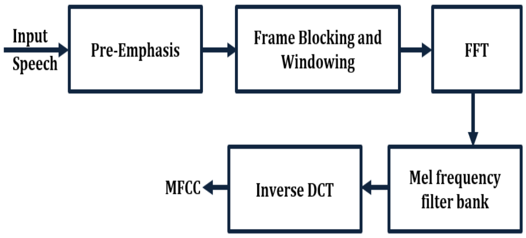


Figure 2. Stepsto extract MFCC features.

MFCC is an auditory kind of feature set which depends on the auditory properties of our ear. Any input signal is first filtered by using first order FIR filter in order to remove any unwanted noise present in the signal [17]. Followed by this pre-emphasized input, it is divided into several frames. In general, any realtime signal will not be stationary for a long time and hence to analyze these signals it should be converted into many frames. In order to eliminate the signal discontinuities, the signal is windowed using a hamming window which is frequently employed according to the literature. Then N- point FFT is calculated for each frame and it is filtered by a set of 40 band pass filters and for each band the power is calculated. Then for each frequency ‘f’ mel frequency is calculated using the Equation (1).

$$mel(f) = 2595 * \log(1 + f(Hz)/700) \quad (1)$$

If we are using only spectrum for our CNN, then we can use these spectrum coefficients directly as a spectrogram. But in order to estimate MFCC, Mel-Frequency cepstrum is derived using Equation (2).

$$C_n = \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\pi / K] \quad (2)$$

n=1, 2 **...** L

In this, S_k is k^{th} filter output power of the filter bank and total no. of cepstrum is given by L. In our work, the frame length is 20msec with 10msec overlapping and 13 coefficients are taken for each frame.

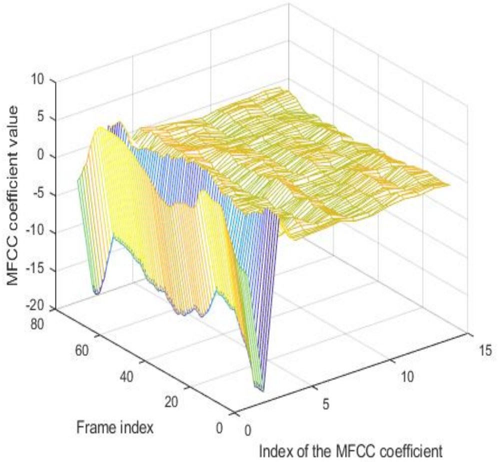


Figure 3. MFCC coefficient value of a male speaker for the tamil word ondru.

4. CNN Model

CNN is a kind of supervised neural network, which consists of several layers in addition to the hidden layer and the output layer in an ordinary neural network. Supervised learning means, the target (output) required is mentioned earlier and according to that, weight and bias values of the network are changed accordingly till exact output is reached. If it exactly matches, the amount of error will be less. The specialty of CNN is, there is no need to manually extract the features and instead raw waveforms can be given directly to the network. In general, in image processing applications, for image enhancement, convolution operation is carried out by various sizes of mask or kernel or filter (2x2 or 3x3 etc.,). The mask slides over the entire image step by step similar to scanning operation and depending on the kind of mask, pixel values are changed. Similarly, the same convolution operation is done on CNN and the objective of the Convolution Operation is to extract the high-level features such as edges, from the input image. Followed by this convolution layer, pooling layer is introduced which is responsible for reducing the spatial size of the convolved feature. By way of dimensionality reduction, it will reduce the computational power necessary to process the data. Besides, it is useful for extracting distinct features which are rotational and positional invariant, thus maintaining the process of effectively training the model. In pooling operation, max or average type can be utilized in which max pooling returns the maximum value of the portion of the image covered by the mask. While on average pooling, it generates the average of all the values from the portion of the image covered by the mask. Since max pooling also performs noise suppression, it is mostly involved with CNN. Hence our input image is converted into a suitable form for our multilayer perceptron, first image is flattened into a column vector by the flatten layer. This flattened output is fed to a feed-forward neural network and back propagation algorithm is applied to every iteration of training the model. Over a couple of epochs, the model is able to classify the inputs using the SoftMax Classification technique.

In speech processing applications, spectrogram of the input speech signal is given as input and CNN processes that as an input image. Whereas in our proposed work, MFCC features have been extracted and that is resized into 32x32 image and it is given as input to the CNN network. The proposed CNN architecture is shown in Figure 4.

In our CNN approach, it is composed of multiple hidden layers containing three convolution layer, pooling layer, batch normalization, and a flattened and fully connected layer. For convolution function, 3*3 and 7*7 filters are used and for the pooling 2*2, 7*7 is used accordingly. An additional batch normalization technique is also used followed by Rectified Linear Unit (RELU) which is used as an activation function. The output of this layer is given to the flatten layer to convert the 2D value into 1D values. Then the function of a fully connected layer is to learn non-linear combinations from the output of the convolutional layer after proper conversion by the flatten layer. Using the softmax function, the neural network, then learns to classify images through several iterations and it converts the values that are applied at the input into probabilities. No. of neurons in the output layer is equal to the no. of classifications and here we have taken as ten.

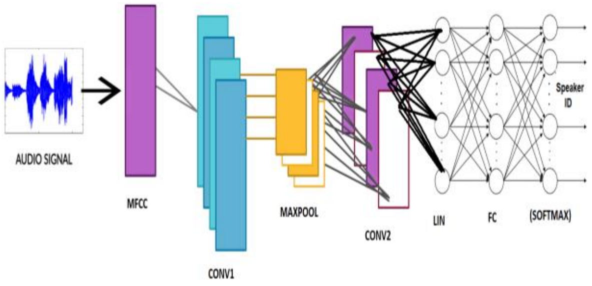


Figure 4. Structure of convolutional neural network architecture.

5. Auto Associative-Back Propagation Neural Networkarchitecture (BPNN)

Already CNN architecture has been discussed which is a variant of BPNN. In order to show how CNN performs well compared to BPNN, the speaker identification system is also developed using BPN network. This network shown in Figure 5 comprises of 3 hidden layers in addition to input and output layer [9]. No.of neurons are 1000, 70, 10 for hidden layers and 13 is considered at the input with one neuron at the output for classification. These 13 inputs are the MFCC coefficients used for training the neural network, simultaneously calculating the error at the output layer, and then modifying the weight and bias values of the network to minimize the error**.** In a multilayer feedforward network, after finding the error it is back propagated to the input to adjust the weights and biases. Hence it is called as back propagation algorithm and often it is stuck with local minimum due to the weights which are initialized randomly at the beginning. Local minimum means during training the error will be reduced and again, it is increasing after some iterations and this fluctuation in the error with many minimum values is known as local minimum. In order to get global minimum, some optimization algorithms are used.In general some of the optimization algorithms utilized to optimize the weights are gradient descent, conjugate gradients, and quasi-Newton. In our work Rprob (Resilient Backpropagation) is employed since it has much faster convergence than the standard steepest descent algorithm and also it requires less memory. Because everytime when the weights and biases are changing it has to be stored which requires increased memory requirements.

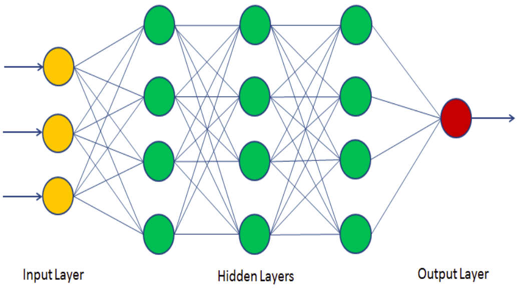


Figure 5. Architecture of back propagation neural network.

Tan sigmoid or logsigmoid is mostly used as transfer functions for hidden layers and linear transfer function is usedfor the output layer. The target for ten speakers isgiven as 1 to 10 sequentially. Input feature vectors are concatenated to form a matrix and applied to the network.

5.1. Training Algorithm for Auto Associative-BPNN

Initially random values are considered for the weights and given to all the nodes in all the layers and the target is also assigned to all the input words [10]. The features extracted are given as Input vectors to this network.

The net output of BPNN layers are determined by,

$$n^m = w^m p + b \tag{3}$$

Here *m* denotes the layer no., which is taken as 4 here.

b ,*p* and *W* are bias, input feature vector and weight matrix

Final output of each layer is given by,

$$a^m = log\,sig(n^m\,) \tag{4}$$

Then the output of last layer is found by,

$$a^m = pureline(n^m) \tag{5}$$

In order to find the performance of the network, squared error is calculated using the formula

$$e = t - a^m \tag{6}$$

Where *t* denotes the target and the error is back propagated using the sensitivity. It is calculated from the sensitivity matrix. The sensitivity of the final layer is calculated by using the formula,

$$s^m = -2F^m(n^m)(t-a) \tag{7}$$

Where *F* denotes the derivative of the transfer function of the final layer and the sensitivity for each layer is found by using,

$$s^m = F^m(n^m)(w^{m+1})^T s^{m+1} \tag{8}$$

At the same time at each individual layer, the weight value and bias values are updated by using the following equations,

$$w^m(k+1) = \gamma w^m(k) - (1-\gamma)\alpha s^m(a^{m-1})^T \tag{9}$$

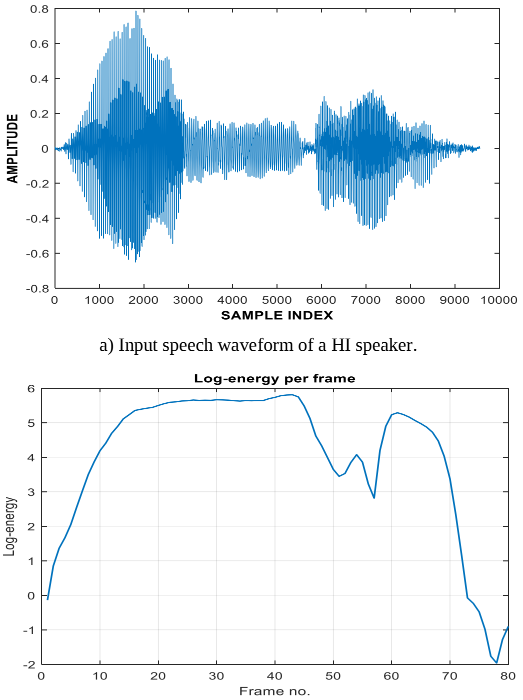
$$b^m(k+1) = \gamma b^m(k) - (1-\gamma)\alpha s^m \tag{10}$$

α, *k* and γ are learning rate, iteration no. and momentum coefficient and we can change the values according to our requirements.

6. Experimental Results and Discussion

6.1. CNN Model Results

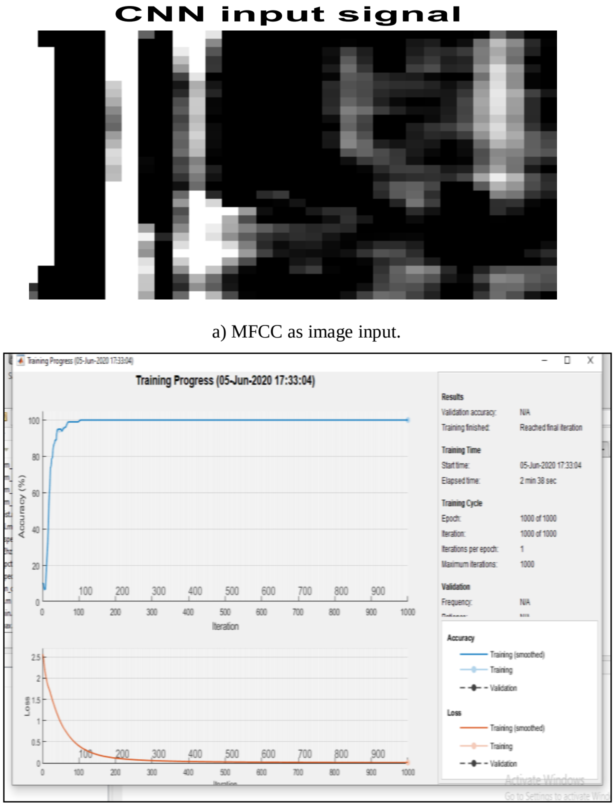
Figure 6-a) shows our sample signal for the same testing utterance “ondru” for a male speaker and Figure 6-b). shows the log energy for each frame of the test signal which has 80 frames. After extracting the MFCC features, it is resized into 32x32 image input, which is shown in Figure 7-a) and progress of training is shown in Figure 7-b).



a) Input speech waveform of a HI speaker.

b) Log energy of the Input speech waveform.

Figure 6. Speech signal and log energy of the input speech waveform.



a) MFCC as image input.

Results
Validation accuracy: N/A
Training finished: Reached final iteration

Training Time
Start time: 05-Jun-2020 17:33:04
Elapsed time: 2 min 38 sec

Training Cycle
Epoch: 1000 of 1000
Iteration: 1000 of 1000
Iterations per epoch: 1
Maximum iterations: 1000

Validation
Frequency: N/A

b) Training progress of CNN.

Figure 7. Input image and training progress of CNN.

Our database consists of 200 utterances from 10 HI speakers and among this 50% is used for training and 50% is used for testing. In order to test the network performance, first MFCC features are extracted as in training and it is simulated using the simulated model and our network produces 80% accuracy i.e., out of 100 testing utterances 80 are correctly identified. It is depicted in the Figure 8.

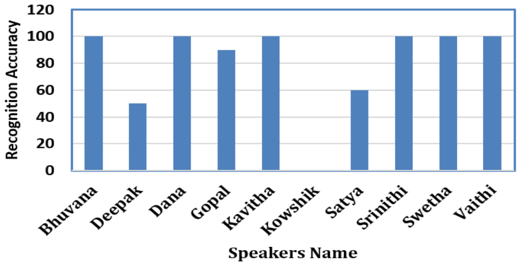


Figure 8. Performance chart of speaker identification system of HI speakers.

The confusion matrix of the recognition accuracy is given in Table 1. Though for a normal speaker, increase in the no. of classification leads to misrepresentative output, but our proposed network identified the speaker with 80% accuracy for the same utterance spoken by all of them. Here traditional features MFCC have been given instead of raw speech waveform. In future, different utterances spoken by all of the hearing-impaired speakers will be attempted. A confusion matrix is a table like format, used to visualize the performance of a system. Here it is utilized to validate the performance of the system in identifying the HI speaker. Each column of the matrix denotes the occurrences in a predicted class, while each row represents the occurrences in an actual class. It is very easy to check how the system is confusing the actual class with others. The meaning is, commonly mislabeling one as another. If one speaker has been removed, then 90% accuracy has been produced.

In our present work 10 HI speakers have been considered and 4 out of them are male children, 6 are female children. Technically speaking, all are profoundly deaf and only one female child Kavitha is Hard Of Hearing (HOH). Profoundly deaf means, the degree of deafness is high, but if they have been brought to the school earlier by identifying them as deaf, using speech therapy their speech can be improved compared to others. But though their deafness level is less for HOH children, if they are not trained earlier, their speech will be very worse [18]. In our case except the profoundly deaf male child kowshik, all were admitted in the deaf school earlier. Since he doesn’t know how the sounds will be, his speech is very bad. This is reflected in our result which is clearly seen in our confusion matrix. None of his utterance is identified by the system. From the confusion matrix, it is very clear that, except kowshik, Deepak and Satya the accuracy is 100%. Among these, recognition accuracy is 50 and 60% for Deepak and Satya respectively. If we train them properly to utter a particular sentence or word there is a chance of getting more accuracy. However, here we have trained and tested the system only for the same word in Tamil by all the children.

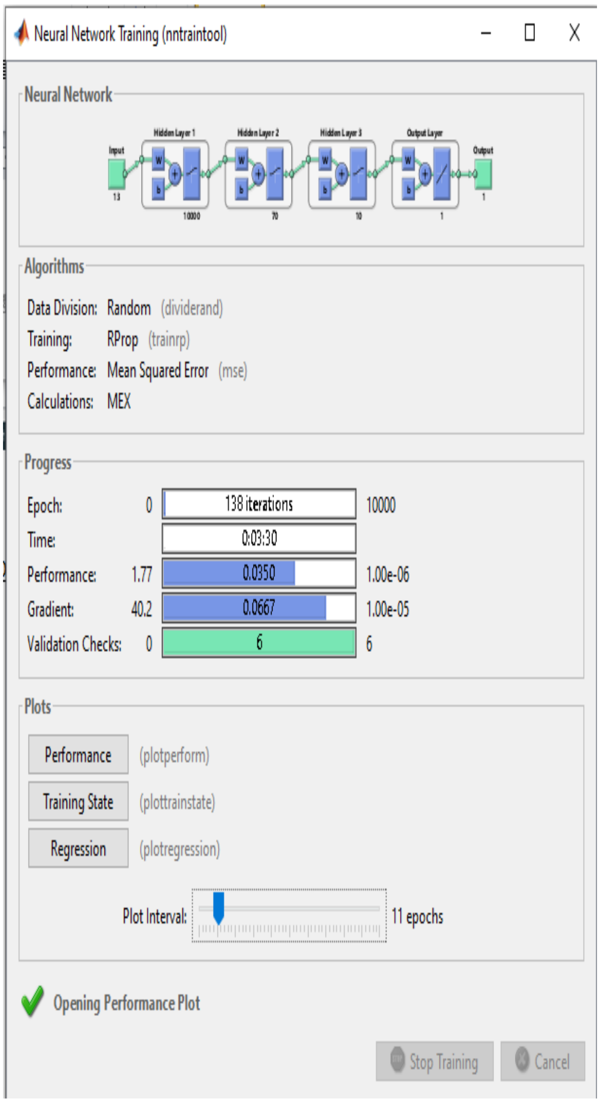
In order to verify the system performance for text independent case, we have tested the system for two different words. But most of the time it identified the speaker as speaker 10 i.e, vaithiyanathan. This is like false acceptance, may be that speaker voice is dominating, system mostly identified as vaithiyanathan. From the above results we can confirm that unlike the normal person, only text dependent speaker identification is possible for HI speakers. Because they were known only to the specific sentences what we taught, they don't know all the phonemes and syllables. It is tough for them to utter and difficult for us to teach all the sounds, how it will be.

Nevertheless, the HI speaker identification implemented using CNN and its performance is high compared with the basic Multilayer perceptron neural network using back propagation algorithm.

6.2. BPNN Results

After initializing the network parameters, architecture andfeature vectors of the input training, speech is given to the network for training. Like CNN network, here also 100 speeches are taken for training and 100 speeches are considered for testing.

The neural network model is stored for testing and now the MFCC features of test speech is simulated. The speaker is identified based on the speaker ID 1 to 10 at the output. The Figure 9 shows the training performance of the BPNN for our speaker identification task. Though the no. of neurons and training algorithms changed and verified, the performance of the network -mean squared error is not reduced below 0.129. The BPNN network simulation model is shown in Figures 9 and 10-a), 10-b) shows the performance plot of the proposed network.



Data Division: Random (dividerand)
Training: RProp (trainrp)
Performance: Mean Squared Error (mse)
Calculations: MEX

Epoch:	0	138 iterations	10000
Time:		0:03:30	
Performance:	1.77	0.0350	1.00e-06
Gradient:	40.2	0.0667	1.00e-05
Validation Checks:	0	6	6

Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

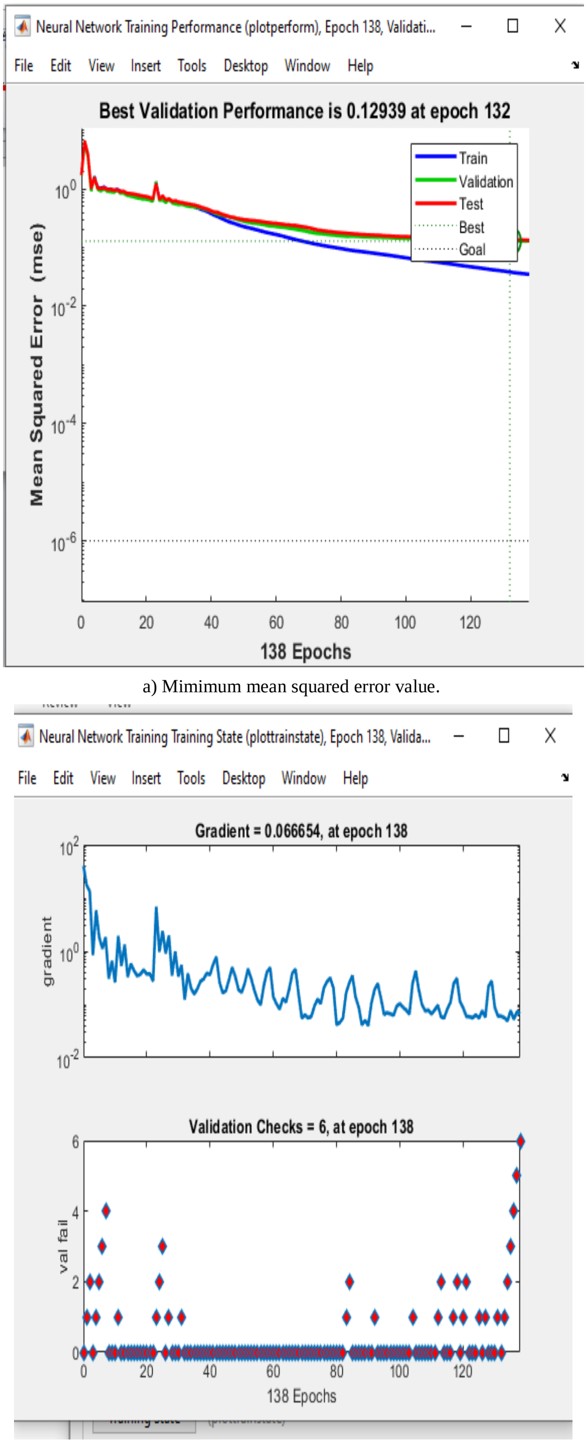
Plot Interval: 11 epochs

✓ Opening Performance Plot

Figure 9. Simulation model of BPNN.

Table 1. Confusion matrix for speaker identification accuracy for same word.

Speaker Name	Bhuvana	Deepak	Dana	Gopal	Kavitha	Kowshik	Satya	Srinithi	Swetha	Vaithi
Bhuvana	**10**									
Deepak		**5**		1						4
Dana			**10**							
Gopal				**9**						1
Kavitha					**10**					
Kowshik	10					**0**				
Satya				1			**6**		1	2
Srinithi								**10**		
Swetha									**10**	
Vaithi										**10**


Neural Network Training Performance (plotperform), Epoch 138, Validati...
File Edit View Insert Tools Desktop Window Help
Best Validation Performance is 0.12939 at epoch 132
Mean Squared Error (mse)
10^{0}
10^{-2}
10^{-4}
10^{-6}
0 20 40 60 80 100 120
138 Epochs
Train
Validation
Test
Best
Goal
Neural Network Training Training State (plottrainstate), Epoch 138, Valida...
File Edit View Insert Tools Desktop Window Help
Gradient = 0.066654, at epoch 138
gradient
10^{2}
10^{0}
10^{-2}
Validation Checks = 6, at epoch 138
val fail
0 2 4 6
0 20 40 60 80 100 120
138 Epochs

a) Mimimum mean squared error value.

b) Gradient value and validation check.

Figure 10. Training progress of BPNN.

After training, similar to CNN network, it is tested by the trained simulation model by extracting the MFCC features. The overall performance of text dependent speaker identification system is only 9%.

This is clearly depicted in Table 2 through confusion matrix.

From the results it is clear that except the first speaker, none of the speakers are not identified properly. Most of the speeches are classified as speaker 1. This shows that the network is not trained properly. Even for normal speakers, if the no. of classification is more, it is tough to classify correctly. But the accuracy is very poor using BPNN.

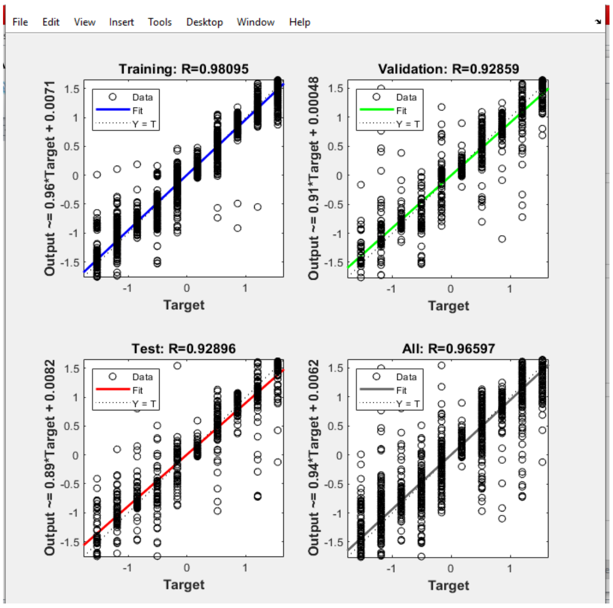


Figure 11. Regression plot of the proposed BPNN.

The database has been collected from Maharishi Vidya Mandir school for deaf in our area and utilized that for our work [18]. Though high variability among the HI speakers would affect the classification accuracy, due to reasonable performance we can utilize CNN for speaker identification or recognition purpose, even for voice or vocal impaired persons. By fine tuning the network further using some other preprocessing techniques and methods we can achieve, 100% accuracy in the future. So the proposed system finds application mainly for biometric applications [19] of HI speakers and for speaker verification or authentication.

Table 2. Confusion matrix for speaker identification accuracy using BPNN.

Speaker Name	Bhuva	Deepak	Dana	Gopal	Kavitha	Kowshik	Satya	Srinithi	Swetha	Vaithi
Bhuvana	**8**						2			
Deepak	9	**1**								
Dana	5		**0**				5			
Gopal	9	1		**0**						
Kavitha	7	3			**0**					
Kowshik		10				**0**				
Satya	9					1	**0**			
Srinithi	10							**0**		
Swetha	10								**0**	
Vaithi	10									**0**

7. Conclusions

A challenging task in HI Speaker recognition or identification is the huge variation in their speech production characteristics. Due to the inability of hearing the sounds, it is tough for them to know different sounds and to pronounce it properly. In this paper HI speaker recognition has been proposed using deep learning networks for text dependent case and 80% accuracy has been achieved using CNN with MFCC features. Though in recent papers recognition accuracy of 97.83% and 87.5% has been achieved for normal speakers, for hearing impaired speakers research is not much explored. The challenge faced is the input with unclear features as speaker is hearing impaired. Due to late intervention of deafness of a HI child, accuracy has been reduced by 10%,otherwise our system could produce 90%.Results are also validated with their speech characteristics and speech data has been collected from special school intended for HI children in our area in classical Tamil language. Also, it has been investigated and analysed that, ordinary BPNN is not suitable for identification of HI speakers. In general, voice biometrics denotes the recognition of an individual person based on their voice. Since it is unique for each individual, in this article for HI speaker it has been tried. Speaker identification systems find application in variety of different domains such as telephone banking, E-Commerce and forensics. In addition, the proposed system can also be used by the HI for accessing and controlling their smartphones and their personal virtual assistants. In future, recognition accuracy will be enhanced by incorporating suitable pre-processing techniques and dominant features for HI speech.

References

[1] Andy-Jason C. and Kumar S., "An Appraisal on Speech and Emotion Recognition Technologies based on Machine Learning," *in International Journal of Recent Technology and Engineering*, vol. 8, no. 1, pp. 2266-2276, 2020.

[2] Ahmad R., Naz S., Afzal M., Rashid S., Liwicki M., and Dengel A.,"A Deep Learning based Arabic Script Recognition System: Benchmark on KHAT," *The International Arab Journal of Information Technology*, vol. 17, no. 3, pp. 299-305, 2020.

[3] Al-Hassani R., Cagdas-Atilla D., and Aydin C., "Development of High Accuracy Classifier for the Speaker Recognition System," *Applied Bionics and Biomechanics*, vol. 2021, 2021.

[4] Ashar A., Shahid-Bhatti M., and Mushtaq U., "Speaker Identification Using a Hybrid CNN-MFCC Approach," *in Proceedings of International Conference on Emerging Trends in Smart Technologies*, Karachi, pp. 1-4, 2020.

[5] Bai Z. and Zhang Z., "Speaker Recognition Based on Deep Learning: an Overview," *Neural Networks*, vol. 140, pp. 65-99, 2021.

[6] Chowdhury A. and Ross A., "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616-1629, 2019.

[7] Dhakal P., Damacharla P., Javaid A., and Devabhaktun V., "A Near Real Time Automatic Speaker Recognition Architecture for Voice-Based User Interface," *Machine learning and Knowledge Extraction*, vol. 1, no. 1, pp. 504-520, 2019.

[8] Ganvir S. and Lal N., "Automatic Speaker Recognition using Transfer Learning Approach of Deep Learning Models," *in Proceedings of 6th International Conference on Inventive Computation Technologies*, Coimbatore, pp. 595-601, 2021.

[9] Hagan M., Demuth H., and Beale M., *Neural Network Design*, Campus Pub. Service, 2002.

[10] Https://www.datacamp.com/community/ tutorials/neural-network-models-r, Last Vested, 2022.

[11] Irum A., and Salman. A, "Speaker Verification Using Deep Neural Networks: A Review," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 20-25, 2019.

[12] Jeyalakshmi C. and Revathi A., "Efficient Speech Recognition System for Hearing Impaired Children in Classical Tamil Language," *International Journal of Biomedical Engineering*

and Technology, vol. 26, no. 1, pp. 84-100, 2018.

[13] Liang M. and Hu X., "Recurrent Convolutional Neural Network for Object Recognition," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 3367-3375, 2015.

[14] Lukic Y., Vogt C., Dürr O., and Stadelmann T., "Speaker Identification and Clustering Using Convolutional Neural Networks," *in Proceedings of IEEE 26th International Workshop on Machine Learning for Signal Processing*, Vietri sul Mare, pp. 1-6, 2016.

[15] Meftah A., Mathkour H., Kerrache S., and Ajami-Alotaibi Y., "Speaker Identification in Different Emotional States in Arabic and English," *IEEE Access*, vol. 8, pp. 60070-60083, 2020.

[16] Polur P. and Miller G., "Investigation of an HMM/ANN Hybrid Structure in Pattern Recognition Application Using Cepstral Analysis of Dysarthric (Distorted) Speech Signals," *Medical Engineering and Physics*, vol. 28, no. 8, pp. 741-748, 2006.

[17] Revathi A. and Jeyalakshmi C., "Robust Speech Recognition in Noisy Environment using Perceptual Features and Adaptive Filters," *in Proceedings of International Conference on Communication and Electronics Systems*, Coimbatore, pp. 692-696, 2018.

[18] Revathi A. and Jeyalakshmi C., "A Challenging Task in Recognizing The Speech of The Hearing Impaired Using Normal Hearing Models in Classical Tamil Language," *Journal of Engineering research*, vol. 5 no. 2, pp. 110-128, 2017.

[19] Revathi A., Jeyalakshmi C., and Thenmozhi K., "Person Authentication Using Speech As A Biometric Against Play Back Attacks," *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 1569-1582, 2019.

[20] Shi Y., Zhou J., Long Y., Li Y., and Mao H., "Addressing Text-Dependent Speaker Verification Using Singing Speech," *Applied Sciences*, vo. 9, no. 13, pp. 2636, 2019.

[21] Sidi-Yakoub M., Selouani S., Zaidi B., and Bouchair A., "Improving Dysarthric Speech Recognition Using Empirical Mode Decomposition and Convolutional Neural Network," *EURASIP Journal on Audio, Speech, and Music Processing*, no.1, pp. 1-7, 2020.

[22] Tripathi S. and Bhatnagar S., "Speaker Recognition," *in Proceedings of 3rd International Conference on Computer and Communication Technology*, Allahabad, pp. 283-287, 2012.

[23] Zhao Y., Lin X., and Hu X., "Recurrent Convolutional Neural Network for Speech Processing," *in Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing*, New Orleans, pp. 5300-5304, 2017.

[24] Zhang Z., Sun Z., Liu J., Chen J., Huo Z., and Zhang X., "Deep Recurrent Convolutional Neural Network: Improving Performance for Speech Recognition," *arXiv preprint arXiv:1611.07174*, pp. 1-10, 2016.

[25] Zhao H., Zarar S., Tashev I., and Hui- Lee C., "Convolutional-Recurrent Neural Networks for Speech Enhancement," *in Proceedings of IEEE International conference on Acoustics Speech and Signal Processing*, Calgary, pp. 2401-2405, 2018.

Jeyalakshmi Chelliah received B.E degree in Electronics and Communication Engineering from Bharathidasan University in 2002 and M.E. degree in Communication systems from Anna University, Chennai in 2008. She served as a faculty for 11 years in the Department of ECE,Trichy Engineering college, Tamilnadu. Since 2016 she has been with K.Ramakrishnan college of Engg., where she is working as Professor in ECE dept. She has obtained PhD degree from Anna University, Chennai in the field of Speech recognition of hearing-impaired people in 2015. Her research interest also includes speech processing, Image processing, Machine learning. She has published 35 papers in Reputed International journals and presented papers in more than 10 International Conferences.

KiranBala Benny Presently working as a Head of the Department, Department of Artificial Intelligence and Data Science, K.Ramakrishnan College of Engineering (Autonomous), Trichy, TamilNadu, India. He received his Bachelor degree in B.Tech Information Technology, Master Degree in M.E Computer and Communication Engineering, Management degree in M.B.A Human Resource Management and Doctorate Degree in Ph.D Computer Science and Engineering (Field of Image Processing). He has having 10 years of Teaching & Research Experience and also published more than 50 papers in peer reviewed journal.

Revathi Arunachalam has obtained B.E (ECE), M.E (Communication Systems), and Ph.D (Speech Processing) from National Institute of Technology, Tiruchirappalli, Tamilnadu, India in 1988, 1993 and 2009 respectively. She has been serving on the faculty of Electronics and Communication Engineering for 30 years and she is currently working as a Professor in the Department of ECE, SASTRA Deemed University, Thanjavur, India. She has published 40 papers in Reputed International journals and presented papers in more than 50 International Conferences. Her areas of interest include Speech processing, Signal processing, Image processing, Biometrics and Security, Communication Systems, Embedded Systems and Computer Networks.

Viswanathan Balasubramanian is currently working as Associate Professor in Department of ECE, K. Ramakrishnan College of Engineering. He completed his B.Tech degree in Electronics Engineering from Madras Institute of Technology, Anna University, his M.S and PhD from Institute of Microtechnology and Swiss Federal Institute of Technology (EPFL) respectively. After completing his PhD, he worked as a Senior Analog Design Engineer in microelectronics companies (Semtech and Kandou) working on Sigma delta ADCs, linear regulators and SAR ADCs. His broad research interests are towards, Analog/Digital integrated circuit/systems design for biomedical and wireless transceiver applications. In total, he has more than 10 years of industrial and research experience in IC design and 3 years of academic experience.