Telephone Voice Speaker Recognition Using Mel Frequency Cepstral Coefficients with Cascaded Feed Forward Neural Network

M. F. FRANKLIN NISSY[1], G. RENISHA[2]
[1] *M.E Student Government College of Engineering Tirunelveli*
[2] *Associate Professor Government College of Engineering Tirunelveli*

Abstract- ***Speaker recognition is the process of identification of the person from the characteristics of his voice. It provides service such as database access services, information services and security control for confidential information areas. However, the accurateness of speaker recognition often drops off quickly because of the low-quality speech and sound. To overcome this problem a new speaker recognition model based on Mel frequency cepstral coefficients (MFCC) are used for feature extraction. Feature extraction means that the speech signal is converted into a series of feature vector coefficients. These features only include the information needed to identify the speaker and discarding all other stuff which carries information like background noise, emotion etc. Features extracted from MFCC are given as the input to the Cascaded Feed Forward Neural Network (CFFNN) which identifies the speech signal of the corresponding speaker. MFCC is an efficient way to extract features from the signal and the Mel scale based feature extraction gives better accuracy in the clean and noisy environment.***

Indexed Terms- Speaker recognition, Mel Frequency Cepstral Coefficients (MFCC), Cascaded Feed Forward Neural Network (CFFNN), Feed Forward Neural Network (FFNN)

I. INTRODUCTION

Speaker recognition is the process of identifying the person who is speaking by the uniqueness of his/her voice (voice biometrics). This is also called as voice recognition. Human speech conveys different types of information. It carries information like gender and also carries other information like identity of the speaker and the emotional state of the speaker. Speaker recognition help to recognize an unknown speaker among the set of reference speakers using speaker specific information present in their speech waves. The goal of speaker recognition is therefore to extract this speaker-specific information and use them for identification purposes.

Speaker recognition is a branch of biometric authentication. It focuses on security system of controlling the access to secure data or information from being accessed by anyone. Speaker detection is the process of using the voice of spokesperson to verify their uniqueness and control access to services such as voice dialing, mobile banking, database access services, voice mail or security organize to a secured system.

In telephone, the usable voice frequency band ranges from approximately 300Hz to 3.4 KHz. It is for this reason that the ultra-low frequency band of the electromagnetic spectrum between 300Hz and 3 KHz is also referred to as tone frequency. The bandwidth owed for a single voice broadcast channel is usually 4 KHz, together with guard bands, allowing a sampling rate of 8 KHz to be used as the basis of the pulse code modulation. By Nyquist sampling theorem, the sampling frequency (8 KHz) must be at least double the highest component of the voice frequency.

The bandwidth of a signal depends on the quantity of information contain in it and the quality of it. The range of frequencies necessary for an analog voice signal, with a fixed telephone line quality (recognizable speaker), is 300Hz – 3.4 KHz. This means that the bandwidth of the signal is 3.1 KHz. A human voice contains much higher frequencies, but this bandwidth gives a good compromise between the quality of the signal and the bandwidth. Bandwidth, jointly with noise, is the major factor that determines

the information-carrying ability of a telecommunications channel.

In order to allow more long-distance calls to be transmitted, the frequencies transmitted are restricted to a bandwidth of about 3 KHz. All of the frequencies in telephone tone below 400 Hz and above 3.4 KHz are eliminated. The person will be able to hear 1 KHz sound clearly. The person will also be able to have the sense of hearing the 2 KHz and 3 KHz tones. However, the person will have problem on hearing the 4 KHz tone, and will not able hear the 5 KHz or 6 KHz tones.

II. RELATED WORK

Feature extraction is the process of extracting the features from the speech signal. There are various feature withdrawal technique such as Linear Prediction Coding (LPC) and Linear Predictive Cepstral Coefficients (LPCC). LPC removes the effects of formants from the speech signal and estimates the intensity and frequency of the buzz [1]. Drawback of this technique is that performance degradation in occurrence of noise. LPCC techniques gives smoother spectral envelop and stable representation as compare to LPC [2]. The drawback of this technique is that linearly spaced frequency band. In classification stage the patterns are classified into dissimilar classes. There are many classifiers are used such as DTW, VQ, etc. Dynamic Time Warping algorithm calculates the distance between two sequences which may vary in time or speed. Then time normalization distance is calculated between patterns. Genuine speaker is recognized with minimum time normalized distance. It requires less storage space [3]. In vector quantization the extracted speech feature of spokesman are quantized to a number of centroids. These centroids compose the codebook of that speaker. It is used for data compression and requires less storage [4]. VQ is computationally less complex.

III. PROPOSED SYSTEM

A. MFCC Based Feature Extraction

A Mel frequency cepstral coefficient is the proposed system used to extract useful features from the speech signal. MFCC give better removal compared to LPC and LPCC. It is the fastest and best method. The MFCC characteristics are rooted in the recognized discrepancy of critical bandwidths of the human ear with frequency filters spaced linearly at low frequencies and used logarithmically at high frequenciesto retain the phonetic vital characteristics of a speech signal. The Mel-frequency scale has a linear spacing of frequencies under 1000 Hz and a logarithmic spacing above 1000 Hz. Tonal pitch 1 kHz and 40 dB above the perceptual audible threshold is defined as 1000 mels, and used as reference point. MFCC is based on signal disintegration with the help of a filter bank.

a. Digitization

Analog speech signal is segment into small frames by sampling. The sampling is done to obtain an amplitude value from the sound waveform at certain interval of time. It convert continuous signal to discrete signal. According to Nyquist sampling theorem, the sampling frequency is usually between 15 KHz to 20 KHz.

b. Removing silence

In original speech signal there is a silent portion which doesn't contain any useful information regarding to the speaker. This silent portion is removed in order to get better accuracy rate. Silence removal also compresses the signal. The MFCC block diagram is shown in the following Fig-1

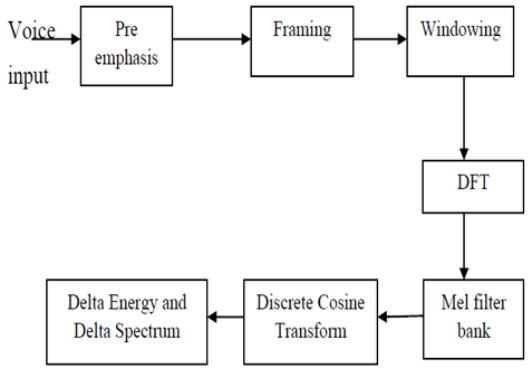

Fig-1 Block diagram of proposed MFCC Method

- STEPS IN MFCC

1. Pre-emphasis

It is the first step in Mel frequency cepstral coefficient extraction. High frequency signal are distorted more by noise than low frequency signal. Preemphasis only boosts components of high frequen cy, while leaving the signal components of low frequ

ency in their original state. Hence high frequency signal contain more useful information that are necessary to recognize the original speaker.Pre-emphasis works by enhancing the energy of the high frequency each time a data transfer takes place.

2. Framing

In framing the signal is divided into small fragment because speech signal changes over time. When it is fragment into small frames then it is assumed to be constant throughout the frame for simplicity purpose. If the frame is long then it changes over time. In order to make it constant, framing is done in the signal.

3. Windowing

Windowing technique is usedto eliminate discontinuit y of the signal. It minimizes spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If the window is defined as w(n), 0≤n≤N-1. Where 'N' is the number of samples in each frame then the signal is the product of windowing.

$Y_1(n) = X_1(n)\, W(n) \quad 0 \le n \le N-1$ (1)

4. Fast Fourier Transform

Fast Fourier Transform (FFT) is used to convert each frame of N samples from time domain to frequency domain. FFT is performed to obtain magnitude of frequency response of each frame. Which is defined on the set of N samples $\{x_n\}$, as follow

$X_k=\sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}$ k=0,1,2,…N-1 (2)

After this the output is often referred to as spectrum or periodogram.

5. Mel Filter Bank Processing

Pitch in the speech signal is measured on a scale called the 'Mel' scale. The Mel-frequency scale is a linear spacing of less than 1 KHz and a logarithmic spacing above 1 KHz. The output of FFT is multiplied by a set of 20 triangular band pass filter to get log energy of each triangular band pass filter. The Mel filter bank has a triangular band pass frequency response, and a constant Mel frequency interval defines the spacing as well as the bandwidth. The number of Mel spectrum coefficients, is typically chosen as 20.

6. Discrete Cosine Transform

We will be converting the log Mel spectrum back into time in the final step. The result is called Cepstrum coefficients (MFCC) for the Mel frequency. For the specified frame study the cepstral representation of the speech spectrum provides a good description of the local spectral properties of the signal. Because the Mel spectrum coefficients are real numbers we can use Discrete Cosine Transform (DCT) to convert them to the time domain. Therefore Mel power spectrum coefficients are

$\tilde{S}_0$, k = 0, 2… k-1 (3)

The MFCC can be calculated, $\tilde{c}_n$ as

$\tilde{c}_n=\sum_{k=1}^{k}(log\tilde{s}_k)\cos[\, n(k-\frac{1}{2})\frac{\pi}{k}]$ (4)

The first component, $\tilde{c}_0$ excluded, since it represents the mean value of the input signal, which carried little speaker specific information.

7. Delta Energy and Delta Spectrum

The first order derivative of cepstral coefficients is called Delta coefficients, and hereby the second order derivative of cepstral coefficients is called Delta-Delta coefficients. Coefficients of Delta tell about the speaking rate and coefficients of Delta-Delta are similar to speaking speed.

The extracted features are given as the input to the Feed Forward Neural Network (FFNN) and cascaded Feed Forward Neural Network (CFFNN) for training and testing.

B. Feature Classification Using Neural Networks

Classification is the process of identifying unknown speakers using a classifier to match their feature with existing databases. In this paper artificial neural network is used as a classifier. Artificial Neural Network (ANN) works in the same way like human brain and consists of different neurons which are used to carry message from one layer to other. Artificial Neural Network mainly consists of three layers-Input layer, hidden layer and output layer. The network has varying neurons input n, which receive input of different sets features. The number of hidden layer

varies from 1 to 4 and neurons in each hidden layer vary from 10 to 60.

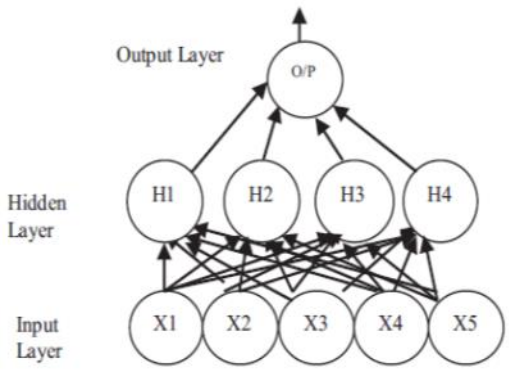

Fig-2 Artificial neural network

a. Input Layer

This is the first component of a network of neurons. It's used to provide the network with the input data or features.

b. Output Layer

This layer gives out the predictions. The output layer represents the speakers for voice recognition. Various output layer size is used with different task.

c. Hidden Layer

It computes highly complex functions by this cascading effect. The concealed unit most commonly used is the one that uses a Rectified Linear Unit (ReLU) as the activation function. Rectified Linear Unit allows only positive values to pass through it. The negative values are mapped to zero. Based on the arrangement or connection there are different types of neural network such as FFNN and CFFNN.

1. Feed Forward Neural Networks

Feed Forward Neural Networks (FFNN) consists of a series of layers. The first layer has a link from the network input. Each subsequent layer has a correlation from the previous layer. The last layer produce the network's output.

2. Cascade Feed Forward Neural Networks

Cascade Feed Forward Neural Networks (CFFNN) is similar to Feed Forward Networks, but includes a connection from the input and every previous layer to following layers. A CFFNN consists of number of layers and all layers take its input from all preceding layers. Hence this network is more multifaceted in terms of interconnection. In CFFNN all inputs and outputs of all preceding layers are included in the input to any layer. This results in a cascaded interconnection between layers leading to more compact structures. While the FFNN consists of 70 number of layers and each layer takes its input as the output of previous layer and signal flows in one direction only, the CFFNN architecture include a weight connection from the input to each layer and from each layer to the successive layers. The commonly used hyperbolic tangent sigmoid activation function is used for all hidden layers while pure-linear function is used for output layer.

- Backpropagation Algorithm

The gradient descent method involves determining the loss function derivative with respect to the network weights. This is usually done using backpropagation. Considering one output neuron, the feature square error is

$E = L(t,y)$

Where

t is the destination output for a training sample,

y is the real output of the output neuron and

E is the thrashing for the output y and end value t

For each neuron j, the output o_j is set as

$$o_j=\varphi(net_j)= \varphi(\sum_{k=1}^{n} wk_j o_k) \quad (5)$$

Where the activation function φ is nonlinear and differentiable (even if the ReLU is not in one point). The logistic function is a historically used activation function:

$$\varphi(z) = \frac{1}{1+e^{-z}} \quad (6)$$

This has a convenient derivative of:

$$\frac{d\varphi(z)}{dz} = \varphi(z)(1- \varphi(z)) \quad (7)$$

The input net_j to a neuron is the weighted sum of outputs o_k of before neurons. If the neuron is in the initial layer after the input layer, the o_k of the input layer are simply the inputs x_k to the network. The number of neuron input units is n. The variable w_{kj} denotes the weight between neuron k of the earlier layer and neuron j of the present layer.

IV. RESULTS AND DISCUSSION

All signals that are transmitted through telephone consist of multiple frequencies. The frequency range which a signal occupies is called the signal bandwidth. The bandwidth is Hertz (Hz) measured. The audio was recorded in a room by using the mobile phone. The system is tested a small speech database that consists of five speakers. We sampled the analog speech of 8 KHz rate, and then pre-emphasized the digital speech signals. The first step is to record the voice of five speakers with each speaker pronounces a predefined word pattern by using the mobile phone. The recording contains alphabet (A to Z) word pattern for each speaker. Word pattern of each speaker is stored and then extract the features by using MFCC method. All speakers word pattern are used in a training stage and testing stage. The performance of the speaker recognition system depends on the amount of test data and data train.

A. Without Silence Removal

The speech signals frequently contain a lot of areas of silence or noise. That area will not contain any useful information. Silence signal will make the processing signal larger and take more time and space when getting information or features from the signal. Hence the accuracy of the signal gets reduced compared to silence removal signal. The below two Fig-3& 4 shown the speech signal spoken by a girl among one of five speakers.

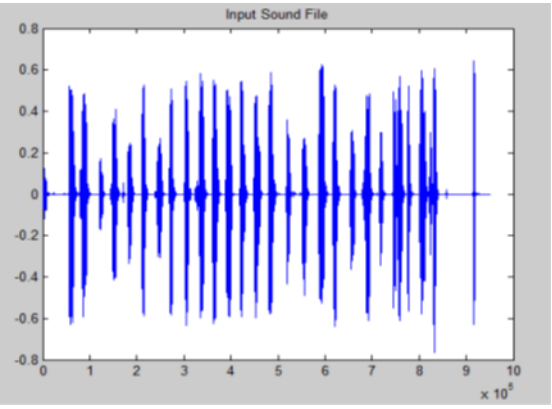

Fig-3 Speech signal with Silence

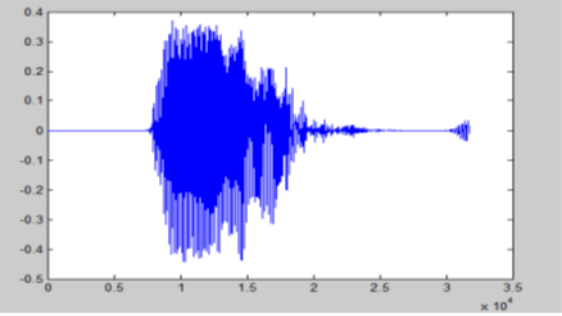

Fig-4 Framed signal

The speech signal is spoken by a female of age 23. The word spoken by them is ‘a’.

Table-1: Feed forward and cascaded MFCC values for three speakers without silence removal

Number of Speakers	FFNN	CFNN
3	12.32	40.09

From the above table it is observed that accuracy of Feedforward neural networks is very low compared to Cascade-forward neural networks.

B. Silence Removal

Silence removal is the process of removing silent portion from the speech signal. In speech analysis it is needed to first apply a silence removal method, in order to identify clean speech segments. The below two Fig-5& 6 shows the speech signal after silence removal.

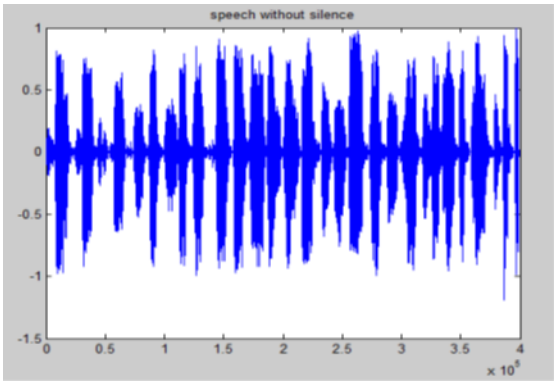

Fig-5 Speech Signal without Silence

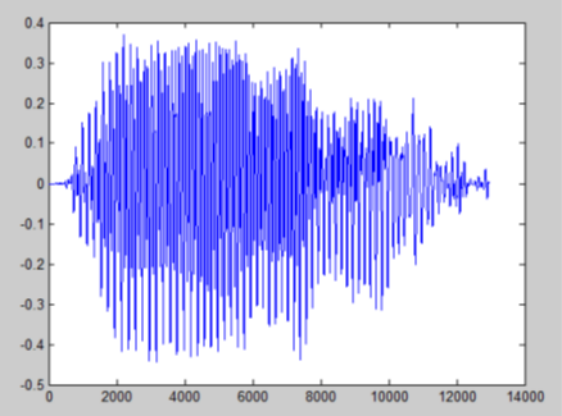

Fig-6 Silence removed speech signal frame

After framing, windowing is applied followed by fast Fourier transform this is shown in below figure 7& 8.

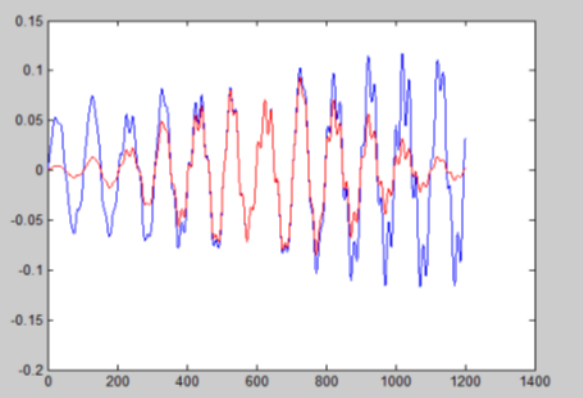

Fig-7 Windowing frame

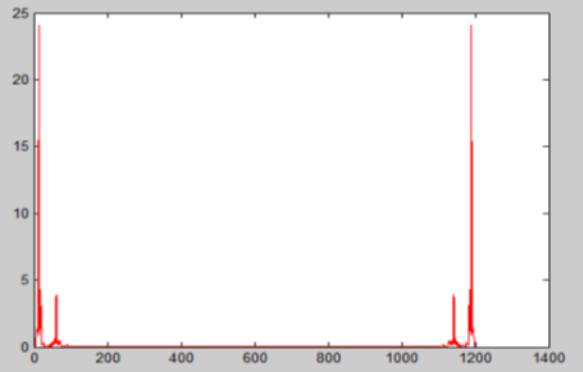

Fig-8 Frame in Frequency domain

FFNN and CFFNN accuracy value when five speakers are speaking after silence removal is shown in the table

Table-2: Feedforward and cascaded MFCC values for five speakers

Number of speakers	FFNN	CFFNN
5	48	52.3

From table 1 & 2 it is observed that after silence removal accuracy value for both FFNN and CFFNN get increased.

C. Delta and Delta-Delta Coefficients

It is the first and second order derivative of cepstral coefficient. FFNN and CFFNN MFCC delta accuracy value when five and three speakers are speaking after silence removal is shown in the table 1.3&1.4

Table-3: Feedforward and cascaded MFCC delta values for five speakers

Number of speakers	FFNN	CFFNN
5	22.76	51.23

Table-4: Feedforward and cascaded MFCC delta values for three speakers

Number of speakers	FFNN	CFFNN
3	76.54	78.43

From the above tables it is observed that when the number of speakers get decreases then the accuracy value get increased. This is shown in the below bar chart Fig-9

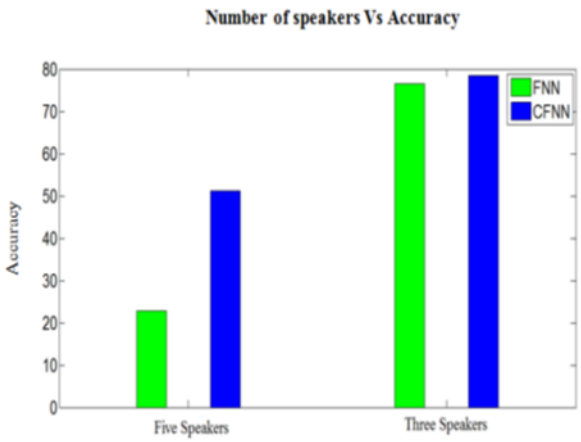

Fig-9 Number of speakers Vs accuracy

After the silence removal the accuracy value get increase this shown in the below bar chart.

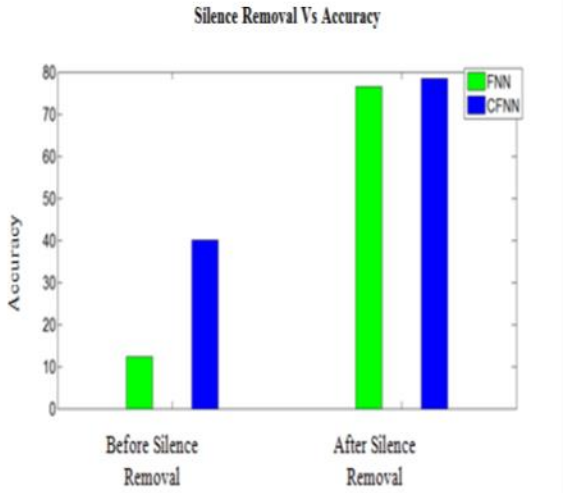

Fig-10 Silence removal Vs accuracy

V. CONCLUSION

In this paper a new speaker recognition model was proposed known as Mel frequency cepstral coefficient. MFCC is used to extract features from the speech signal. The extracted features are used for training and testing the feedforward neural network and cascaded feedforward neural network. When comparing the output of FFNN and CFFNN, CFFNN yield better accuracy value compared to FFNN around 78.43%. Thus the recognition performance accuracy is improved by MFCC. Hence the proposed method is superior to conventional methods from experimental data.

REFERENCE

[1] Chougala, M. and Kuntoji, S., 2016, March. Novel text independent speaker recognition using LPC based formants. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 510-513). IEEE.

[2] Li, P., Hu, F., Li, Y. and Xu, Y., 2014, July. Speaker identification using linear predictive cepstral coefficients and general regression neural network. In Proceedings of the 33rd Chinese Control Conference (pp. 4952-4956). IEEE.

[3] 3.Nair, R. and Salam, N., 2014, December. A reliable speaker verification system based on LPCC and DTW. In 2014 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-4). IEEE.

[4] 4.Ilyas, M.Z., Samad, S.A., Hussain, A. and Ishak, K.A., 2007. Speaker verification using vector quantization and hidden Markov model. In 2007 5th Student Conference on Research and Development (pp. 1-5). IEEE.

[5] 5. Bansal, P., Imam, S.A. and Bharti, R., 2015, October. Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy. In 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI) (pp. 41-44). IEEE.

[6] 6. Chauhan, N. and Chandra, M., 2017, March. Speaker recognition and verification using artificial neural network. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 1147-1149). IEEE.

[7] Chauhan, N., Isshiki, T. and Li, D., 2019, February. Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 130-133). IEEE.

[8] Weng, Z., Li, L. and Guo, D., 2010, July. Speaker recognition using weighted dynamic MFCC based on GMM. In 2010 International Conference on Anti-Counterfeiting, Security and Identification (pp. 285-288). IEEE.

[9] Sukhwal, A. and Kumar, M., 2015, October. Comparative study between different classifiers based speaker recognition system using MFCC for noisy environment. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 955-960). IEEE.

[10] Revathy, A., Shanmugapriya, P. and Mohan, V., 2015, March. Performance comparison of speaker and emotion recognition. In 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN) (pp. 1-6). IEEE.