Automatic Speaker Recognition using MFCC and Artificial Neural Network

Kharibam Jilenkumari Devi, Ayekpam Alice Devi, Khelchandra Thongam

Abstract*: Automatic speaker recognition is the process of identification of a person automatically from his/her voices. A robust feature extraction algorithm is required for effective and efficient classification. In this paper, a new method is proposed for identifying the speaker using an artificial neural network. Here mel- frequency cepstral coefficient(MFCC) is used as a feature extraction technique that provides useful features for the recognition process. Using these extracted features value, input samples are then created and finally, classification is performed using Multilayer Perceptron (MLP) which is trained by backpropagation. This proposed method gives an accuracy of 94.44%.

***Keywords*: Speech Signal, Mel-frequency Cepstral Coefficient, Feedforward Neural Network, Back-propagation.**

I. INTRODUCTION

Automatic Speaker Recognition in the field of study that deals with digital signal processing correlated to the recognition of the people based on their voice. Voice/speech signal is the most commonly used biometric technique for authenticating and monitoring of Identity. The objective for speaker recognition is to recognize the speaker through the method of processing, characterizing and identifying the data contained in the voice signal. Speaker recognition comprises of speaker identification and speaker verification [1]. Speaker identification is identifying who is speaking from a set of registered speakers whereas speaker verification is the process of accepting and rejecting the claimed identity. Speaker identification is classified into a text-dependent and text-independent system. In text-dependent, the speaker is speaking on a predefined or prompted text transcription and, a similar text is spoken on both training and testing phases. In text-independent, there is no limitation on the text being spoken, the test speaker is asked to utter any words or phases of his own and, could be different for training and testing phases. This work focuses on the speaker-independent system. Best matches of voices of the input voice sample are found in speaker identification. Enrolment and verification stage are the stages of speaker identification [2]. Feature extraction and pattern classification are involved in this process. There are different methods of feature extraction techniques such as MFCC (Mel Frequency Cepstral Coefficients) LPC (Linear prediction coefficient), PLP (Perceptual Linear Prediction) [3], PNCC (Power Normalized Cepstral Coefficients) [4], wavelet decomposition and Transform domain such as Discrete Wavelet Transform(DWT) and Daubechies wavelets(mother wavelet) of speech signal[5]. So far MFCC is the most frequently used feature extraction technique as compared to other techniques as it is less complex in implementation, more effective and robust under various conditions [6]. In this paper, we used the MFCC technique. These extracted features are used for every speaker to create a model and stored into a database so that it can perform a comparison during testing. There are several methodologies of recognition of a speaker, which are an improved i-vector[7], GMM(Gaussian Mixture Models)[8], HMM(Hidden Markov Models), etc. In this work, recognition is perform using MLP with feed-forward backpropagation. This proposed method is better to classify the speech signal by comparing it with existing state-of-art methods. The block diagram of the proposed method is shown in fig. 1. The paper is organized as follows: Section II explains the literature review. Section III and IV give a description of the database and the Proposed method. Recognition using MLP feedforward backpropagation is explained in section V. Finally Result and Analysis is shown in section VI.

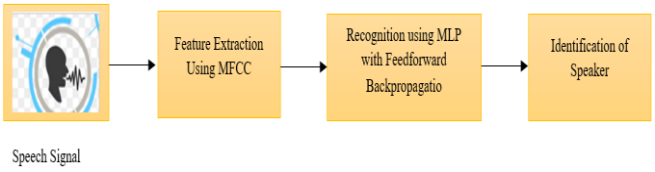

Fig.1: Block diagram of the proposed method

II. LITERATURE REVIEW

This chapter shows the overview of existing and other related work on speaker recognition.

Brucal et al[9] used MFCC to extract voice features using MATLAB VOICEBOX toolbox. The extracted MFCCs were used as inputs to the multilayer Artificial Neural Networks (ANN) for the female voice recognition algorithm. This study explored the recognition performance of the neural networks using the variable, number of hidden neurons and layers, and determine the architecture that would provide the optimum performance in terms of high recognition rate.

Sharma et al[10] proposed MFCC, PLP and neural networks-based Speech Recognition System for Hindi language and also tested against different neural network techniques. Using MFCC and PLP for feature extraction with feedforward BPN for training and testing gives 79% and it was observed that the method was giving better accuracies as compared to other conventional methods like HMM and SVM.

Revised Manuscript Received on November 22, 2019.
* Correspondence Author
Kharibam Jilenkumari Devi, Department of Electronics and Communication Engineering, National Institute Of Technology Manipur, Imphal, India. Email: jilenkumari@gmail.com.
Ayekpam Alice Devi, Department of Computer Science and Engineering, National Institute of Technology Manipur, Imphal, India. Email: ayekpamalice4@gmail.com
Dr. Khelchandra Thongam Department of Computer Science and Engineering, National Institute of Technology Manipur, Imphal, India. Email: thongam@gmail.com

Shen et al [11] used the technique of latent factor analysis (LFA) to study the channel factors in the speaker's Gaussian Mixture Model (GMM). In the endpoint detection phase of speaker recognition, GMM speech modeling is introduced to sort out the beginning and ending points of the speech segment. The Factor Analysis technique is used to fit the differences between the speaker's characteristics of space and channel space. This removes the channel factor in the speaker's GMM. The upper vectors of GMM are extracted as the input of the Support Vector Machine (SVM) to get recognition results. But computing power decreases while the large dataset is used.

Pawar et al[12] discussed the text-dependent speaker identification that detects a specific speaker from a known population. The feature extraction is performed by means of LPC coefficients, AMDF, and DFT calculation. By implementing these characteristics as input parameters, the neural network is trained. The findings achieved show that in different cases the modification in the outcomes is negligible for the same speaker talk. The software operates well to identify speakers from various speakers.

Doungpaisan et al [13] introduced an algorithm for language and text-independent speaker recognition systems on the basis of fuzzy logic and ANNs are evaluated. MFCCs are extracted for all the speakers. Then, ANN for a particular speaker is trained and this process is repeated for all the speakers one by one to get a feature matrix of the same size and then the same is done with the Fuzzy Logic Technique. The database used has addressed only noise and channel mismatch problems but still, there are several problems.

Pawar et al [14] proposed a new method of feature extraction using speaker pitch, stationary Wavelet, and multilayered Neural Networks. Implementation of a performance test is done with the recorded database for text-dependent and text-independent. Stationary wavelet with the multilayered neural network showed better accuracy and faster identification time in comparison with traditional MFCC, discrete, and continuous wavelet transform approaches.

III. DATABASE

The choosing of the database for implementing any proposed method is one of the most important steps to be taken by the researchers. In this paper, we used a database of Indian scenario named as IITG Multivariability Speaker Recognition Database [15]. The database is classified into Part-I, Part-II, Part-III respectively. In this work, Part-III of IITG-MV phase-IV speaker recognition is used for the purpose, the subject was asked to read some text.

IV. PROPOSED METHODOLOGY

A. Feature extraction using MFCC.

Feature extraction is the process of extracting the features from the input speech signals for identifying the speaker and also the process of computing a series of features for each short –time frame of the input speech signals, by considering a small segment of speech is adequately stationary to develop significant modeling. In this paper, the extraction of features is performed by using the mel frequency cepstral coefficient as the computation of the MFCC method is based on short-time power.

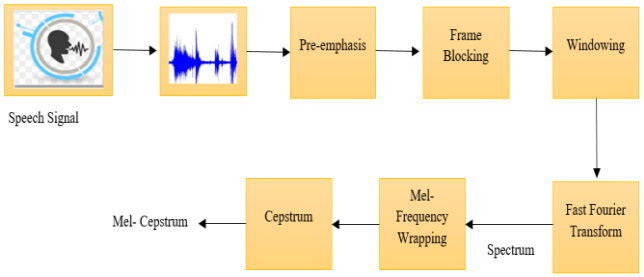

Fig.2: steps of MFCC Feature Extraction.

The spectrum produced from the human vocal tract and it also maps the known variation of the human ear's critical bandwidth frequencies with two filters which are linearly spacing at low frequencies below 1kHz and logarithmically at high frequencies above 1k Hz to capture the important characteristics of speech. The steps of MFCC feature extraction is shown in the above fig. 2.

- **Pre-emphasis**

It is required to compensate for the high-frequency part that was suppressed during the sound production mechanism of humans by passing through a filter. The speech signal is pass to high pass filter as

$$x_1(n) = x(n) - \alpha * x(n-1) \quad (1)$$

Where $x_1(n)$ represents the output signal $x(n)$ and $x(n-1)$ denotes present and past signal respectively. The value α lies between 0.9 to 1.

- **Frame Blocking**

It splits the continuous speech signal into small frames of N samples, with adjacent frames are separated by M samples (M<N) and are overlapping by N-M samples. This process continues until the whole signal is broken into small frames.

- **Windowing**

Windowing reduces spectral distortion by tapering the signal to zero at both the beginning and end of each frame. The extracted signal is obtained by multiplying signal $x(n)$ with a window $w(n)$ at time n, is represented by

$$y_2(n) = x(n) * w(n), \quad 0 \le n \le N-1 \quad (2)$$

Where N represents the number of samples in each frame. Here Hamming window is used as it minimizes the frequency resolution of spectral analysis while sinking sidelobe level of window transfer, is represented by

$$w(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right], \ 0 \le n \le N-1 \quad (3)$$

- **Fast Fourier Transform**

It transforms the N number of samples from the time domain to the frequency domain. FFT is the commonly used algorithm to implement Discrete Fourier Transform (DFT) which is defined on a set of N Samples $\{y_n\}$, as

$$Y_n = \sum_{n=0}^{N-1} y_n e^{-j2\pi kn/N}, \quad k = 0,1,2,\dots\dots\dots,N-1 \quad (4)$$

The result thus obtained from FFT is referred to as a spectrum or periodogram.

- **Mel-frequency wrapping**

Mel frequency is mainly based on the study of frequency perceived by a human. Human hearing shows differential sensitivity to all frequency bands. It is less sensitive at higher frequencies above 1000 Hz. and Mel-frequency is linear frequency spacing below 1 kHz and the speech signal is given as

$$Mel(f) = 2595 * log_{10}(1 + f/700) \quad (5)$$

- **Cepstrum**

This is the final step of the MFCC process, in which the log mel spectrum is transformed back to the time domain and this conversion is commonly done by DCT as its output contains an important amount of energy. The result obtained from DCT is known as Mel Frequency Cepstral coefficient(MFCC) and is represented by,

$$C[n] = \sum_{n=0}^{N-1} log \left| \sum_{n=0}^{N-1} x(n) \exp(\frac{-j2\pi kn}{N}) \right| \exp(\frac{j2\pi kn}{N}) \quad (6)$$

Where n= 0,1,2,…. N-1. $C[n]$ denotes the MFCC and n represent the number of coefficients, n=12, thus 12 cepstral coefficients are extracted from each frame.

B. Recognition with MLP Feedforward backpropagation

The features thus obtained from MFCC will be used as input for recognition through our approach Multilayer perceptron(MLP) feedforward neural network. Fig. 3 shows the structure of the Multi-layer FNN (feedforward neural network). A feedforward network (FNN) is an artificial neural network that does not form a cycle in relation to the nodes of an *input layer*, one or more *hidden layers*, and an *output layer*. The learning of the Multilayer perceptron Feed-forward neural network is carried out by using the backpropagation algorithm. The network needs to offer the output called the target for a specific input to train the network. The network is initialized first by a small random value called weights. Once the network is trained, any of the input patterns will be produced as output and it is called as forward propagate. And there is an error in output and target. We propagated backward to decrease this error and raised the value of weight. The backpropagation algorithm uses a method called the delta rule or gradient descent to search for the minimum value of the error function in weight space. In other words, the output of each neuron will be nearer to its target. The following are the steps of the backpropagation algorithm.

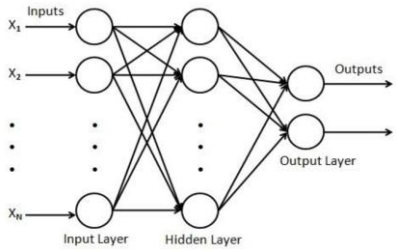

Fig. 3: Structure of multi-layer Feedforward Neural network

- Identify the number of layers and the number of neurons

The neuron of hidden layers, input and output say

X_i: The i[th] input

Y_j: The output of the j[th] hidden neuron

Z_k: The output of the j[th] output neuron

- Initialize all weights and biases into small random say Wn=weight and 1=biases.
- Calculate the actual output as

$$Y_j = \Sigma(X_i * W_n) + 1 \quad (7)$$

$$Z_k = f\Sigma(X_i * W_n) + 1 \quad (8)$$

Where f is the activation function

- Calculate the error 'E' between actual output and target output Tκ.

$$E = \sum \frac{1}{2(T_k - Z_k)^2} \quad (9)$$

- Now propagate backward to reduce the error i.e. update the value of weights and biases and repeat above step3 and step4.

$$\frac{\partial E}{\partial W_n} = \frac{\partial E}{\partial Z_k}\frac{\partial Z_k}{\partial W_n} = \frac{\partial E}{\partial Z_k}\frac{\partial Z_k}{\partial net_k}\frac{\partial net_k}{\partial W_n} \quad (10)$$

$$\frac{\partial Z_k}{\partial net_k} = \frac{\partial}{\partial net_k} f(net_k) = f(net_k)(1 - f(net_k)) \quad (11)$$

$$\frac{\partial net_k}{\partial W_n} = \frac{\partial}{\partial W_n}(W_n X_i) = Z_k \quad (12)$$

$$f(net_k) = f(\Sigma W_n Z_k) \quad (13)$$

$$W_n{}^+ = W_n - \eta \frac{\partial E}{\partial W_n} \quad (14)$$

Where η is the learning rate.

- If the error is smaller that means the output is nearer to target output, then stop.

V. Results and Discussions

In this proposed method simulation is done in Matlab 2018a. Firstly, the input speech signal is given to the system and then MFCC 12 coefficients for each sample of speech signals are extracted. Finally extracted features are applied to the MLP feed-forward neural network. The backpropagation algorithm performs the learning on this network(FNN). The MLP network will have 36 outputs because of 36 speakers. The output is represented by binary string (0 and 1) of 36 bits. For speaker 1, the first bit of 36 bits is 1 and other bits are 0 and for speaker 2, the second bit is 1 and others are 0 and so on.

The network which has been trained with 156 samples for 36 speakers at 10000 epoch, the error drops to 0.00243 as shown in fig.4. and the network that has been tested with 36 samples, 34 samples are predicted correctly and two speakers are not predicted clearly as any of the speakers. This system achieves an accuracy of 94.4 %. Fig. 5 shows the neural network regression plot of training, testing, validation. Fig 6 shows the neural network performance of the system and the best validation performance obtained is 0.022032 at 10000 epochs.

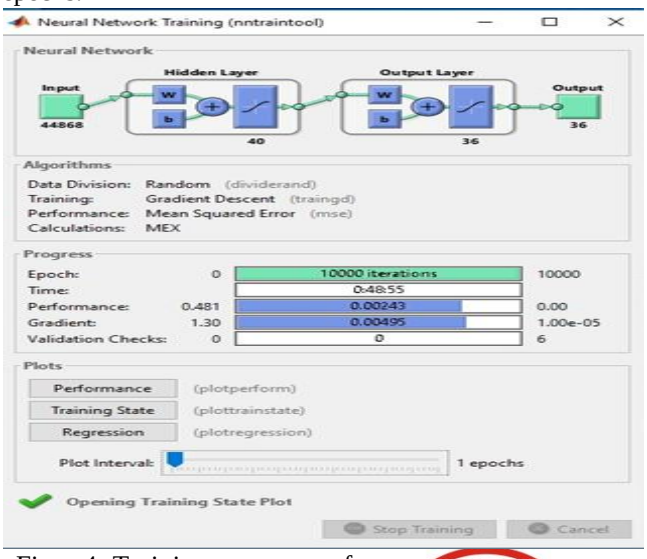

Fig.: 4 Training progress of MLP at 10000 epochs

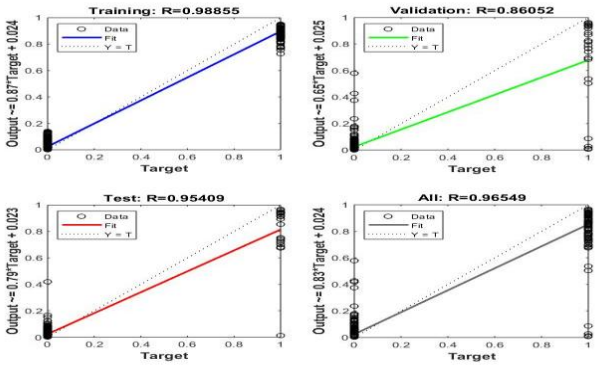

Fig.: 5 Regression plot of neural network

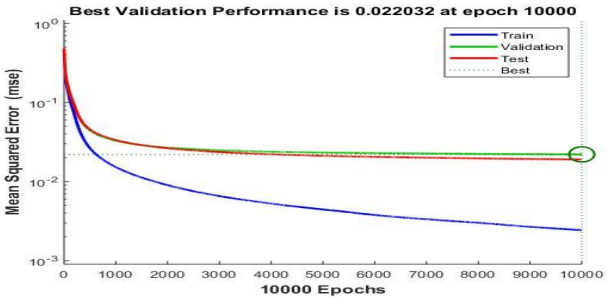

Fig.: 6 Validation performance.

Table- I: Comparison with other methods

Sl. No.	Methodology	Accuracy
1.	MFCC+PLP+ neural network[10]	79%
2.	MFCC + neural network [16]	77%
3.	MFCC+MLP-Feedforward back propagation[**proposed**]	**94.44%**

Table I shows the comparison of our proposed method(MFCC+MLP-feedforward backpropagation) with other state of-art-methods [10],[16]. It has been found that the best performance /Recognition rate is achieved by the proposed method in this comparison.

VI. CONCLUSION

In this work, a method of automatic speaker recognition is proposed. The features are extracted using MFCC and using these extracted features, input samples are created. Finally, recognition is performed by a Multilayer perceptron feed-forward neural network which is trained by backpropagation.

The network has been trained and tested using a database of IITG Multivariability Speaker Recognition. The proposed method provides an accuracy of 94.44 % in comparison with other methods. In the future, we will be implementing ASR using Deep learning for better recognition.

REFERENCES

1. Jain, Anil K., Arun Ross, and Salil Prabhakar, "An introduction to biometric recognition." *IEEE Transactions on circuits and systems for video technology* vol.14, Jan. 2004.
2. Kraljevski, I., Bissiri, M. P., & Hoffmann, R, " Text independent speaker identification with coded speech," *Elektronische Sprachsignalverarbeitung*, 2013, pp. 239-246. Dave, Namrata, "Feature
3. extraction methods LPC, PLP and MFCC in speech recognition." *International journal for advanced research in engineering and technology* vol.1, July 2013, pp.1-4.
4. Chanwoo Kim, Richard M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," IEEE/ACM Transactions on Audio, Speech and Language Processing .vol. 24, July 2016, pp. 1315-1329.
5. Hsieh, C-T., E. Lai, and Y-C, Wang, "Robust speech features based on wavelet transform with application to speaker identification," IEEE Proceedings-Vision, Image and Signal Processing, vol.14, April, 2002,pp.108-114.
6. Poonkuzhali, C., R. Karthiprakash, S. Valarmathy, and M. Kalamani, "An approach to feature selection algorithm based on ant colony optimization for automatic speech recognition," *International Journal of Advanced Research in Electrical, Electronics, and Instrumentation Engineering* vol. 2, 2013 ,pp. 5671-5678.
7. Wei Li, Tianfan Fu, Jie Zhu, "An improved i-vector extraction for speaker verification," EURASIP Journal on Audio, Speech, and Music Processing 2015, pp. 1-9.
8. Virendra Chauhan, Shobhana Dwivedi, Pooja Karale and Prof. S.M. Potdar. Speech to text converter using Gaussian Mixture Model (GMM). International Research Journal of Engineering and Technology (IRJET) vol. 3, 2016, pp.160-164.
9. Brucal, Stanley Glenn E., Aaron Don M. Africa, and Elmer P. Dadios, "Female Voice Recognition Using Artificial Neural Networks and MATLAB Voicebox Toolbox," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol.10, 2018, pp.133-138.
10. Sharma, Poonam, and GARG ANGALI, "Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks," *International Journal of Computer Applications*, vol.142, 2016, pp. 12-17.
11. Wang, J., Ji, A., Johnson, M. T, (2012, July). "Features for phoneme independent speaker identification. In *2012 International Conference on Audio, Language and Image Processing,"* July 2012, pp. 1141-1145.
12. Pawar, R. V., P. P. Kajave, and S. N. Mali. "Speaker Identification using Neural Networks," *IEC (Prague)*. 2005.
13. Ge, Z., Iyer, A. N., Cheluvaraja, S., Sundaram, R., & Ganapathiraju, A, "Neural network-based speaker classification and verification systems with enhanced features," In *2017 Intelligent Systems Conference (IntelliSys) Sep. 2017*, pp. 1089-1094.
14. Pawar, M. D., & Kokate, R. (2019), "A Robust Wavelet Based Decomposition and Multilayer Neural Network for Speaker Identification," In *Innovations in Electronics and Communication Engineering* pp. vol., 2019197-209.
15. Haris B. C. et al, "Multivariability speaker recognition database in Indian scenario," International Journal of Speech Technology, vol.15, 2012, pp. 441-453.
16. Seddik, Hassen, Amel Rahmouni, and Mounir Sayadi, "Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier," First International Symposium on Control, Communications and Signal Processing, March 2004, pp. 21-24. IEEE, 2004

AUTHORS PROFILE

Kh. Jilenkumari Devi is presently working as Lecturer in NIT Manipur in the Dept. of ECE. She received her M.Tech (ECE) from NERIST, Arunachal Pradesh and, currently pursuing a Ph.D. in NIT Manipur. Her area of interest Includes Image Processing, Signal Processing, and Neural Networks.

A. Alice Devi received her M.Tech (CSE) from NIT Manipur, B. Tech (CSE) from Manipur Institute of Technology, Manipur. Her area of interest includes Speech processing and Neural Network.

Dr. Khelchandra Thongam is an Assistant Professor in NIT Manipur in the Dept. of CSE. He received his Ph. D from the University of Aizu, Japan, M.Tech (CSE) from the University of Aizu, Japan. His area of interest includes Intelligent System Design, Soft Computing, Hybrid Intelligent System, and Robotics.