Speaker Recognition Using MFCC and Combination of Deep Neural Networks

Keshvi Kansara[1], Dr. A.C. Suthar[2]

[1] *ME. Student, L.J Institute of Technology, Ahmedabad, Gujarat, India*
[2] *Guide and Director, L.J Institute of Technology, Ahmedabad, Gujarat, India*

ABSTRACT

Speaker Recognition is a process of validation of a person's identity based on his/her voice. The two major steps in speaker recognition are: Extraction of Features and Matching of Features. A direct analysis and synthesizing of the complex voice signal is difficult as a large amount of information is contained in the signal. Several methods have been used to obtain speaker related features and matching of the features. In the following paper we combine the use of different methods to obtain a speaker recognition system. The conventional method can be combined with other methods to obtain better efficiency of the system. In the proposed approach the feature extraction is done using conventional Mel Cepstral Coefficients and a Butterfly structure Deep Neural Network (also called as Deep Autoencoder). After obtaining the coefficients the Deep Neural Network (DNN) is trained for the classification purpose. DNN can be directly used to pull out features and then classify speakers using same DNN but the MFCC and Auto-encoder are used at first for data compression and to get maximum number of features thus getting better efficiency and faster results. The Deep Autoencoder undergoes unsupervised training and DNN undergoes supervised training. Features are obtained for three different classes and these features are then used to train the Deep Neural Network. After the training phase deep neural network can be used as classifier but only for the classes that have been used during the training and for the same text or phrase used during the training time.

Keyword: - *Auto-encoder, Butterfly Structure, DNN, MFCC, Speaker Recognition.*

1. INTRODUCTION

Speaker Recognition is a biometric manner that takes into account the Speaker, or voice, for recognition purposes. The speaker recognition process relies on features subjective to both the physical structure of a person's vocal tract and the behavioural characteristics of the individual. The physiological part of speaker recognition is the physical shape of the person's voice tract. The behavioural component is related to the physical movement of jaws, tongue as well as larynx.

The speaker recognition method should not be confused with speech recognition. The speech recognition system would recognise the words as they are spoken, which is not a biometric method. Also the speaker recognition has two major forms that is speaker identification and the other one being speaker verification. Briefly, speaker verification makes the binary decision of whether the speaker is who he claims to be or not. On the other hand Speaker identification is comparing the voice of speaker with a given set of database in order to identify who the speaker is. In this paper speaker identification is the focus of research work.

The Speaker recognition technique has two main modules also which are: Feature Extraction and Feature Matching. The figure below shows the two modules and the methods we that are used in the project.

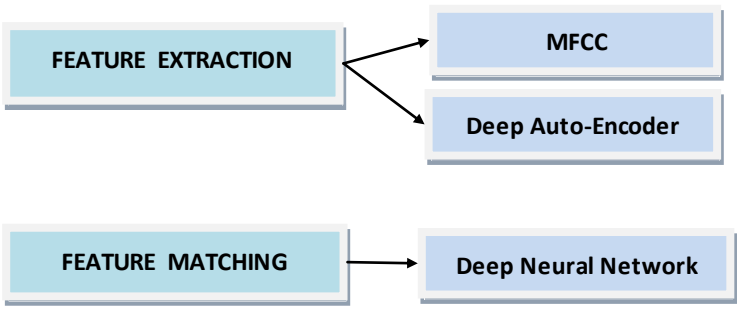

Fig-1: Speaker Recognition Modules

The two modules are explained below:

Feature Extraction: This module converts the speech signal into a set of features or feature vectors which are then used for further analysis.

Feature Matching: The feature extracted in the previous stage are matched with the already stored reference features and based on matching recognition is done.

The Features are first extracted and then the extracted features are then compared with the reference features in the model database. The feature with the maximum similarity is taken as the output and its label is considered as the identity of the person. Thus the speaker can be recognized using the feature or set of features

In the proposed approach we have combined the classical way and heuristic way of extracting features. In the first part, we want to extract features. It is very intuitive in the field of learning and philosophy that extracting the correct features for a particular task is the most difficult part in any speech or speaker recognition based system. Humans can easily recognize the features, but for machines to learn to extract features is very hard and time cumbersome. In our approach, we are proposing a hybrid way of extracting features. We use autoencoder (bottle-neck) features and classical MFCC features in order to represent a speaker. We call it as speaker – based features.

2. FEATURE EXTRACTION BY MFCC

The Process of Obtaining Mel Frequency Cepstrum Coefficients is given below:

2.1 Pre-emphasis

In speech processing, the original signal has a high content of lower frequency components, and processing the signal to accentuate higher frequency energy is necessary. This step passes the signal via a filter which enhances higher frequency ranges. This process increases the energy of signal having higher frequencies. Then each value in the signal is again evaluated using this formula:

$$Y[n] = X[n] - \alpha * X[n-1]$$

2.2 Framing

The input speech signal is divided into small 20-30 ms frame with an overlap of one half of the frame size. The speech signals are not the stationary signals, but can considered stationary for short period of time. Generally the size of frame is taken equal to power of two in order to make possible the use of FFT.

2.3 Windowing

The window function is used to smoothen each frame of the signal. By multiplication of the frame with window function to attenuate both of the ends of the signal towards zero smoothly, the unwanted transients can be avoided. The window that is applied here is hamming window. The equation for hamming window is given as:

$$w(n) = \alpha - \beta * \cos(2\pi n/N-1) \text{ where,}$$

$$\alpha = 0.54 \text{ and } \beta = 1-\alpha = 0.46$$

2.4 Fast Fourier Transform (FFT)

The next step to the process is the Fourier Transform. It converts each frame from the time domain into the frequency domain. It is usually performed to obtain the magnitude frequency response of each frame. When DFT is applied to a frame, its implicit that the signal within a frame is periodic and also continuous when wrapping around. If that isn't the case, then also Fourier Transform can be performed but the discontinuity at the frame's first and last points can introduce unwanted effects in the Frequency Response. To overcome this, we have two options which are given below:

1. Multiplying each frame with a Hamming window to increase the continuity at the first and the last points.

2. Taking a frame of non consistent size so that it always has an integer multiple of the fundamental periods of signal.

2.5 Mel Filter Bank Wrapping:

Human sensitivity of the frequency contents of sound for speech signals does not trail a linear scale. For each tone with Frequency (f), measured in Hz, a subjective pitch is calculated on the scale called Mel scale. The Mel-frequency scale has a linear frequency spacing less than 1000Hz and a log spacing above 1000Hz. The equation to compute the Mel value for a given frequency f in Hz is:

$$\text{Mel}(f) = 2595 * \log10(1+f/700)$$

2.6 Cepstrum Construction:

In the last step, the log Mel spectrum is converted to time domain. The result is called as Mel frequency cepstrum coefficients of that signal. The cepstral representation of the speech spectrum provides a wise representation of the local spectral properties of signal for the given frame analysis.

In the proposed approach the entire process is carried out for all the 6000 frames and 12 MFCC's are extracted. Along with that, we take delta and delta-delta features of the MFCC as the features in order to use it as features in the feature vector when we train a DNN as classifier. Following were the MFCC parameters.

Table-1: MFCC Parameters

Parameter	Value
Frame length (N)	257
Number of output Coefficients	12
Sampling rate	16kHz
Frame shift (M)	128

The output obtained for 1 class of speaker is shown below. By one class we mean one of the speakers.

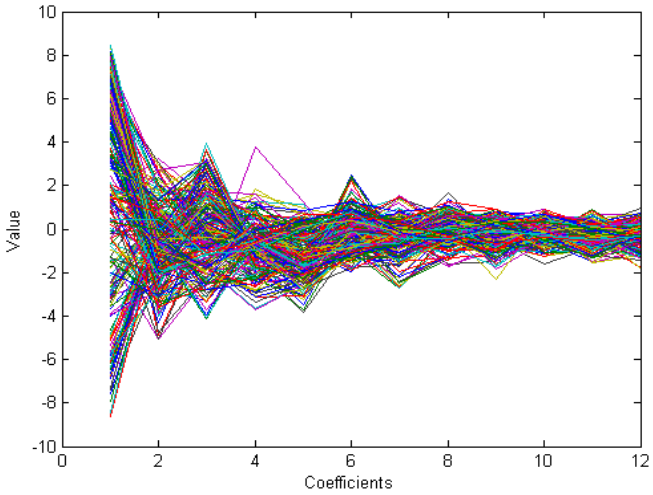

Fig-2: MFCC of one Class.

3. FEATURE EXTRACTION AND DIMENSIONALITY REDUCTION BY USING DEEP AUTO-ENCODER

Deep auto-encoder is a type of network algorithm which undergoes unsupervised learning. This auto- encoder consists of two symmetrical DNNs that usually have few shallow layers representing the encoding half of the net and second set of layers that make up the decoding half.

To extract the autoencoder features, each speech signal is divided into say some number of frames such that we get 6000 frames in total. Now, it is very essential in speech communication that speech processing cannot be done directly on time-series data of speech and specifically when we are dealing speech in Deep learning area. Any speech processing should be done on Short time Fourier transform (STFT) version of the speech frames. Of all these 6000 frames, we obtain STFT and then by using it we extract features from autoencoder in unsupervised way.
Following table shows the autoencoder parameters:

Table-2: Autoencoder Parameters

Parameter	Value
Number of input	257
Number of output	257
Size of hidden layer 1	128
Size of hidden layer 2	40 – (Output of this layer is our Feature vector)
Size of hidden layer 3	128
Structure	257 – 128 – 40 – 128 – 257
Learning rate	0.005
Decay parameter	0.01

Number of iterations	100
Batch size	100
Number of batches	53

After training the autoencoder, we can say that we have 40 features for corresponding frame of any class. After that, we will extract the MFCC features.

The result of the one of the class is given below. It is the same class used to extract MFCC features.

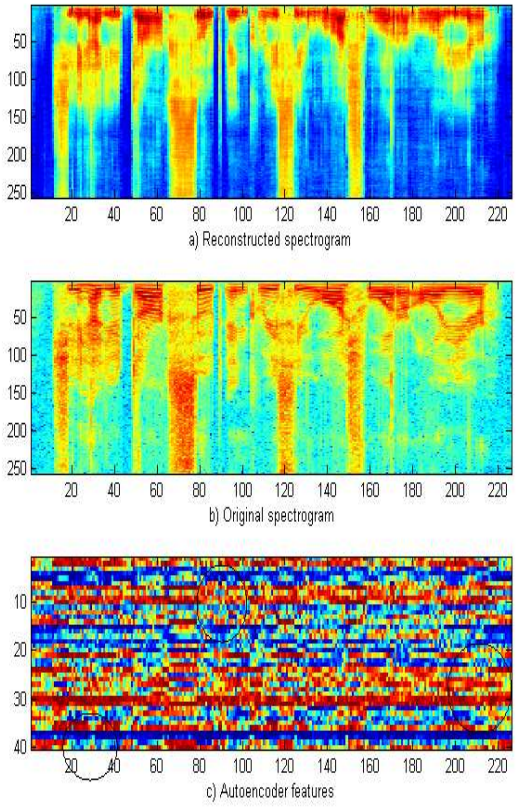

Fig-3: Autoencoder Features of on class.

4. FEATURE MATCHING BY DEEP NEURAL NETWORK

A Deep neural network is nothing but a feed forward artificial neural network with more than one hidden layer. This hidden layer is between the input and the output. One can train any neural network in following stages.
1) feed-forward pass 2) back propagation and 3) weight updating process. Each hidden unit in the network 'j' takes into account a non-linear function to map its total input from the layer before to the following layer.

$$x_j = b_j + \Sigma y_i w_{ij}$$

Where,
y_j = input to the present layer
w_{ij} = weights connecting current layer to the next layer.

b_j = biases of the current layer i.e j th layer,
$y_i = f(x_j)$, here f could have been any of sigmoid, tanh or

A cost function C is associated with a network, which is of the form,

$$C = ||y - f(x)||^2$$

Where, f(x) is the network output and y the labels.

Back-propagation is a process wherein error and error derivatives are back propagated in the network in order to correct the weights which are incorrect, which indirectly gives error or a bad cost function value. After this the weights can be updated using any of the systematic gradient descent methods, LBFGS etc. The stochastic gradient descent methods follow equation of the form:

$$W_j = W_j - \alpha (\partial C / \partial W_j)$$

Where, C is the cost function of the network. Once the network is trained, the value of cost function becomes to be very negligible as desired, nearly 0, weight update stops and we say, we have obtained set of trained weights which set apart our network for a specific use.

5. CONCLUSION

Hence, we can conclude that we have come up with a unique way of extracting features in the field of speaker recognition. We have successfully combined the framework of Deep learning with the classical speaker recognition in terms of feature extraction. We have implemented the feature extraction stage and our further work will include speaker classification which will ultimately lead to recognition.

6. REFERENCES

[1] Douglas Reynolds, Najim Dehak, Fred Richardson, "Deep Neural Network Approaches to Speaker and Language Recognition", IEEE Signal Processing Letters, 2015.

[2] Jun Du, Li-rong Dai, Yan-hui Tu, "speech Separation Based On Signal-noise Dependent Deep Neural Networks For Robust Speech Recognition", IEEE International Conference On Acoustic, Speech And Signal Processing 2015.

[3] Yun Lei And Nicolas Scheffer, "A Noise Robust I-vector Extractor Using Vector Taylor Series For Speaker Recognition", IEEE 2013

[4] Mitchell Mclaren, Yun Lei And Luciana Ferrer: "Advances In Deep Neural Network Approaches To Speaker Recognition", Speech Technology And Research Laboratory, IEEE 2015

[5] Shanthi Therese S and Chelpa Lingam, "Speaker based Language Independent Isolated Speech Recognition System",INSPEC Accession Number:14933464, IEEE 2015.

[6] Xin Lei, Ehsan Variani And Erik Mcdermott: "Deep Neural Networks For Small Footprint Text-dependent Speaker Verification", International Conference On Acoustic, Speech And Signal Processing , ICASSP 2014 6854363, IEEE 2014.

[7] Sailesh Khaparkar, Anand Vardhan Bhalla, "Perfprmance Improvement Of Speaker Recognition System", International Journal of Advanced Research in Computer Science and Software Engineering, March 2012.

[8] Srinivas Govinda, Ritu Pal, Surampudi,"Speech Signal Processing Using Neural Network", IEEE International Advance Computing Conference 2015.

[9] V. Srinivas, Dr. T. Madhu, Dr. Ch. Santhi Rani, "Neural Network Based Classification For Speaker Identification", International Journal Of Signal Processing, Image Processing And Pattern Recognition 2014.