IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 4, April 2016)

Text-Dependent Speaker Recognition using Deep Neural Networks

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

[1]PG Student, [2]Guide and Director, Department of Electronics and Communication Engineering, L.J.I.E.T, Ahmedabad.

***Abstract*— Speaker Recognition is used for identification of a person depending on the characteristics contained in the speech signal. In this paper we propose the use of Deep Neural Network (DNN) for text dependent speaker Recognition system (SRS). Mel Frequency Cepstral Coefficients (MFCC) and Auto-encoder (Butterfly Structure Neural Network) are used to extract the features of speech signal at the initial stage. The previously obtained coefficients are then used to train the DNN to later classify the speakers. DNN can be directly used to extract features and classify speakers but the MFCC and Auto-encoder are used initially for data compression and maximum number of feature extraction thus aiming to get better efficiency and faster results.**

***Keywords*—Deep Auto-encoder, DNN, MFCC, Speaker Recognition**

I. Introduction

Speech signal is the most powerful form of communication because of its rich dimensions. The primary information associated with the speech signal is the message to be conveyed from the speaker to the listener whereas the rich dimensions refer to the gender, emotion, language, health condition and the identity of the speaker. While speech signal deals with extracting the linguistic information present in the signal, speaker recognition deals with preserving features related to the identity of the speaker. With the use of speech as a biometric in access control systems and forensics the task of identifying a speaker from their voices have increased tremendously. There are two types of speaker recognition methods: 1) Text-dependent and 2) Text-Independent. Text-dependent speaker recognition method uses phoneme context information and thus high recognition accuracy can be achieved easily. Text-Independent method operates on unconstraint speech and so is more likely to errors and less accuracy.

The success of a speaker recognition system depends on extracting the right speaker dependent features which should be invariant in the articulation dynamics.

The selection of a feature should be such that it must have large variation between speakers and small variations between different sessions of the same speaker, it must be robust against noise and channel effects and hard to mimic or reproduce.

In past years many methods and algorithms for speaker recognition have been developed. Text-dependent speaker recognition using DNN has been proved to get magnificent results. Motivated by the powerful feature extraction capability and re-cent success of deep neural networks applied to speech recognition [6], we propose a SR technique based on DNN. In this paper a new method is proposed wherein two types of DNN are used out of which one uses unsupervised learning and the one other supervised leaning. The algorithm with unsupervised learning is known as Auto-encoder. An auto-encoder neural network is an unsupervised learning algorithm that pertains back propagation, setting the target values to be equal to the inputs. The auto-encoder along with MFCC is used for feature extraction and simple DNN is used for classification purpose. The MFCC extracts the features hidden in high frequency components where as the Deep auto encoder extracts the low frequency components. The different features extracted using MFCC and Auto encoder helps in maximum efficient training of Deep neural network and so can be helpful in better classification.

The figure below shows the proposed block diagram of system:

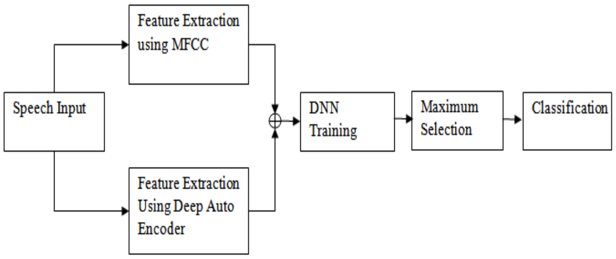
Feature Extraction using MFCC
Speech Input
DNN Training
Maximum Selection
Classification
Feature Extraction Using Deep Auto Encoder

Figure 1: Proposed Block Diagram Of Speaker Recognition System

II. Obtaining Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral coefficients are the most widely used features for speaker recognition. It combines the advantages of Cepstral analysis with the perceptual frequency scale based on critical bands. MFCC is obtained on the basis of human hearing perceptions which cannot accept frequency more than 1Kilohertz. In other words MFCC is based on human years critical frequency bandwidth. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000 Hz. It is obtained by the steps following:

A. Pre-emphasis

In speech processing, the original signal usually has much of lower frequency energy, and processing the signal to emphasize higher frequency energy is necessary. This step processes the passing of signal through a filter which emphasizes higher frequencies. This process increases the energy of signal at higher frequency. Then each value in the signal is re-evaluated using this formula:

$$Y[n] = X[n] – \alpha * X[n-1]$$

B. Framing

The input speech signal is segmented into small 20-30 ms frames with an overlap of one half of the frame size. The speech signals are non-stationary signals, but are considered stationary for short period of time. Generally the frame size is taken equal to power of two in order to facilitate the use of FFT.

C. Windowing

The window function is used to smooth each frame of the signal. By multiplying the fame with window function to attenuate both ends of the signal towards zero smoothly, the unwanted artefacts can be avoided. The window that is used here is hamming window. The hamming window is given as:

$$w(n) = \alpha - \beta *\cos(2\pi n/N-1), \text{ with } \alpha = 0.54 \text{ and } \beta = 1-\alpha = 0.46$$

D. Fast Fourier Transform

The next processing step is the Fourier Transform, which converts each frame of from the time domain to the frequency domain. We usually perform FT to obtain the magnitude frequency response of each frame. When we apply DFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around.

If that is not the case, we can still perform Fourier Transform but the discontinuity at the frame's first and last points is likely to introduce unwanted effects in the FR. To deal with this problem, we have two strategies [7]:

1) Multiply each frame by a Hamming window to increase its continuity at the first and last points.
2) Take a frame of a variable size such that it always contains an integer multiple number of the fundamental periods of the speech signal.

E. Mel Filter Bank Wrapping

Human perception of the frequency contents of sounds for speech signals does not follow a linear scale. For each tone with an original Frequency, f, measured in Hz, a subjective pitch is calculated on the Mel scale. The mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz.The formula to compute the mel for a given frequency f in Hz is:

$$Mel(f) = 2595 * \log 10 (1+f/700)$$

One way of simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel frequency component. That filter bank has a triangular band pass FR, and the spacing as well as the bandwidth is determined by a constant Mel-frequency interval. The modified spectrum hence consists of the o/p power of these filters. The number of Mel spectrum coefficients, K, is typically chosen as 20.

F. Cepstrum Construction

In the final step, the log Mel spectrum has to be converted back to time domain. The result is called as Mel frequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. As the Mel spectrum coefficients are real valued numbers and so are their logarithms, they may be converted to the time domain using the Discrete Cosine Transform (DCT). It is known that the logarithm has the effect of changing multiplication into addition [7]. Thus we can simply convert the multiplication of the magnitude of the Fourier transform into addition. Finally by taking the DCT of the logarithm of the magnitude spectrum, MFCC can be obtained.

III. Feature Dimensionality Reduction Using Deep Auto-Encoder

Deep auto-encoder is a type of neural network algorithm that undergoes unsupervised learning.

This auto- encoder consists of two symmetrical deep neural networks that typically have four or five shallow layers representing the encoding half of the net and second set of four to five layers that make up the decoding half. The figure below shows the simplified structure of Deep Auto-encoder:

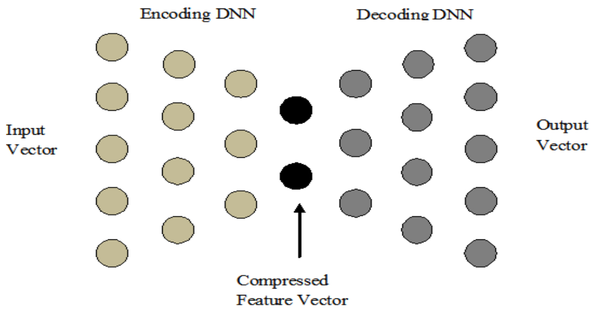

Figure 2: Deep Auto-Encoder

As it belongs to unsupervised learning the training example set $\{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}...., x^{(n)}\}$ where $x^{(i)} \in \mathrm{R}^n$ is unlabeled. The auto encoder applies back propagation, setting the target values to be equal to the input values i.e. $y^{(i)} = x^{(i)}$, where, $y^{(i)}$ is the target set. The aim of training this network is to reduce the error of reconstruction useful coefficients of speech signal. Often a reduced set of hidden units is use, creating an information coefficient set. The weights between the input and hidden layer are often related with hidden layer and the output.

IV. Supervised Learning For Classification

A Deep neural network is a feed forward artificial neural network (ANN) with more than one hidden layers between the input and the output. One can train any ANN in three stages. a) feed-forward pass b) back propagation c) weights updating process. Each hidden unit 'j' uses a non-linear function to map its total input from layer before to the next layer.

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

where,
b_j = biases of the current layer i.e j th layer,
y_j = input to the present layer
w_{ij} = weights connecting current layer to the next layer.
$y_i = f(x_j)$, here f could have been any of sigmoid, tanh or ReLU (rectified linear unit), as non-linear function.

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

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

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

Back-propagation is an algorithmic process where error and error derivatives are propagated backwards in the network in order to correct the incorrect weights, which indirectly is responsible for the error or bad cost function value. The weights can be updated using any of the systematic gradient descent methods, or LBFGS and so on. 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 cost function value becomes very small as desired, hypothetically 0, weight update stops and we say, we have optimized set of trained weights which characterize our network for a specific application.

V. Conclusion

The Mel Frequency Cepstral coefficients have been obtained in a vector form. These coefficients along with the coefficients obtained by Auto-encoder act as features to the DNN.

VI. Future Work

A Deep Neural Network based Speaker recognition system will be implemented, where MFCC and auto-encoder will be the crucial features extracted instead of direct speech signal for the training purposes. MFCC and auto-encoder are used in order to reduce he dimensionality as speech is a highly complex data.

It will be concluded on the basis of experimental results that MFCC, auto-encoder and DNN based SR is better than the classical SR and can be used for the practical purposes of SR in various tools.

REFERENCES

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

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

[3] Fred Richardson, Douglas Reynolds And Najim Dehak: "Deep Neural Network Approaches To Speaker And Language Recognition" Ieee Signal Processing Letters, Vol. 22, No. 10, INSPEC Accession Number: 15071019, IEEE 2015.

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

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

[6] Ehsan Variani, Xin Lei 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] Anand Vardhan Bhalla, Sailesh Khaparkar, “Perfprmance Improvement Of Speaker Recognition System”, International Journal of Advanced Research in Computer Science and Software Engineering, March 2012.

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

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

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