A Review of Mel Frequency Cepstral Coefficient (MFCC) Analysis of Speech and Neural Network

Ibrahim Khalil
Nanjing University of Science and Technology, China

Abstract— Human voice is the most natural and easiest way of communication with each other. But communication with a machine is not a simple task to do. After lots of hard works people have been developed computer to communicate with machine. Automatic speech recognition permitting a natural and simple to use method of communication between human and machine is an active area of research. Mel Frequency Cepstral Coefficient (MFCCs) are the most prominent and commonly used features in most of the speaker and speech recognition applications. Recently, another acoustic model based on deep neural networks (DNN) has been presented. This paper gives a review of MFCC's improvement techniques that are applied in the speech recognition system. Here also discussed neural networks.

Keywords: MFCC, Neural Network

I. INTRODUCTION

Speech is an acoustic signal which contains information of impression that is formed in the speaker's brain. Speech could be a helpful interface to interact with machines. Researchers are working for quite a long time to improve this kind of communication. From the evolution of computational power, it has been conceivable to allow a system to speak in real time conversations. But in spite of good development made in this field, the speech recognition is still facing a lot of problems and error. These problems are because of the variations occurred in speaker including the variations as a result of age, sex, a variety of language, speed of speech signal, the emotional state of speaker can cause the difference in the pronunciation of various persons. Surroundings can add noise to the signal. Sometimes speaker causes the additional of noise itself [1]. In speech recognition process, an acoustic signal captured by microphone or telephone is transformed to a set of characters. A view about automatic speech recognition (ASR) is given by describing components of future human computer interface. Hence for the communication with machines human could use speech as a useful interface. Speech recognition mainly focuses on training the system to recognize an individual's unique voice characteristic. There are many techniques that can be used for speech recognition. For feature extraction Mel Frequency Cepstral Coefficients (MFCC) as it is fewer complex in implementation and more effective and robust under various conditions. MFCC is intended using the knowledge of human auditory system. It is a standard method for feature extraction in speech recognition [2]. Perceptual Linear Prediction (PLP), Relative Spectra of Log Domain Coefficients PLP (RASTA-PLP), Linear Predictive Coding (LPC), Predictive Cepstral Coefficients (LPCC), Deep Neural Network (DNN) technique also can be used for speech recognition.

II. TYPE OF SPEECH RECOGNITION SYSTEM

A. Speaker Based Recognition Systems:

1) Speaker Dependent Models:

It is an acoustic model that has been altered to distinguish a specific person's speech. They are primarily more accurate for the specific speaker. These systems are usually easy to develop and require not so much capital but rather more accurate[3].

2) Speaker Independent Acoustic Models:

Speaker Independent system can recognize any of speaker without any prior training. (IVRS) Interactive Voice Response System used this model. Independent Acoustic model accepts input from extensive number of various clients. The development of this system is extremely difficult, and the cost of development is exceptionally high. Its accuracy is lower than speaker dependent acoustic systems.

3) Speaker adaptive Models:

Speaker adaptive recognition system uses the speaker dependent data and to the best appropriate speaker to distinguish the speech and reduces error rate by adaption [4].The adapt operation according to the characteristics of speakers.

B. Utterances Based Recognition System:

1) Isolated Words:

Isolated word recognition system which recognizes single utterances word. Isolated word recognition is appropriate for circumstances where the user is required to give only one-word response or commands. It is very simple and most straightforward for implementation because word limits are distinct, and the words tend to be clearly pronounced which is the major advantage of this type.

2) Connected Words:

A connected words system is like isolated words, but it allows separate utterance to be "run -together" with an insignificant interruption between them. Utterance is the pronunciation of a word or words that represent a single meaning to the computer.

3) Continuous Speech:

Continuous speech recognition system permits users to speak almost normally, while the computer determines its content. The development of Continuous speech recognition system is very hard.

4) Spontaneous Speech:

Spontaneous Speech recognition system recognizes the natural speech. Spontaneous speech is natural that suddenly came out of the mouth. An ASR system with spontaneous speech can deal with a variety of natural speech features for example words being run together, Spontaneous speech may include mispronunciation, false-starts, and non-words[4].

III. FEATURE EXTRACTION MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

Feature extraction in ASR is the computation of a sequence of feature vector that provides a compact representation of the specific speech signal. It is basically performed in three main stages. The first phase is known as the speech analysis or the acoustic front-end, which performs spectra-temporal analysis of the speech signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second phase, the compilation of an extended feature vector composed of static and dynamic features. Finally, the last phase of these extended feature vectors transforms into progressively reduced and robust vectors which are then provided to the recognizer.[5]

MFCC is based on the perception of human hearing which cannot distinguish frequencies over 1khz. In other words, the MFCC is based on known alteration of the human ear's critical bandwidth with frequency. MFCC has two types of filter that are linearly spaced at the low frequency below 1000Hz and have logarithmic spacing above 1000Hz. A particular pitch is present on Mel Frequency Scale to capture the important characteristic of phonetic in speech.[6][7]. The overall process of the MFCC is shown in "Fig.1."

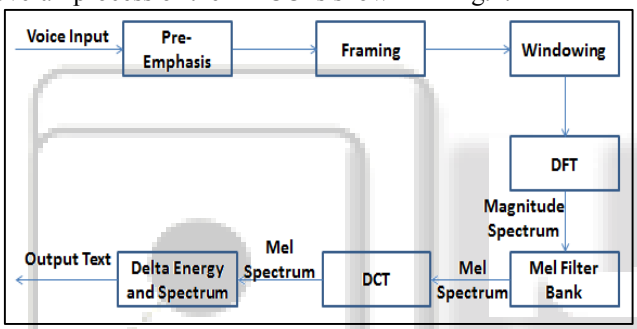

Fig. 1: MFCC Block Diagram [8]

As shown in figure 1, MFCC consists of seven computational steps. Each step has its function and mathematical approaches as discussed briefly in the following.

A. *Pre-Emphasis:*

On this step, passing of the signal through a filter which emphasizes higher frequencies. This procedure will increase the intensity of signal at higher frequency.

$$h(z) = 1 - az^{-1} \qquad 0.9 < a < 1 \qquad (1)$$

The most typical value of a is about 0.95. However, the signal spectrum is boosted approximately 20 dB/decade by pre-emphasis filter[6].

B. *Framing:*

The speech signal is typically isolated into small duration blocks, called frames and the spectral analysis is performed on these frames. This is because the signal of human speech changes over time gradually and can be treated as a quasi-stationary process. The common frame length and framing of the speech recognition task are 20 – 40 msec. The acoustic signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N)[6]. The values used are usually M = 100 and N = 256.

C. *Windowing:*

Hamming window is used as window shape by considering the following block in the extraction processing sequence and integrating all the nearest frequency lines. The Hamming window is usually used, it specifies the accuracy of the frequency resolution of the spectral analysis while minimizing the level of the window transfer function[9].

$$y(n) = x(n)w(n) \qquad (2)$$

Hamming window is used for speech recognition task as:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n}{N-1}\right) \qquad (3)$$

D. *Fast Fourier Transform:*

To change over each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement the equation below:

$$y(w) = \text{FFT}\,[h(t) * X(t)] = H(w) * X(w) \qquad (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

E. *Mel Filter Bank Processing:*

A set of triangle bandpass filters that simulating human's ear features is applied to the spectrum of the speech signal. This process is called Mel filtering. The human ears analyze the sound spectrum in groups grounded on several overlapped critical bands. These bands are distributed in such a way that the frequency accuracy is higher in the low frequency region and lower in the high frequency region as illustrated in "Fig. 2,"[9].

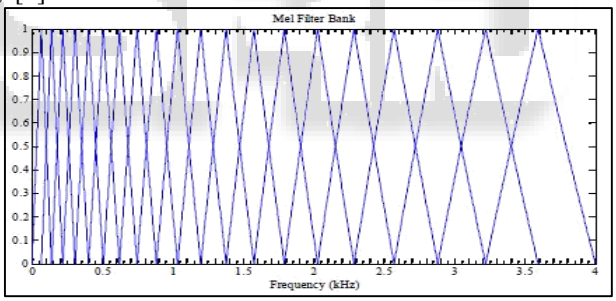

Fig. 2: The Mel-scale filter bank [9]

The Mel frequency is computed from the linear frequency as:

$$f_m = 2525 \times \log 10\left(1 + \frac{f}{700}\right) \qquad (5)$$

Where f_m = is the Mel frequency for the linear frequency f. The filter bank energy is obtained after Mel filtering. This leads to the definition of MFCC, a baseline acoustic feature set for speech and recognition applications[9].

F. *Discrete Cosine Transform:*

This is the process to change over the log Mel spectrum into time domain using Discrete Cosine Transformation (DCT). The higher order coefficient represents the excitation information, or the periodicity in the waveform, while the lower order cepstral coefficients represent the vocal tract shape or smooth spectral shape[12]. DFT can also be used to calculate Coefficients. DFT is commonly used for spectral analysis whereas DCT used for data compression as DCT signals have more information concentrated in a small number of Coefficients and thus, it is anything but difficult to

speak Mel spectrum in a comparatively modest number of coefficients and requires less capacity. This instead of using DFT DCT is desirable for the coefficient computation as DCT outputs can contain significant amounts of energy. [10]The output after applying DCT is known as MFCC.

$$C_n = \Sigma_{k-1}^{k}(\log D_k)\cos[m(k-\frac{1}{2})\frac{\pi}{k} \quad (6)$$

Where m = 0,1 …. K-1

Where, C_n represents the MFCC and m is the number of the coefficients.

G. Delta Energy and Delta Spectrum:

The voice signal and frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy) and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2 is represented at the equation below:

$$\text{Energy} = \sum x^2[t] \quad (7)$$

Each of the 13 delta features represents the change between frames in the equation 8 corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features[6].

$$d(t) = \frac{c(t+1)-c(t-1)}{2} \quad (8)$$

IV. Neural Network (NN) in Speech Recognition Process

There are many similarities between neural networks and Markov models. These two models are statistical models which are represented as graphs. Where Markov models using for state transitions and probabilities, neural networks using for connection strengths and functions these days. A key difference is that neural networks are basically parallel while Markov chains are sequential. Neural networks act like a human brain. These networks perform learning phenomenon. Neural networks are a computational model inspired by an annual central nervous system which is capable of machine learning and pattern recognition. The artificial neural networks are usually presented as systems in which multiple neurons are associated with one other. These systems have been used to solve a wide assortment of assignment that is difficult to solve using common rule-based programming including computer version, speech recognition. Neural networks perform very well at learning phoneme probability from exceptionally parallel audio input, while Markov models can use the phoneme observation probabilities that neural networks provide to produce the most probable phoneme sequence or word. This is at the fundamental hybrid method to natural language understanding. Classification process in the Neural network(NN) shown in figure 3 [13].

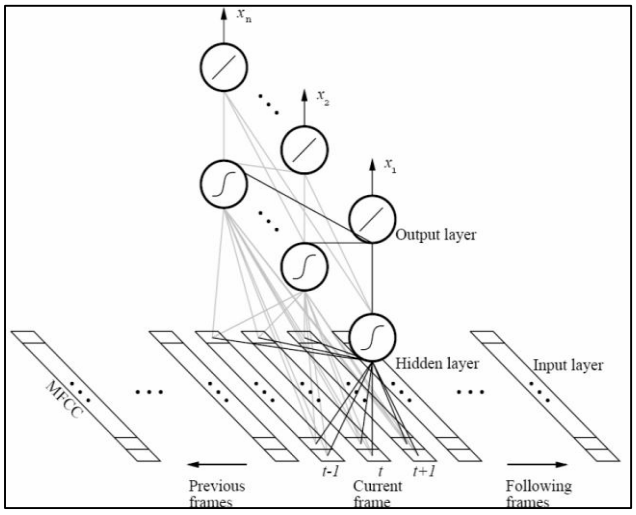

Fig. 3: Classification process in the NN[13].

V. Pattern Matching Using Neural Networks

Artificial neural networks (ANNs) are intelligent systems connected to a simplified biological model of the human brain in another way. The neural networks possess the capacity to self-learning capability from fault tolerant and noise immune and can be used to identify the system, recognize pattern, classification, speech recognition, image processing etc. Here is two typical NNs.

A. Multilayer Feedforward Network

The first type of neural nets used for speech classification is a Multilayer Feedforward Network using Back Propagation algorithm for the training session. This kind of NN is the most popular NN and is used worldwide in many different types of application[15].

B. Radial Basis Function Network

Another approach to characterizing the speech samples is to use of Radial Basis Function Network. This network comprises three layers: an input layer, a hidden layer, and an output layer. The main difference between this type of networks is that the hidden layer has (Gaussian) mapping functions. It is mainly used for function approximation but can also solve classification problems. Redial means that it is symmetric around their center, basis functions mean that a linear combination of their functions can generate (approximate) an arbitrary function[16].

VI. Conclusion and Future Work

In this review paper, we have discussed the basic of speech recognition system and different approaches. Some feature extraction pattern matching and their pros and cons have been discussed. MFCC is the most frequently used features extraction techniques in the fields of speech recognition and speaker verification applications. Neural Network and HMM are considered as the most dominant pattern recognition techniques used in the field of speech recognition. In future, there will be focus on the development of a large vocabulary audio recognition system. Developing such systems Artificial Neural Network (ANN) which will be used in a video that can recognize human mouth expression and recognize the speech.

REFERENCES

[1] Wouter Geuaert, Georgi Tsenav, Valeri Mladenov, "Neural Network used for Speech Recognition" Journals Automatic Control, volume.20.1.7, 2010
[2] Parwinder Pal Singh, Pushpa Rani, "An approach to Extract Feature using MFCC" IOSR Journal of Engineering(IOSRJEN), vol. 04, issue 08 (August. 2014)
[3] Suman k. Saksamudre, P.P Shrishrimal, R.R. Deshmukh, "A Review on Different Approaches for Speech Recognition System" International Journal of Computer Applications, Volume 115-No.22,April 2015.
[4] Pratik k. Kurzekar, Ratndeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishimal, "Continuous Speech Recognition System: A Review", Asian Journal of Computer Science and Information Technology,2014.
[5] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", International Journal for Advance Research in Engineering and Technology, Volume 1, Issue VI, July 2013.
[6] Lindasalwa Muda, Mumtaj Begam and Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Volume 2, Issue 3, March 2010.
[7] Zaidi Razak, Noor Jamilah Ibrahim, Emran mohd tamil, modh Yamani Idna Idris, Mohd yaakob Yusoff, "Quranic verse recitation feature extraction using Mel Frequency Cepstral Coefficient(MFCC)" University Malaya.
[8] www.researchgate.net/figure/MFCC-Block-Diagram-7_fig1_ 281652896/download_
[9] X. Xiong, "Robust speech features and acoustic models for speech recognition", PhD. Thesis, 194 p., Nanyang Technological University, Singapore, 2009.
[10] Jamal Price, sophomore student, " Design an automatic speech recognition system using matlab", University of Maryland Estern Shore Princess Anne.
[11] Samuel Kim and Thomas Eriksson, "A pitch synchronous feature extraction method for speaker recognition", IEEE International Conference on Acoustics, speech and signal processing, 2004. Proceedings. (ICASSP '04), vol. 1 pp. I-405-408.
[12] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad and Arpit Bansal, "Feature Extraction Using MFCC" Signal & Image Processing : An International Journal (SIPIJ) Vol.4, No. 4, August 2013.
[13] Sayf A. Majeed, Hafizah Husain, Salina Abdul Samad, Tariq F. Idbeaa, "Mel Frequency Cepstral Coefficients (MFCC) feature extraction enhancement in the application of speech recognition: A comparison study", Journal of Theoretical and Applied Information Technology,10th September 2015. Vol.79. No. 1.
[14] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, "Neural Networks used for Speech Recognition", Journal of Automatic Control, University of Belgrade, vol. 20:1-7, 2010.
[15] Akram M. Othman, and May H. Riadh, "Speech Recognition Using Scaly Neural Networks", World Academy of Science, Engineering and Technology, vol.38, 2008.
[16] Mohamad Adnan Al-alaoui, Lina Al-kanj, Jimmy Azar, and Elias Yaacoub, "speech Recognition using Artificial Neural Networks and Hidden Markov Models", IEEE Multidisciplinary Engineering Education Magazine, VOL. 3, 2008

