Mel Frequency Cepstral Coefficient: A Review

Shalbbya Ali[1], Dr. Safdar Tanweer[2], Syed Sibtain Khalid[3], Dr. Naseem Rao[4]

{shalbbya.ali@gmail.com[1], safdartanweer@gmail.com[2], s.sibtainkhalid@jamiahamdard.ac.in[3] , naseemjmi0786@gmail.com[4]}

Department Of CSE, Jamia Hamdard

Abstract. Speech is the easiest and widely used mode of communication between human beings. The development of a human computer interface for establishing a similar dialogue between the machine and humans was the inspiration behind speech recognition systems. This involves training a system so that voice recognition can be carried out. The machine has to be trained using any of the speech recognition algorithms which would extract the features of the voice and perform speech recognition. One such algorithm is the Mel-frequency Cepstral coefficient. The paper describes all the stages of the MFCC technique along with brief description of each process. It also gives a comparison between linear predictive coding and MFCC technique.

Keywords: Speech Recognition System, MFCC, Mel Frequency, Feature extraction.

1 Introduction

The interaction that takes place between a human and a computer is a common form of communication in today's scenario. This kind of interaction can only be carried out with the help of some hardware devices like keyboard, touchscreen, mouse etc. but humans prefer a more natural form of communication rather than using hardware devices for the communication. Speech is the widest and profound means of communication used by human beings. The inspiration behind Speech Recognition Systems is the human to human interaction, voice being the biggest feature. Speech Recognition can enable a dictation facility by which devices can be controlled and documents can also be created. The creation of documents by speech saves time and effort made by a human. The business organizations have been greatly benefitted by this domain. There exists a voice technology where the callers can input their data and get a response later without actually interacting with a live agent. It has led to cost reduction and optimization in the business domain eliminating the cost that would be used in keeping a live agent. Non-verbal communication between human and a computer already existed but using voice for the same has led to the development of the area of Speech Recognition. A voice sample contains lot of information starting from gender

ICIDSSD 2020, February 27-28, New Delhi, India

DOI 10.4108/eai.27-2-2020.2303173

and age of a person sometimes even reflecting the state of mind of the speaker. The motive of Voice Recognition is to use some of this information and identify the speaker on one or more of these parameters. There are many techniques that can be used to extract features from a voice sample which may fall under spectral or cepstral domain.

Voice Recognition is a technique that identifies a voice sample from some unique properties that maybe acoustic or phonetic. These characteristics can be represented in the form of a speech signal see **Fig. 1.**

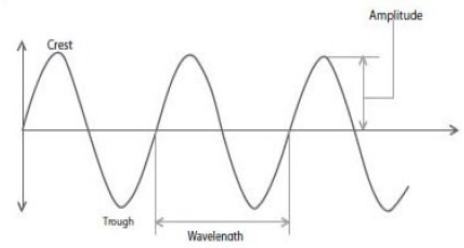

Fig. 1. The Sound wave of Speech.

The block diagram of Speaker Identification System is shown in the figure below see **Fig.2**. A sample of speech is taken and after some initial processing, properties of the speech sample are extracted. These properties maybe acoustic properties or phonetic properties. The features are extracted from the voice sample using some feature extraction technique. There are many extraction techniques that exist. Some of them are MFCC, LPC RASTA etc. The selection of a technique totally depends on what feature we want to exploit, the size of sample, the size of vocabulary, whether the speech is an isolated word or a bunch of continuous words. After the features have been extracted, the training is being carried out on the algorithm.[1]

The data that is offered during training period is very closely related to the data that would be given during the testing period. The data is then stored in the database for future use. During testing period, some data closely based on the data offered during training period is given to the system. Some processing is done on this data and pattern matching is done. If the data matches with the one stored in the database, the system has clearly been trained else retraining would be done or the result would show that the pattern has not been matched.[1]

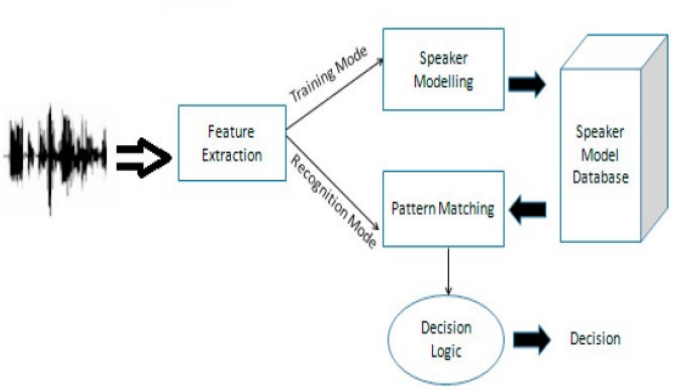

Fig. 2 Block Diagram of Speech Recognition System.

The voice recognition is divided into two categories:
1.1 Speaker Identification
1.2 Speaker Verification
1.1 Speaker Identification: It is the process by which a voice is identified from a group of speakers. The voice which best matches with the voice which is already stored in the database is identified and incase the voice does not match with the one stored in the storage then the new voice becomes another voice in the database which would be matched when a new voice has to be recognized as shown in **Fig 3.** [1]

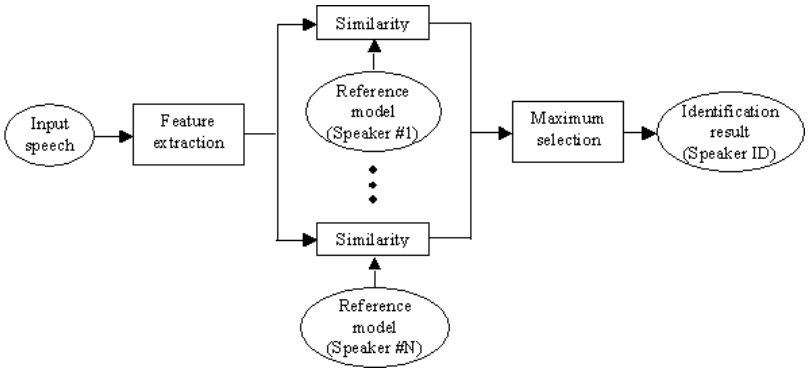

Fig. **3** Block Diagram of Speaker Identification System.

1.2 Speaker Verification: It is the process which authenticates a speaker based on some property present in the voice sample. The identity claim maybe accepted or rejected. This type of recognition is widely used in military,aircraft and other authentication areas. The biometric identification also falls under the same category. The voice that is authenticated is the ones whose parameters match the most with the one present in the database as shown in **Fig 4.** [1]

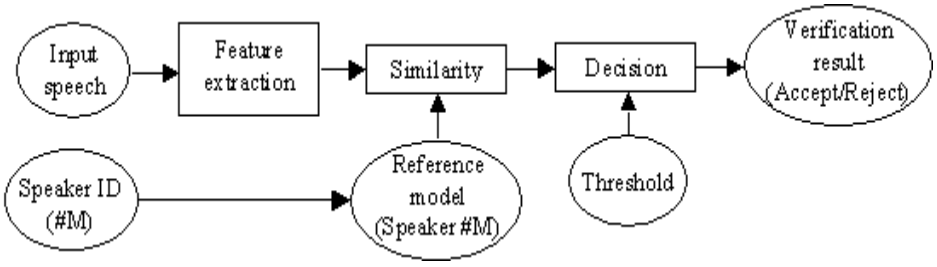

Fig. 4 Block Diagram of Speaker Verification System.

There are many feature extraction techniques which help in extracting some features from a speech sample .One such technique is the Mel-Frequency Cepstral Coefficient. It is the most widely used technique for speaker recognition because of its ease of implementation and flexibility[2].

2 Feature extraction steps in MFCC

The step by step computation of MFCC is described by the block diagram in **Fig 5.**

2.1 Pre-Emphasis- The speech signal x(n) has to be sent through high-pass filter:

$$y(n) = x(n) - a * x(n-1)\ . \tag{1}$$

where y(n) is the output signal and the value of a is usually between 0.9 and 1.0.The processing of signal is done by passing it through a filter which would emphasize its higher frequency. The energy of the signal is increased at higher frequency[2].

2.2 Frame Blocking- The speech signal is divided into small segments of 20-30ms which are known as frames. There may be N no. of frames where each frame that is adjacent to the other frame is segregated by M(M<N). The values of M and N typically used are 100 and 256 respectively [2].

2.3 Windowing - It is a technique usually used in Signal processing where a speech signal is segmented into temporal fragments. The borders of the signal that are repeating themselves are unrelated to the real world signal. Windowing function is a smooth function which goes to zero in the extremes. The idea is to hide the discontinuity at the borders. After the application of the window function, there is a change in the signal but the effect on signal statistics in minimized.The window function usually used in MFCC is the Hamming Window function. It maintains the continuity of the first and last points in the frame.

The hamming window function is defined the equation below where W (n) is the window function [2].

$$Y(n) = X(n) * W(n) \, . \tag{2}$$

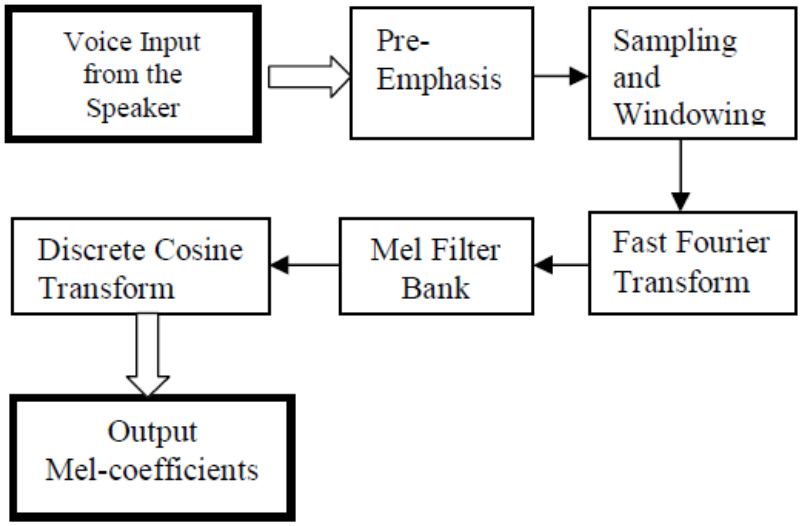

Fig. 5 MFCC Block Diagram.

2.4 Fast Fourier Transform- It is a signal analysis technique which is used to extract and compress some features of the speech signal without losing any relevant information so that speech processing becomes easier. It represents the given signal in a frequency domain. All the relevant information

from the original signal is retained only the representation of the signal is changed. The time-domain signal is converted into frequency domain signal[3].
Triangular Band-Pass Filters- The magnitude frequency response is multiplied by a set of triangular band pass filters so that a smooth magnitude spectrum is attained.

2.5 Discrete Cosine Transform- The formula for Discrete Cosine Transform is

$$C(n) = \sum Ek * \cos\left(n * (k - 0.5) * \frac{\pi}{40}\right)) \quad (3)$$

where n = 0,1,..to N and N is the number of triangular band pass filters, L is the number of mel-scale cepstral coefficients. The DCT is carried out on the output of the band pass filters to generate mel-scale coefficients. The frequency domain signal is transformed into time-domain signal and the features are also termed as the mel-scale cepstral coefficients or mel-frequency cepstral coefficients which is used for speech recognition[3].

3 Linear Predictive Coding

This technique is also widely used for voice recognition. It uses the preceding samples and forms a linear combination. The primary aim of LPC is frame-based analysis of the input voice signal to create observational parameters. To apply LPC and generate the functionality, input voice signals need to be surpassed using a pre-emphasizer. Pre-emphasizer output acts as the blocking input to frame in which the signal is distorted into N test pieces. Further, windowing is performed in which each frame is windowed in a way that the signal interruption at the beginning and end of each frame is reduced. Then, each frame which is windowed is self-correlated and the total auto-correlation value is the value of the LPC calculation[4][6].For the comparison between MFCC and LPC see **Table 3.1**.

3.1 Comparison between LPC and MFCC

1.	Linear prediction coefficients (LPC) mimic human vocal tract and offer a robust speech characteristic.	MFCC is a reproduction of the human hearing mechanism designed to selectively apply the operating theory of the ear assuming that the human voice is a good speaker recognizer.
2.	It supports single speaker and single language and is quick in implementation.	It supports multiple speakers and multiple languages and is a little slow in implementation.
3.	Small vocabulary size makes it reliable.	It is reliable for vocabulary of reasonable to high size.
4.	Linear prediction technique is used to achieve the filter parameters comparable to the vocal tract by lowering the mean square error between the input and the projected speech.	MFCC are coefficients generated from human auditory experience on a distorted frequency scale based.
5.	It is used in speaker recognition systems where the main purpose is to retrieve the property of the vocal tract.	These are commonly used to recognize speakers and understand speech.
6.	This offers very precise estimates of speech parameters and is relatively good for calculation.	Certain pattern recognition problems relating to human voice are said to be accurate in cepstral coefficients.
7.	LPC estimates are highly sensitive to quantizing noise, and may not be appropriate for generalization.	The MFCC features are not totally accurate when background noise is present and might not be well adapted for generalization.
8.	The LPC approach is commonly used in music and technical companies to build mobile robots, in telephone companies, violin tonal evaluation and other instrumental string gadgets.	For security purposes, MFCC is used to classify airline reservations, numbers spoken in a telephone and speech recognition software.

4 Literature Review

Nisha in the year 2017 in her paper discussed Speaker identification and Speaker Verification techniques. The identification of voice falls under the category of Speaker Identification where the given voice is compared with a set of already existing similar sample voices present in the database and then a match is done. However, Speaker Verification is the process of authenticating a speaker based on some parameter or in other words the speaker is identified. Text dependent and Text independent voice recognition systems have also been discussed. Along with this, various feature recognition techniques like Linear Predictive Coding (LPC), Perceptual Linear Prediction(PLP), Relative Spectral Filtering(RASTA) and Mel-Frequency Cepstral Coefficient (MFCC) along with their merits and demerits have been discussed.The merits and demerits can easily give a trade-off to choose the feature extraction technique best suited to the current situation and time. Some classification techniques like Hidden Markov Model (HMM), Neural Network Modelling Approach(NN), Dynamic Time Warping (DTW) and Vector Quantization(VQ) have also been discussed.Ayushi T. N. P et. al. (2018) have highlighted the use of technology in the field of communication. Applications have been created that can take part in this communication. The functions used in establishing such a link include acoustic and articulatory speech recognition, converting speech signal to text and then again to a signal, language translation etc. The techniques for achieving the able mentioned functionalities have also been discussed in the paper. The speech recognition systems can be classified into various parameters like speaker, vocal sound, vocabulary etc. The general prerequisite for any feature extraction technique involves some preprocessing followed by feature extraction by any of the methods be it LPC,MFCC or Dynamic Time Warping. It also brings forth the acoustic models and a basic idea about the phonemes.The pattern classification approach includes template based, knowledge based, statistical and neural network based approach. The speech to text conversion methods include Hidden Markov Model HMM, Artificial Neural Network Classifier based on Cuckoo Search Optimization. The text to speech conversion phases include Text processing, articulator synthesis, formant synthesis and concatenative synthesis. The language translation techniques include rule based machine translation, statistical machine translation, example based machine translation and hybrid machine translation. The findings and issues of each functionality has been clearly mentioned and explained. In speech to text HMM has been classified as the best technique despite of the drawback that are present in it and in text to speech formant synthesis has been concluded as the best approach for conversion. Hybrid translation finds itself to be popular because of its fast learning ability and data acquisition.Ayushi Y. K. et. al. (2017) have explained various Speech Recognition Algorithms along with the techniques and challenges involved in the same. The issue with recognition involves the utterance approach because no two people can speak in the same way, there is always a difference it maybe because of their age, gender, geographical location or any other reason. Some styles of speaking are very quick and some of them are really slow. The same word pronounced in the American accent might sound different in the British accent. All these issues have to be analyzed before developing an algorithm. There are various types of speaker models, which may either be text dependent or text independent. The size of vocabulary also has an impact on speech recognition along with the background in which the word or sentence is uttered. Different speech recognition techniques like MFCC, PLC etc have also been briefly discussed. The approaches to classification

include acoustic phonetic approach, pattern recognition approach and artificial intelligence approach.Pattern matching approaches for speech recognition have also been discussed. The future scope regarding large vocabulary systems and speaker independent continuous systems have also been discussed in the paper.Shaik Riyaz, B. et. al.(2019) in their paper have used speaker recognition for Urdu language. The dataset used in doing so consists of 250 different Urdu words which are spoken by 20 different speakers which comprises of 8 male speakers and 12 female speakers.The reason for using HMM and MFCC for accomplishing the same has been found to have the highest accuracy as compared to other feature extraction techniques and models. The dataset contains varied speakers whose accent in Urdu language varies from Kannada accent to Uttar Pradesh accent. Vector Quantization is used to compress the size of the feature vector and improve accuracy. For classification, Hidden Markov Model has been used. The performance is based on the accuracy of the Urdu words spoken with turns out to be 96.4% even after using minimum no of features as mentioned in the paper[7][8].

5 Conclusion

The Mel-Frequency Cepstral Coefficient algorithm is a good choice for speech recognition where multiple speakers exist. The algorithm supports multiple languages, multiple speakers and a high vocabulary size.

Moreover, the implementation of this algorithm is quite easy as it successfully extracts features from the voice sample. In computational context MFCC is far more expensive because of the use of Fast Fourier Transform when measuring its range. However, the noise present in the background environment does affect the recognition and impacts the performance of the algorithm. The change from a noisy environment to a noise free location can help in improving the results and performance.

References

[1] Nisha. Voice Recognition Technique: A Review. International Journal for Research in Applied Science Engineering Technology (IJRASET). May 2017;Volume 5(Issue 5)

[2] Koustav Chakraborthy, Asmita Talele & Prof. Savitha Upadhya.Voice Recognition Using MFCC Algorithm. International Journal of Innovative Research in Advanced Engineering (IJIRAE). November 2014;Volume 1 (Issue 10).

[3] Parvinder Pal Singh,Pushpa Rani.An Approach to Extract Feature Using MFCC.IOSR Journal of Engineering (IOSRJEN). August. 2014;Volume 04, (Issue 08).

[4] Ayushi Y.Vadwala, Krina A. Suthar, Yesha A. Karmakar, Nirali Pandya.Survey Paper on Different Speech Recognition Algorithm:Challenges and Techniques. International Journal of Computer Applications.October 2017 Volume 175.

[5] Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik, Supriya Agrawal. Speech to text and text to Speech Recognition Systems-A Review.IOSR Journal of Computer Engineering(IOSR-JCE). March-April 2018;Volume 20(Issue 2).

[6] Ayushi Y.Vadwala, Krina A. Suthar, Yesha A. Karmakar, Nirali Pandya.Survey Paper on Different Speech Recognition Algorithm:Challenges and Techniques. International Journal of Computer Applications.October 2017 Volume 175.
[7] Riyaz, Shaik et al. Automatic Speaker Recognition System in Urdu using MFCC & HMM.International Journal of Recent Technology and Engineering (IJRTE) February 2019; Volume-7(Issue-5S4).
[8] F. Leu and G. Lin.An MFCC-Based Speaker Identification System.IEEE 31st International Conference on Advanced Information Networking and Applications *(*AINA*)*, Taipei, 2017, pp. 1055-1062.