*Indian Journal of Natural Sciences* www.tnsroindia.org.in ©IJONS

*Vol.13 / Issue 76 / February / 2023* *International Bimonthly (Print)* *ISSN: 0976 – 0997*

**RESEARCH ARTICLE**

# MATLAB based Automatic Speech Recognition using MFCC

Anitha Bujunuru[1]*, C.Silpa[2] and B.Mythily Devi[3]

[1]Associate Professor, Department of ECE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

[2]Associate Professor, Department of ECE, Malla Reddy Engineering College, Hyderabad, Telangana, India.

[3]Assistant Professor, Department of ECE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India.

Received: 26 Aug 2022 Revised: 09 Dec 2022 Accepted: 11 Jan 2023

***Address for Correspondence**
**Anitha Bujunuru,**
Associate Professor,
Department of ECE,
Guru Nanak Institutions Technical Campus,
Hyderabad, Telangana, India



## ABSTRACT

In this paper a novel process of Automatic speech recognition technique using Mel Frequency Cepstrum Coefficient (MFCC) is presented. Speech recognition is a procedure of transforming signal of speech to a word sequence. Now a days the important challenges in research and development is, designing a more accurate speech recognition system, it consists of several constraints including representation of speech, speech classifier, feature extraction and feature matching. In this work, a MATLAB code is developed to analyze the relation between linear frequency and Mel frequency, frequency characteristics of speech signal using Mel- Frequency Cepstrum Coefficient (MFCC), feature vector of the given speech signal. By comparing of two speech signals the similarity has been calculated.

**Keywords:** Speech recognition, HMM, MFCC, Feature Extraction, Feature matching, Mel frequency.

## INTRODUCTION

Basic communication is in the form of speech. It plays vital in the development of emotional and social skills. Speech is produced by the biologic organs of human vocal tract and articulators. Each has its own frequency for a particular person are affected by emotional, gender etc. Various speech recognition methods are used for different types of accents, pronunciation and roughness of voice that includes dynamic programming. Phoneme is a basic unit of sound and it is a minimal unit that can be used to differentiate meanings of words [2]. Speech recognition is basically dividing into two types.

i) Speaker Recognition
ii) Speech Recognition
Speaker Recognition is the method of finding out the speaker who has been convey the particular speech and it does not any training [9]. Speech recognition is the process of calculating the parameters of speech wave such as fundamental frequency, amplitudes, power etc. and it needs the training of signal [3][4]. A basic speech recognition system begins with a pre-processing stage whose input is a speech signal and processes the signal and produces a feature vector which gives the necessary information required for recognition. Section II represents literature survey, section III explains the speech recognition system based on Hidden Markov Model (HMM), section IV describes the realization of speech recognition system, simulation results are described in section V section and VI describes the conclusion.

**Literature Survey**
The speech recognition basic concepts were begun in 1960's with investigation of voice print analysis that was close to fingerprint concept. In 1980's Mel Frequency Cepstrum Coefficients (MFCC) representation for speech recognition was described by Davis & Mermelstein as a useful approach to speech recognition and it becomes a widely used technique for feature extraction [1]. In 1984 George Orwell's describes that a machine can
identify the human voice [ 6].

**Hidden Markov Model (Hmm) Based Speech Recognition System**
Speech signal can be converted into computer readable signal by using automatic speech recognition system (ASR), to obtain this a continuous speech signal converted into discrete parameter vector sequence. Speech recognition systems are described by Hidden Markov Models and these are statistical models that can generate series of symbols or quantities [6][7].
HMMs are utilized in speech recognition because
1) Speech signal can be represented as a 10 – 25 m sec of piecewise stationary signals.
2) Speech signal are quite simple and train automatically to perform calculations.
HMM output consists of a series of n- dimensional real valued vector (n must be a small integer i.e.,10) and generating one of these for every 25msec.

**Proposed Methodology**
The schematic approach of Automatic speech recognition system using HMM is shown in Fig1. Two main blocks of speech recognition system are feature extraction and feature matching. Feature extraction section will transform speech signal into another form of representation that will process and generate message, this extracted data is referred as feature vector. There are three basic methods for extracting feature vector, namely MFCC (Mel-Frequency Cepstrum Coefficient), LPC (Linear Predictive Coding) and perceptual linear prediction (PLP). In modern speech Recognition technique, commonly used feature extraction methods are MFCC and PLP. In feature matching, the extracted feature vector from unknown voice signal is tested with respected to the acoustic model and the model with maximum score is considered as recognized word. Feature matching is done by using two methods, namely VQ (Vector Quantization) code book and Gaussian mixture model (GMM).

**MFCC (Mel-Frequency Cepstrum Coefficient)**
The performance of speech recognition system is examined with extricating and choosing the relevant parameters of the speech signal. A compact representation of the set of coefficients resulting from the real logarithmic cosine transformation of the short-term energy spectrum expressed on the Mel frequency scale is produced [10]. The MFCC algorithm is routinely used to find the relationship between critical bandwidth and human ear frequencies. The analyses and extraction of feature vectors is done with this algorithm. Flowchart of MFCC is shown in Fig 2, the operations of MFCC that will perform on speech signal are Pre-emphasis, Framing and blocking, Windowing, DFT, Mel filter bank, DCT and Delta and energy spectrum . A continuous speech waveform is divided into frames of fixed length with overlapping. To derive a set of feature vectors from each frame, different signal processing operations have been performed at different stages of MFCC algorithm. The process of MFCC is shown in

**Pre-emphasis**

The main task of pre-emphasis is to suppress the high-frequency components of the speech signal that were conquer by the human sound production system. Speech spectrum is flattened by using pre-emphasis filter priori to spectral analysis. A high pass filter takes an audio input x(n) and produces the output y(n) is given below.

$y(n) = x(n) - a * x(n-1)$-------(1)

Fig2, and explained in detailed in below [11]. where a is a constant and its values ranges between 0.9 and 1.0. By applying z-transform to equation (1), H(z) is

$H(z) = 1 - a * z^{-1}$--------- (2)

**Framing**

Speech signals are non-stationary signals and continuously varies. A speech signal is split into a series of frames and each frame is examined separately and viewed by only one feature vector. The audio wave is divided into frames of M samples, with adjoining frames isolated by N samples. where M is always greater than N. The next frame has N samples after the first frame and overlaps M-N samples and each frame overlaps with two subsequent frames. A typical value for M and N is 256,100.

**Windowing**

Since each frame should have steady motion, the trade-off for achieving frame blocking is to use 40 ms windows given at an interval of 25 ms (frame rate is 100 frames/ seconds, and the overlap between adjoint windows is about 25%). A tapered window is applied to each to minimize discontinuities in the audio wave at the corners of each frame. The Hamming window is frequently used. The output of Hamming window is

$x(n) * w(n)$ -------- (3)

Where x(n) is signal of input speech and w(n) is window function and it is defined by

$w(n) = 0.54 - 0.46(\frac{2\pi n}{N-1})$------ (4)

Where n ranges from 0 to N-1.

**Fast Fourier Transform (FFT)**

The frequency domain representation of each frame obtained is by applying Fourier Transform. Equation of DFT is given by

$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}$ -------- (5)

Where0≤k≤N-1.

The FFT (Fast Fourier transform) is computationally more efficient algorithm to find the Discrete Fourier transform (DFT). In general, the length of FFT is powers of 2 (N=2 n ) if not zero padding technique is used to make it. The frequency response of each frame is obtained by applying FFT and then calculate the magnitude response. FFT output consists of both the real and imaginary parts Re(X(k)) and Im(X(k)). Only real data is used in speech recognition system. The speech signal magnitude response is obtained by using equation (6). Each frame spectral magnitudes are stored in one matrix as row with 256 columns. The undesirable frequency responses of each frame can be solved by multiplying with Hamming window.

$|X(k)| = \sqrt{(Re(X(k)))^2 + (Im(X(k)))^2}$ ---- (6)

**Mel Frequency Filter Bank**

Analysis of speech signal by Mel-frequency is depends on human perception experiments. The human ear is delicate and has better resolution at low frequencies than at high frequencies. The Mel filter bank was developed to emphasize low frequencies over high frequencies. An audio wave does not accompany a linear scale, but spreads out over a frequency range. The Mel frequency scale is linearly spaced below 1000 Hz and logarithmically spaced above 1000 Hz. The Melscale frequency is proportional to the linear frequency and is given by equation (7).

$Mel(f) = 2595log_{10}(1 + \frac{f}{700})$ --------(7)

The output is given by the sum of its filtered spectral components [5].

**Discrete Cosine Transform (DCT)**

Transform the logarithmic power spectrum to the time domain using the discrete cosine transform (DCT), resulting in Mel-frequency cepstrum coefficients (MFCC). DCT consists of most of the information in lower order coefficients, which results in cost reduction. When done by cepstrum analysis, it is good to represent the local spectral properties of the signal together with the frame analysis.

**Implementation of Speech Recognition System**

The Mel-Frequency Cepstrum Coefficients (MFCC) are measured from the speech wave specified by the vector S and sampled at FS (Hz). Applied speech wave is pre-amplified using a first-order FIR filter with coefficient of predistortion ALPHA. The pre-distorted audio signal performs short-term Fourier transform analysis with a frame duration of TW (ms), a frame shift of TS (ms), and an analysis window function specified as a function handle with window. This is followed by computation of the amplitude spectrum by filter bank design using M-triangular filters equally spaced on the mel scale between the lower and upper frequency limits denoted by R (Hz). A filter bank is given to the magnitude spectral values to produce the filter bank energy (FBE) M for each frame [11]. The logarithmically compressed FBE is decorrelated by the DCT to produce the cepstral coefficients.

The basic idea behind the feature matching is that feature extraction output is comparing with the reference speech signal and the signal which matches maximum is recognized. The amount of similarity can be measured by using the simple method of cross correlation technique.

## RESULTS AND DISCUSSION

The speech signal has been recorded using 'Audacity' software and audio read function in MATLAB to read the given input signal and produce its values and speech inputs are recorded in wave format. By considering different frequencies in Hz, obtain the Mel frequencies. The relationship between linear frequencies in Hz and Mel frequencies are plotted and is shown in Fig3.From the plot, non-linear relationship is existed between frequency in Hz and Mel frequencies. Mel frequency cepstral coefficients of audio wave has been obtained and are shown in Fig4. The simple cross correlation method is used to recognize the speech signal. The given input speech signal is compared with the reference signal. By using MATLAB software the similarity of two signals is calculated and is shown in Fig5.

## CONCLUSION

The basic Hidden Markov Model with MFCC feature extraction has been discussed. The relation between linear frequencies and Mel frequencies are plotted. The Mel frequency cepstral coefficients of speech signal are obtained and are plotted using MATLAB. The comparison of speech signal with reference signal has been done and the amount of similarity is also calculated.

## REFERENCES

1. Davis.S, Mermelstein.P Comparison Of parametric Representations for mono syllabic Word Recognition in Continuously Spoken Sentences. IEEE Transaction on Acoustics, Speech and Signal Processing, Vol.28 No.4.
2. X.Huang, A.Acero and H.Hon.Spoken Language Processing: A guide to theory, Algorithm and system development. Prentice Hall, 2001.
3. en.wikipedia,org/wiki/speech_recognition
4. Aseem Saxena, Amit Kumar Sinha, Shashank Chakrawarti, Surabhi Charu Speech Recognition Using MATLAB, International Journal of Adances In Computer Science and Cloud Computing, 2013.
5. Siddhant C.Joshi and Dr.A.N.Cheeran MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition,2014.
6. Vidwath R Hebse, and Anitha G, The Learning Method of Speech Recognition Based on HMM, 2015.
7. Abdulla Waleed, and Kasabov Nikola, The concept of Hidden Markov Models in Speech Recognition. Technical Report Information Science Department, University of Otago, 1999.
8. Richardson M, Bilmes J, and Diorio C Hidden Markov Models for Speech Recognition. In Speech Com munication Vol 4, 2003.
9. Vibha Tiwari, MFCC and its applications in Speaker Recognition,International Journal On Emerging Technologies, 2010.
10. Lindasalwa Muda, Mumfaj Begam and I.Elamvazuthi, Voice Recognition Algorithm using MFCC and DTW Techniques, Journal of Computing, Volume 2, Issue 3, March 2010.
11. Gunjan Jhawar, Prajacta Nagraj, and P. Mahalakshmi, Speech Disorder Recognition using MFCC, International Conference on Communication and Signal Processing, 2016.

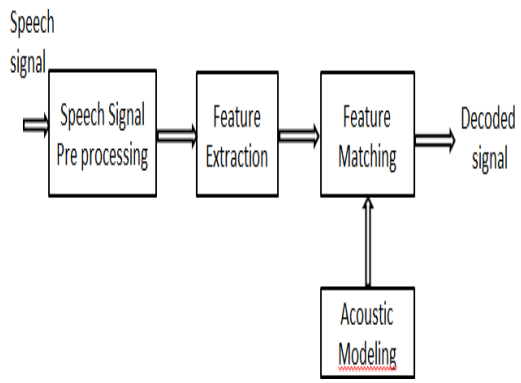


**Fig.1:Schematic approach of Speech Recognition System**

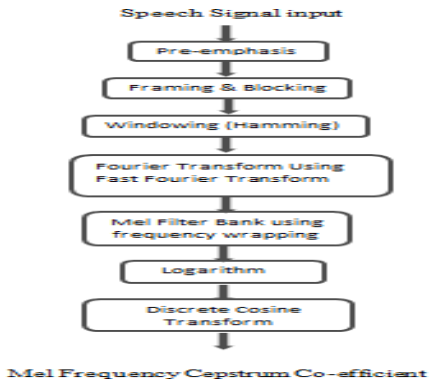


**Fig.2: Flowchart of MFCC**

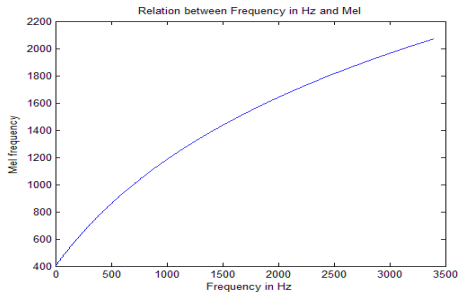


**Fig.3: Relation between linear frequency and Mel frequency**

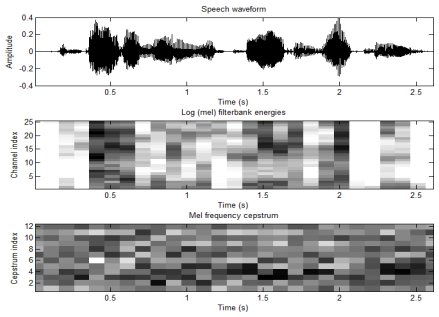


**Fig.4: Mel frequency cepstral coefficients of speech signal.**

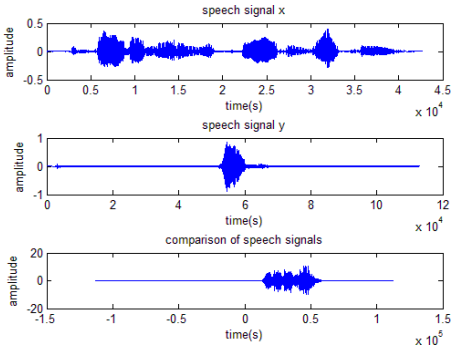


**Fig.5: Similarity of two speech signals**