International Journal for Modern Trends in Science and Technology, 8(05):86-90, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0805013
Available online at: http://www.ijmtst.com/vol8issue05.html

Speech Recognition using Full Length Hidden Markov Model

GannavarapuTanuja | BoppudiNagaManjusha | Bodepudi Akhil Kumar

Department of Electronics and Communication Engineering, R.V.R & J.C College of Engineering, Guntur, A.P, India.
Corresponding Author Email ID: tanujagannavarapu796@gmail.com

To Cite this Article
GannavarapuTanuja, BoppudiNagaManjusha and Bodepudi Akhil Kumar. Speech Recognition using Full Length Hidden Markov Model. International Journal for Modern Trends in Science and Technology 2022, 8(05), pp. 86-90. https://doi.org/10.46501/IJMTST0805013

Article Info
Received: 26 March 2022; Accepted: 25 April 2022; Published: 01 May 2022.

ABSTRACT

Speech is probably the most efficient way to communicate with each other. In speech recognition model mainly two steps are involved. The first one is feature extraction from the speech input signal and second step is training the speech recognition model with the features extracted from the input signal. In this Mel frequency cepstral coefficients (MFCC) extraction technique is used to extract the features from the signal. Before extracting the features from the signal, the speech pre-processing steps has to be performed on the input signal. Hidden Markov Model is used for training the characteristics parameters of the signal. Hidden Markov Model is of two types Full length model, Bakis model. Here full-length model is used for training the characteristics parameters since MFCC is used.

KEYWORDS: Speech Recognition, MFCC, HMM.

1. INTRODUCTION

In speech recognition model mainly two steps are involved.The first one is feature extractionfrom the speechinputsignal and second step is training the speech recognitionmodel with the features extracted from the input signal. Inthis Mel frequency cepstral coefficients (MFCC) extractiontechnique is used to extract the features from the signal withthe help of MATLAB. Before extracting the features fromthesignal,thespeechpre-processingstepshastobeperformed on the input signal. The MFCC feature extractiontechnique basically includes applying the DFT, taking thelog of the magnitude, and then warping the frequencies on a Melscale, followed by applying the inverse DCT. The speech recognition model is trained using the features extracted from the signal through Hidden Markov Model. Hidden Markov Model is used for training the characteristics parameters of the signal. Hidden Markov Model is of two types Full length model, Bakis model. Here full-length model is used for training the characteristics parameters since MFCC is used.

2. SIGNAL ANALYSIS

Voice signal samples into the recognizer to recognize the speech directly, because of the non- stationary of the speech signal and high redundancy of the samples, thus it is very important to pre- process the speech signal for eliminating redundant information and extracting useful information. The speech signal pre-process step can improve the performance of speech recognition and enhance recognition robustness.

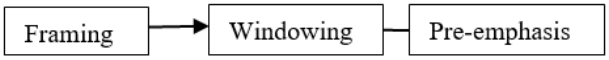


Figure1: Pre-processingStructure

Framing, Windowing, Pre-emphasis are the basic steps used for processing the speech signal before feature extraction.

A. Framing

The speech voice belongs to time-varying signal, which means the speech signal is a nonlinear signal with time changes. Framing is a process of segmenting the sampled speech samples into a small frame. Typically, a frame of 20 - 30 ms could be considered for the short time Time-Invariant property of speech signal. Framing can be classified as non-overlapping and overlapping frames. Overlapping the frames help avoiding information loss in between adjacent frames. Here the overlapping between the adjacent frames will be of 1/3 - 1/2 part of the frame size.

B. Windowing

After slicing the signal into frames, window function such as Hamming window to each frame. Hamming window is more effective to decrease frequency spectrum leakage with the smoother low pass effect. The mathematical solution is given by:

$$w(k) = 0.54 - 046\cos(2\pi n \div (n-1))$$

for; 0≤ n≤N-1

where;

N Represents the length of Window.

n is the sequence of samples

After calculating the hamming window [w(k)], the next step is to multiply each frame by a windowing function to cover the entire speech sequence using the next formula:

x[n] = x[k]w[k–p]

where,

x[k] is the speech sequence.

x is the windowed speech frame at time t.

w[k-p] is the time shifting signal

C. Pre-emphasis

In order to compensate the high-frequency part of the speech signal pre-emphasis process is chosen. Pre-emphasis process often represented by first order High-pass filter (FIR), in order to flatten speech spectrum, and compensate the unwanted high frequency part of the speech signal. The transform function of pre-emphasis can be defined as:

$$H(Z)=1-\alpha Z^{-1}$$

H(z) is the transform function of Pre-emphasis Parameter α is usually between 0.94 and0.97.

3. FEATURE EXTRACTION

The first step in any automatic speech recognition system is to extract features. Mel Frequency Cepstral Coefficient (MFCC) extraction technique is used toextract the features from the signal after the preforming the pre-processing steps on the input signal.

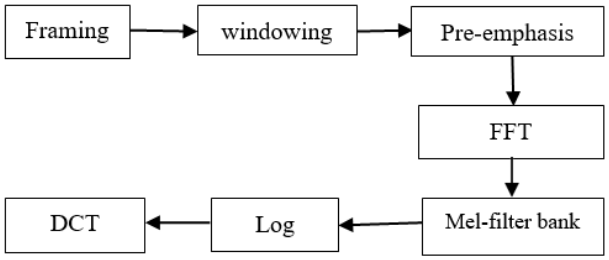


Figure 2: MFCC BlockDiagram

The MFCC feature extraction technique basically includes applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT.

A. Applying FFT

Fast Fourier transform (FFT) algorithm is a fast implementation of the DFT, which converts N- samples of frames into frequency spectrum. After pre- emphasis process is completed, the next step is to apply discrete Fourier Transform on each frame in order to transfer time domain samples into frequency domain. We would generally perform a 512-point FFT and keep only the first 257 coefficients

B. Mel Filter bank

Human hearing is less sensitive at frequencies above 1000 Hz. Therefore, the spectrum is warped using a logarithmic Mel scale to emphasize the low frequency over the high frequency. To achieve the goal, Mel filter bank spectrum used a set of overlapping triangular bandpass filter under a non- linear frequency scale. This filter bank is a set of band pass filters having spacing along with bandwidth decided by steady Mel frequency time. In general, 24- 30 (26 standard) are standard no of triangular filters used for speech recognition. After this

step Logarithm is taken. The relationship between Mel-scale and linear frequency scale is given as

$$Mel(f) = 2595\log(1 + (f \div 700))$$

C.Discrete Cosine Transform

Take the Discrete Cosine Transform (DCT) of the 26 log filter bank energies to give 26 cepstral coefficients. For Automatic speech recognition (ASR), only the lower 12-13 of the 26 coefficients are kept. The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients. Hence DCT is applied in order to obtain Mel Frequency Cepstral Coefficients.

4. HIDDENMARKOVMODEL

HMM is used to classify the features and generate the correct decision. HMM considered the powerful statisticaltool used in speech recognition and speaker identificationsystems, due to the ability to model non-linearly aligningspeech and estimating the model parameters. HMM modelsare of two typesfull length model and linear model.In this we use full length model because MFCC is used for extraction. Figure 3 is fully connected 3-state Hidden Markov Model.S1, S2 and S3 represent the hidden states, O1, O2, and O3 are observations emitted by the hidden states, and aij represents the probability of transitioning from state it o state j. Generally, HMM is characterized by following Number of state N, Number of distinct observation symbol perstate M, State transition probability.

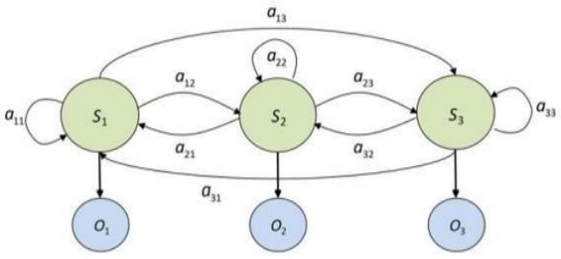


Figure3:Full length model

HMM model can be described by these three sets of parameters a, b and πand the model of N states and M observations referred to by:

$$\lambda = (A, B, \pi)$$

Where,

$A = \{a_{ij}\}$, $B = \{b_j(w_k)\}$ AND $1 < I, J \leq N$ AND $1 \leq K \leq M$

The three cases for HMM are Evaluation, Decoding and Training. The purpose of evaluation is to compute the probability of given sequence O = O_1, O_2, O_3….Ot-1, Ot with given HMM $\lambda = (A, B, \pi)$ that λ has generated the sequence O. Decoding calculates the most likely sequence of hidden states Si of O = O_1, O_2, O_3……..Ot-1, Ot that produced this observation sequence O. In learning the HMM parameters $\lambda = (A, B, \pi)$ are adjusted to maximize the probability to get the best model that represent certain set of observations.

The Strengths of HMM is its mathematical framework and its implementation structure. HMM method is fast in its initial training, and when a new voice is used in the training process to create a new HMM model.

A. Block Diagram

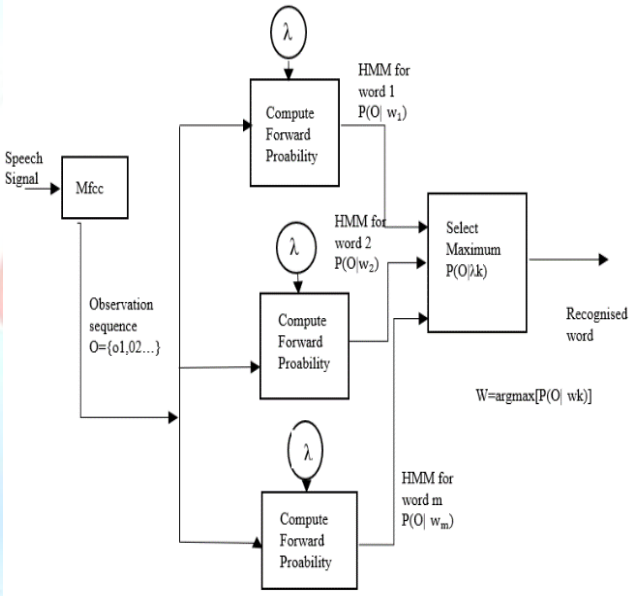


Figure4:Block diagram of speech recognizer

In figure 4 the basic block diagram is presented. The observational sequence is generated after feature extraction of the speech signal. The HMM model has been applied on observational sequence for each word and then by selecting the maximum of P (O| wk) is the recognize word.

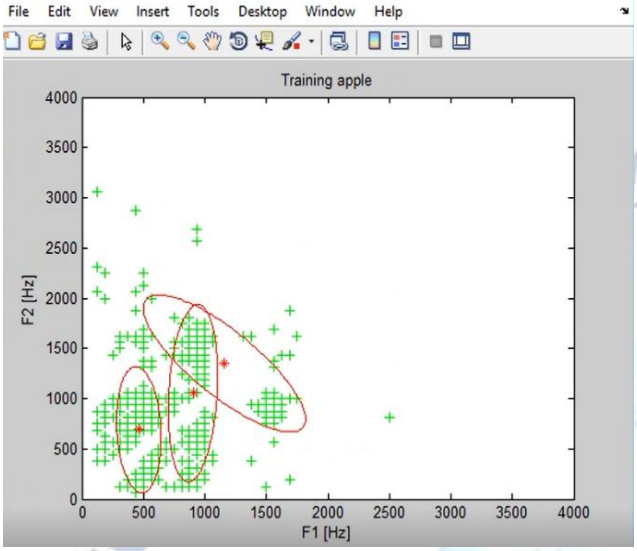
B. Dataset

Database is created of seven fruit names namely Apple, Banana, Kiwi, Pineapple, Orange, Peach, and Lime. For each fruit name 15 utterances are recorded by a single male speaker at sampling rate of 8 KHz. This database is used for training and for testing purpose.

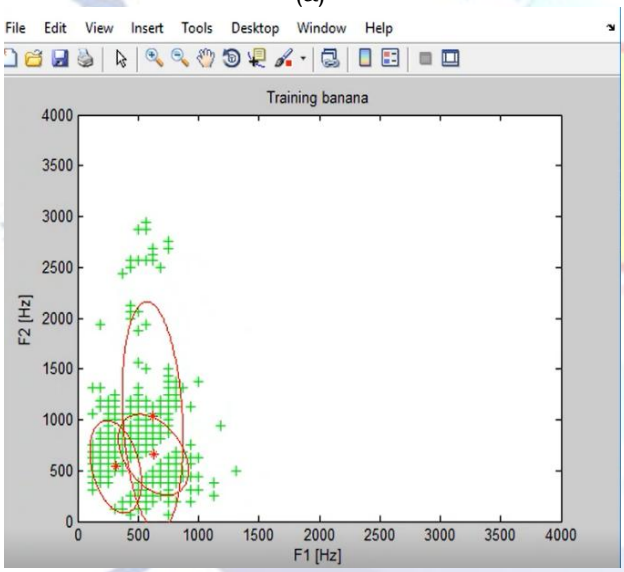
5. EXPERIMENTAL RESULTS

The feature vectors have been extracted for each sound using MFCC algorithm and saved, and the statistical models were generated using Hidden Markov Model

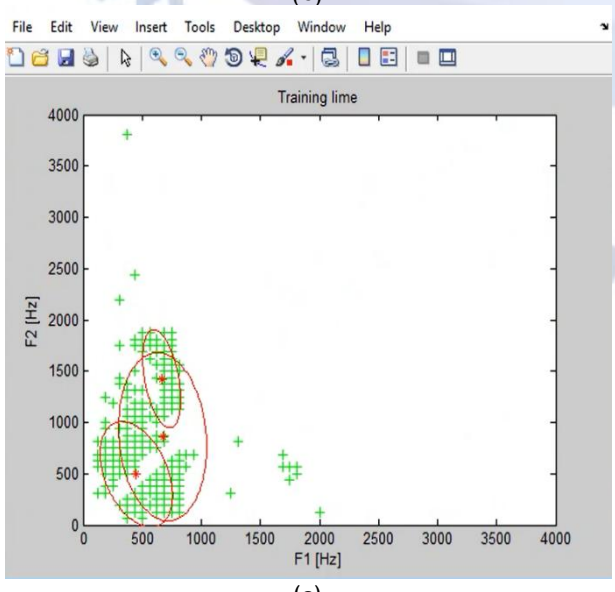
classifier to match the data. In this work, Data set used contain audio files of seven fruit names. Performance evaluation of the Speech Recognition system was obtained by finding the maximum word recognition rate.



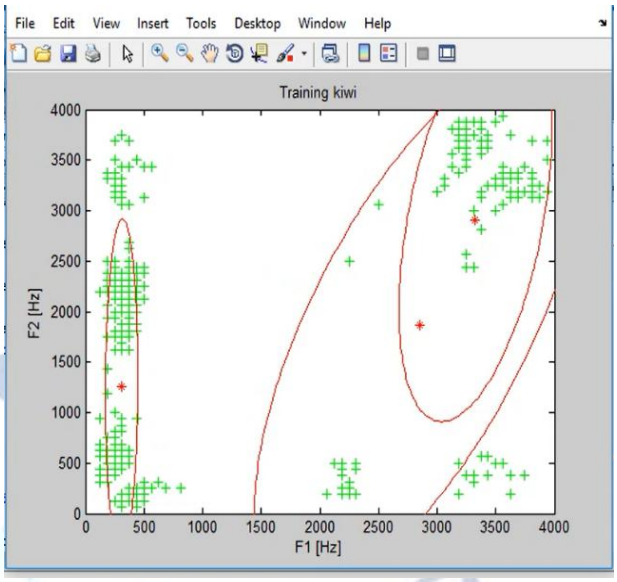
(a)



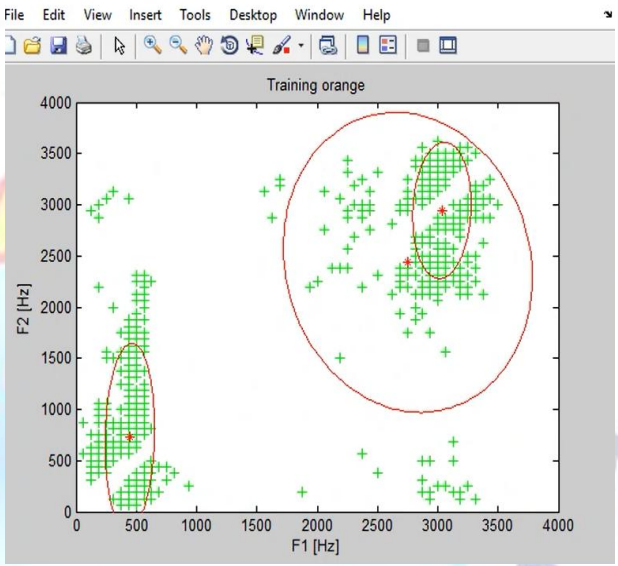
(b)



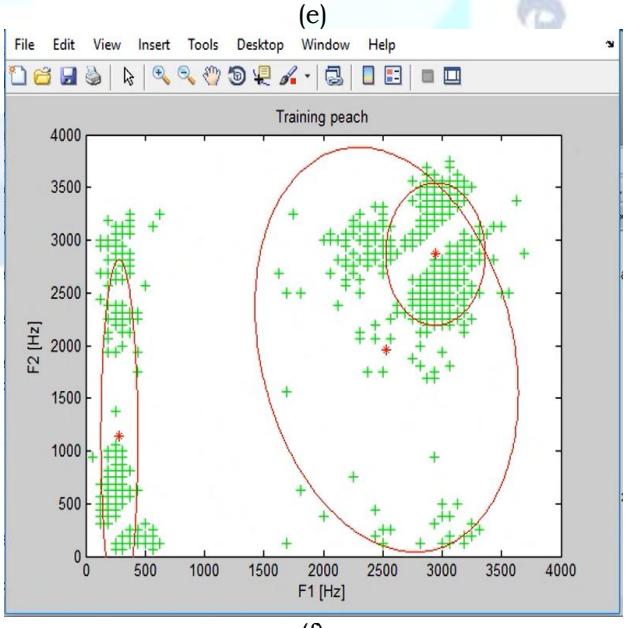
(c)



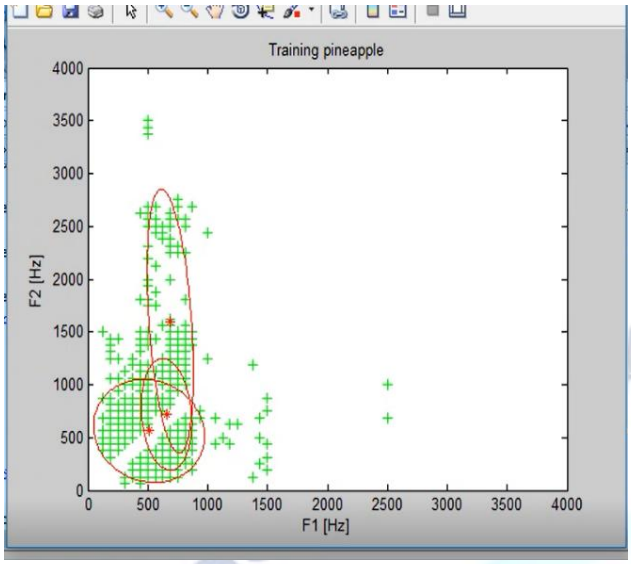
(d)



(e)



(f)



(g)

Figure 5: Training of the words using Baum-welch algorithm.

6. CONCLUSION

The primary contribution of this work is to design Speech recognition using full length hidden Markov Model. The system is designed by MATLAB. In this MFCC is used for feature extraction and HMM is used for training the model. The performance of the system is evaluated using maximum word recognition rate. The accuracy of recognizing the word is 97 percent.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] J. Meng, J. Zhang and H. Zhao,"Overview of the"SpeechRecognitionTechnology,"2012,pp.199-202.

[2] D.O'Shaughnessy," Automatic speechrecognition," 2015 CHILEAN Conference on Electrical,Electronics Engineering,Information and Communication Technologies (CHILECON) 2015, pp. 417-424.

[3] M. A. Hossan, S. Memon and M. A. Gregory,"A novel approach for MFCC feature extraction,"2010 4th International Conference on Signal Processing and Communication Systems, 2010, pp. 1 -5.

[4] M. Sadeghi and H. Marvi, "Optimal MFCC features extraction by differential evolution algorithm for speaker recognition.

[5] S. Boruah and S. Basishtha, "A study on HMM based speech recognition system," 2013 IEEE International Conference on Computational Intelligence and Computing Research, 2013, pp. 1-5.