Speech to Text Conversion in Malayalam

Preena Johnson[1], Jishna K C[2], Soumya S[3]

[1]*(B.Tech graduate, Computer Science and Engineering, College of Engineering Munnar/CUSAT, India)*
[2]*(B.Tech graduate, Computer Science and Engineering, College of Engineering Munnar/CUSAT, India)*
[3]*(Assistant Professor, Computer Science and Engineering, College of Engineering Munnar/CUSAT, India)*

Abstract: The Speech recognition and related tasks for many languages are getting more common in the present scenario. In the case of Malayalam language, recognizing speech is a tedious task. This paper presents a speech to text conversion system for Malayalam language. The system considers only isolated words and is a speaker dependent one with limited vocabulary. The uttered word which is the input for the system is displayed in Malayalam as the output. After the recording phase, features are extracted using Mel-Frequency Cepstral Coefficients (MFCC). Audio files are trained using the Gaussian Mixture Model (GMM).

Keywords: Feature extraction, GMM, Malayalam, MFCC, Speech to text

I. INTRODUCTION

The most common method to interact with computers is using keyboard and mouse. When a large amount of data is to be entered, this is a time consuming process. The mode of interaction can be changed in order to solve this problem. According to human beings, the best way of communication between them is speech. If a system can understand what a human speaks, then it is the best method of interaction between a human and a computer. Many works related to natural language processing are going on these days. Speech to text systems take speech as input, recognize it and convert it into text. A speech to text system support many applications such as an aid for blind persons, telephone directory assistance, in hospitals for health care instruments, in banking, in mobile phones etc.

Malayalam [1][2][3] is one among the Dravidian family languages. It is the official language of state Kerala. The alphabet of Malayalam contains 37 consonants and 16 vowels. The consonants are arranged based on the mode of speech production and the flow of air. Numerous works have been taken place in many of the Indian languages. However, very less work has been recorded in Malayalam. In this paper, we are introducing a speech to text system for Malayalam.

Our system is considering 5 isolated words for the training. This is a speaker dependent system. Initially the words are to be stored and trained. For each word, record a number of samples and store it. The word uttered will be compared with these stored words.

The main stage of a speech recognizer is the feature extraction. Many feature extraction techniques[4][5] are being used such as Linear Predictive Coding (LPC)[2], Linear Discriminant Analysis(LDA), Independent Component Analysis (ICA), Principal Component Analysis(PCA), MFCC[6][1], Kernal based feature extraction, Wavelet Transform and spectral subtraction. The most commonly used technique is MFCC. MFCC considers frequencies with the human perception sensitivity, and therefore it is a best tool for speech recognition.

For the process of recognizing speech, Hidden Markov Model[3][7][8], GMM[6][9], Vector Quantization(VQ), Artificial Neural Networks(ANN)[10] are various techniques[5][8] used. In our system, Gaussian component densities are weighted up to a sum and are represented to a parametric probability density function which acts as GMM. Python is used for the implementation and a user interface is provided.

II. Methodology

2.1. Recording

First, the words are recorded and stored. For recording, in Python, we are using a tool called Pyaudio which has inbuilt audio related functions. For each word, a good number of samples are to be stored as separate files. The parameters are to be specified in the audio stream opened such as frame size, format, channels, rate etc.

2.2. Feature Extraction

The efficient method for the feature extraction is MFCC. Various steps are involved in the MFCC algorithm. After the execution of this phase, Mel-frequency cepstral coefficients are obtained. The steps involved in this algorithm are discussed in the coming section. There are seven steps involved in MFCC:

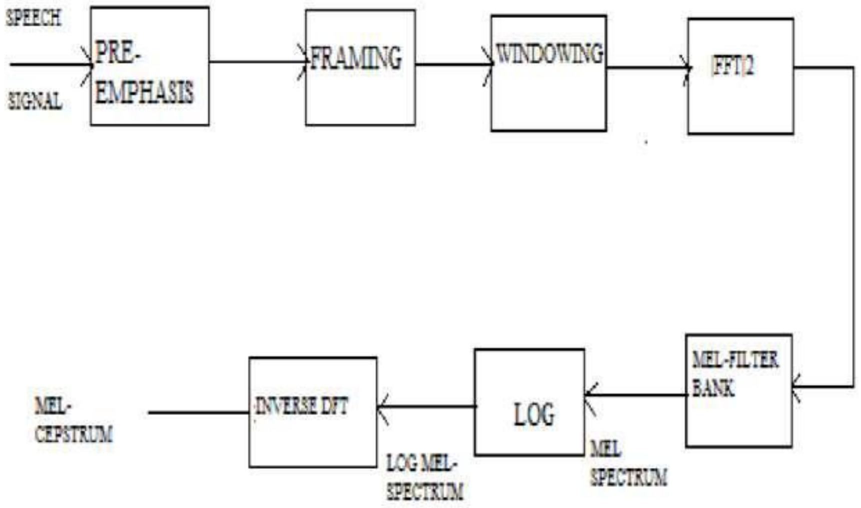


Fig 1.Block diagram of MFCC

2.2.1. Pre-emphasis

Higher frequency parts that are suppressed when human produces sound are compensated in the pre-emphasis phase. Moreover, it can amplify the importance of high frequency syllables *i.e.* the word spoken.

2.2.2.Framing

The input is segmented into frames of 20~30 ms with an optional overlap. To facilitate the use of FFT, usually the frame size is equal to the power of two.

2.2.3.Windowing

Hamming window is applied to each frame in order to keep the last and first points and its continuity in the frame. Let the signal in a frame be denoted by s(n), n=0,.....,N-1, then after Hamming windowing, the signal can be given as s(n)*w(n), where w(n) is the Hamming window.

2.2.4. FFT

In speech signals different timbres correspond to different energy distribution in the spectral analysis over frequencies. Therefore FFT is performed to obtain the magnitude of frequency of each frame. For performing FFT [11] on a frame, we assume that the signal in the frame is continuous and periodic when wrapping around. Even though this is not the case, FFT can be performed but the frame's first and last points' incontinuity can introduce undesirable effect in the frequency responses. We have two strategies to compromise with this problem,

1. Hamming window is multiplied with each frame in order to improve the continuity of last and first points.
2. Take a variable size frame in a way that it always contains integer multiple number of the fundamental intervals of the speech signal.

2.2.5. Mel-Filter Bank

We use a set of 20 triangular bandpass filters. The spectral envelope is extracted using these filters.

2.2.6.Logging

The magnitude frequency response is multiplied by the filters to find the log energy of each bandpass filter. Along the Mel frequency, these filters are equally spaced.

2.2.7. Inverse DFT

Since we have performed FFT, Discrete Cosine Transform (DCT) makes a transformation from the frequency domain into a time-like domain called quefrency domain. The features obtained are similar to a cepstrum, thus it is referred as the mel-scale cepstral coefficients.

2.2. GMM

The GMM[12] is used as a classifier to compare the extracted features from MFCC with stored templates. A Gaussian Mixture Model is a probabilistic model. It assumes all the data points are found from a mixture of a finite amount of Gaussian distributions with unknown parameters. A weighted sum of M component densities b_i is the Gaussian mixture density and it is expressed as:

$$p\left(\frac{\bar{x}}{\lambda}\right) = \sum_{i=1}^{M} p_i b_i(\bar{x}) \tag{1}$$

The mean,vectors, co-variance matrices and mixture weights from all component densities are used for describing GMM. Euclidian distance between various recordings is found for the matching purpose and hence a correct match is found.

III. Implementation

The speech signal is recorded by using a convenient package called Pyaudio as we are developing the system using Python. It is stored as a .wav file. The uttered word is identified by removing the silence. For extracting the features of the speech signal, MFCC is applied. The number of samples chosen in a frame is 256.After this phase, we will get the Mel-frequency cepstral coefficients. As a next step, GMM model parameters are produced. Euclidean distance between the various recordings in the databases is found and matching word is found. The matched word is then displayed in Malayalam. A graphical user interface is provided for recording the voice by the use of PyQt4.

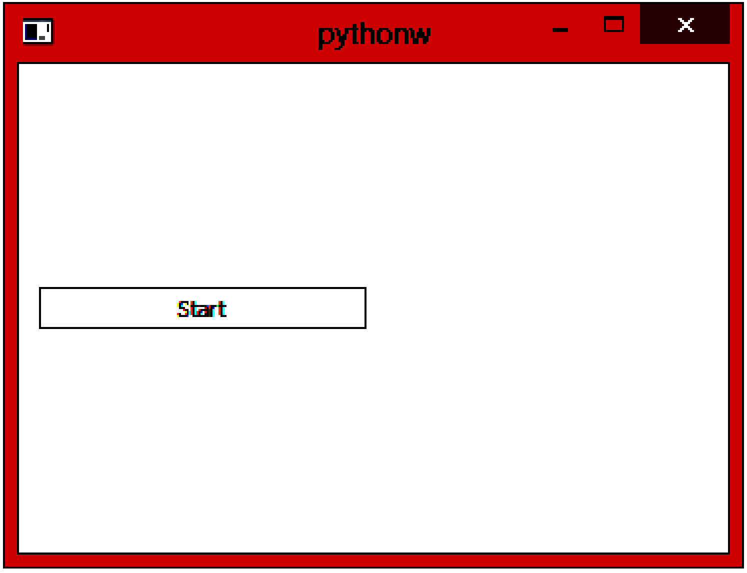
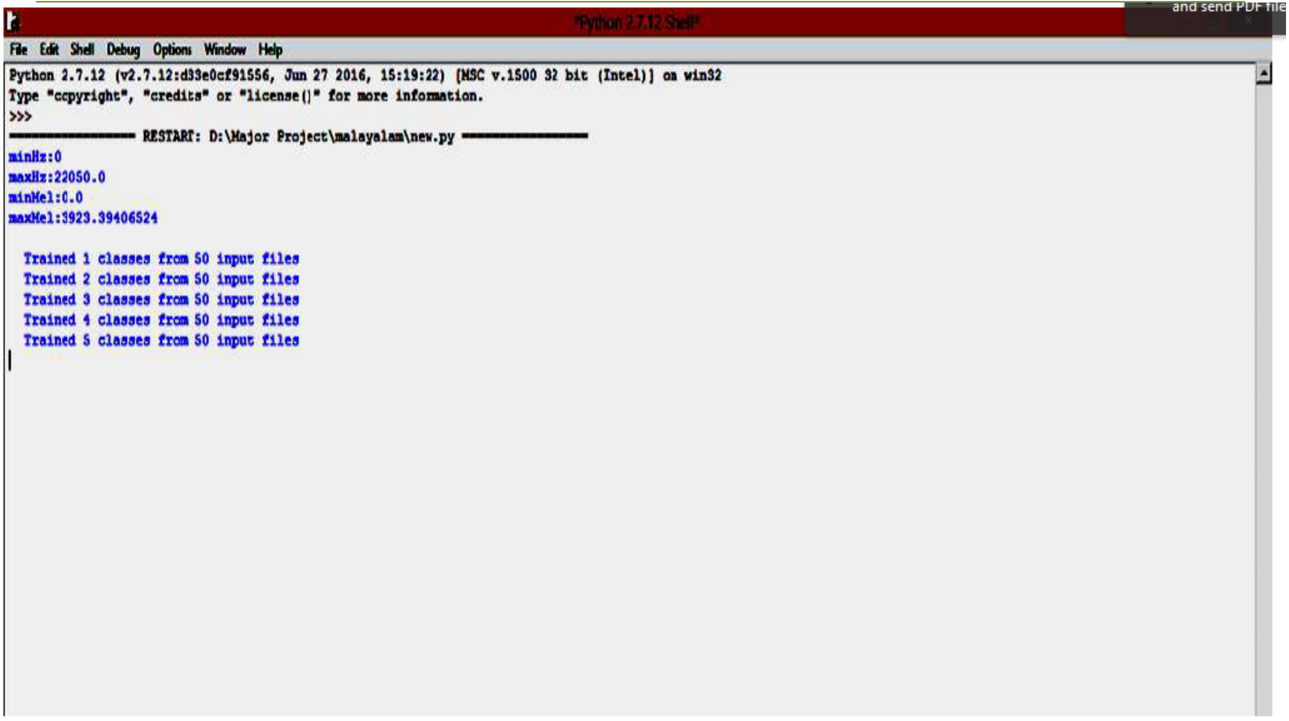


Fig 2.User Interface for recording

IV. Results

The system is trained for five words. The words are അമ്മ, പൃഥി, കഥ, വേഴാമ്പൽ , കേരളം. Each word has 5 samples recorded. The outputs for the word are tested 25 times and accuracy percentage is calculated.



```
Python 2.7.12 (v2.7.12:d33e0cf91556, Jun 27 2016, 15:19:22) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
================ RESTART: D:\Major Project\malayalam\new.py ================
minHz:0
maxHz:22050.0
minMel:0.0
maxMel:3923.39406524

  Trained 1 classes from 50 input files
  Trained 2 classes from 50 input files
  Trained 3 classes from 50 input files
  Trained 4 classes from 50 input files
  Trained 5 classes from 50 input files
```

Fig 3.The word is recorded

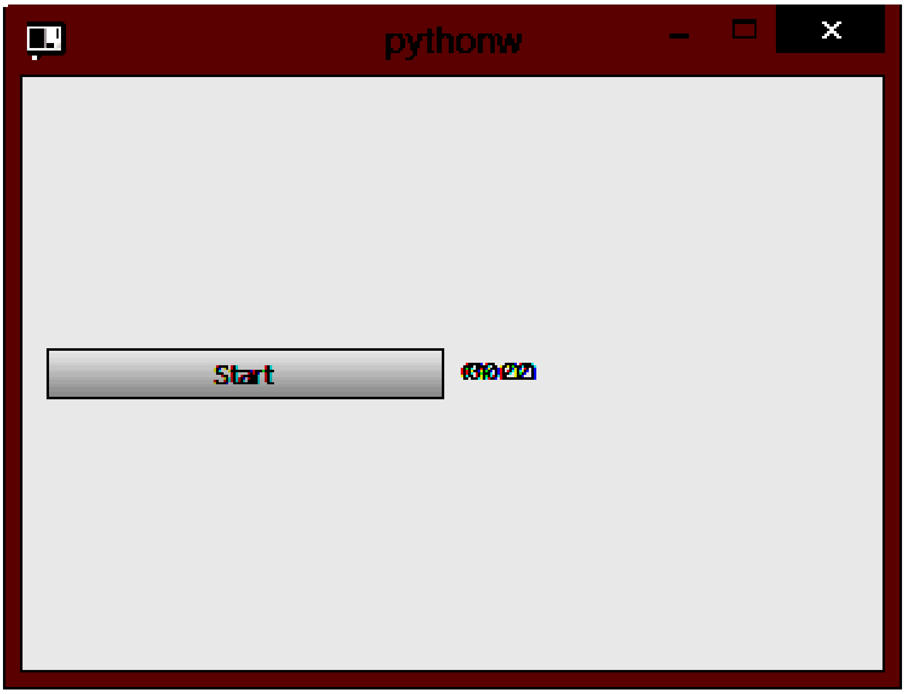


Fig 4.The word is displayed.

The accuracy of the system is found out using confusion matrix as shown in Table 1.

Table 1 Testing and Accuracy of Results

Train Data	Number of test	Number of Correct test	Error	Percentage of accuracy
അമ്മ	25	19	6	76
കേരളം	25	18	7	72
കഥ	25	20	5	80
വേഴാമ്പൽ	25	17	8	68
പൃഥി	25	19	6	76

Conclusion

The system is established as an initiative towards an advanced speech to text conversion system for Malayalam. The system is giving an accuracy of about 75% when modelled using GMM. We would like to enlarge this project to a speaker independent system which also deals with a large vocabulary system.

References

[1]. Cini Kurian, Kannan Balakrishnan, “Speech recognition of Malayalam numbers”,*World Congress on Nature & Biologically Inspired Computing(NaBIC2009),2009*.

[2]. Maya Moneykumar,Elizabeth Sherly ,Win Sam Varghese, “Malayalam word identification for speech recognition system”,*An International Journal of Engineering Sciences,*Special Issue iDravadian,December2014,Vol. 15

[3]. Cini Kurian, Kannan Balakrishnan, “Development & evaluation of different acoustic models for Malayalam continuous speech recognition”, *International Conference on Communication Technology and System Design* 2011.

[4]. R. K. Aggarwal, Mayank Dave, “Implementing a Speech Recognition System Interface for Indian Languages”.

[5]. Miss.PrachiKhilari,Prof.BhopeV.P"A Review on Speech to Text Conversion Methods", *International Journal of Advanced Research in Computer Engineering & Technology* Volume 4 Issue 7, July 2015.

[6]. Tahira Mahboob, Memoona Khanum, Malik Sikandar Hayat Khiyal, Ruqia Bibi4, “Speaker Identification Using GMM with MFCC”, *IJCSI International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015.*

[7]. Nuzhat Atiqua Nafis and Md. Safaet Hossain, “Speech to text conversion in real time”, *International Journal of Innovation and Scientificc Research ISSN 2351-8014 Vol. 17 No. 2 Aug. 2015.*

[8]. L. R. .Rabiner, B. H. Juang ,”An Introduction to Hidden Markov Models”, *IEEE ASSP magazine, January 1986.*

[9]. Virendra Chauhan, Shobhana Dwivedi, Pooja Karale, Prof. S.M. Potdar,” Speech to text converter using Gaussian Mixture Model(GMM)”, *International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 02 | Feb-2016*

[10]. MayaMoneykumar, ElizabethSherly, WinSamVarghese, “Isolated Word Recognition System for Malayalam using Machine Learning”,*2016*

[11]. AnshulGupta, NileshkumarPatel, ShabanaKhan,”*Automatic Speech Recognition Technique For Voice Command”.*

[12]. Virendra Chauhan, Shobhana Dwivedi, Pooja Karale, Prof. S.M. Potdar, “Speech To Text Converter Using Gaussian Mixture Model(Gmm)”, *International Research Journal of Engineering and Technology , Volume: 03 Issue: 02, Feb-2016*